

# RECIIS

Revista Eletrônica de  
Comunicação, Informação  
& Inovação em Saúde

Rio de Janeiro, v.3, n.1, março 2009

ISSN 1981-6278

Número Temático

**Ontologias,  
Web Semântica  
e Saúde**

ISSN 1981-6278



**RECIIS**

Revista Eletrônica de Comunicação  
Informação & Inovação em Saúde

---

*Rio de Janeiro, v. 3, n. 1, mar., 2009*





## Conselho Editorial

Alan Radley, Loughbough University, Loughbough, Grã-Bretanha

Andre Parente, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil

Antonio Fausto Neto, Universidade do Vale do Rio dos Sinos, São Leopoldo, Brasil

Arie Rip, University of Twente, Twente, Holanda

Benoît Godin, University of Québec, Québec, Canadá

Blaise Cronin, Indiana University, Bloomington, Indiana, Estados Unidos

Carlos Morel, Centro de Desenvolvimento Tecnológico em Saúde, Fiocruz, Rio de Janeiro, Brasil

Christer Hogstedt, National Institute of Public Health, Stockholm, Suécia

Dominique Pestre, Centre National de la Recherche Scientifique, Paris, França

Eduardo Albuquerque, Universidade Federal de Minas Gerais, Belo Horizonte, Brasil

Emily Martin, New York University, New York, Estados Unidos

Emir Suaiden, Universidade de Brasília, Brasília, Brasil

Everardo Nunes, Universidade Estadual de Campinas, Campinas, Brasil

Francisco Bastos, Inst. Com. Inf. Cient. Tecnol. Saúde, Fiocruz, Rio de Janeiro, Brasil

Geoffrey Bowker, Santa Clara University, Santa Clara, California, Estados Unidos

Hector Abreu, Instituto Nacional de Câncer, Rio de Janeiro, Brasil

Hiroko Yamane, National Graduate Institute for Policy Studies, Tokyo, Japão

Inesita de Araújo, Inst. Com. Inf. Cient. Tecnol. Saúde, Fiocruz, Rio de Janeiro, Brasil

Joan Fujimura, University of Wisconsin-Madison, Madison, Wisconsin, Estados Unidos

Joanna Chataway, The Open University, Milton Keynes, Grã-Bretanha

João Arriscado, Universidade de Coimbra, Coimbra, Portugal

Jorge Veiga, Instituto de Salud Carlos III, Madrid, Espanha

José Bassani, Universidade Estadual de Campinas, Campinas, Brasil

Kanikaram Satyanarayana, Indian Council of Medical Research, New Delhi, Índia

Kathy Charmaz, Sonoma State University, Rohnert Park, California, Estados Unidos

Lea Velho, Universidade Estadual de Campinas, Campinas, Brasil

Lita Nelsen, Massachusetts Institute of Technology, Boston, Massachusetts, Estados Unidos

Loet Leydesdorff, University of Amsterdam, Amsterdam, Holanda

Luigi Palombi, The Australian National University, Canberra, Austrália

Madel Luz, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brasil

Maged Kamel Boulos, University of Plymouth, Plymouth, Grã-Bretanha

Manuel Limonta, Institute of Hematology and Immunology, Havana, Cuba

Márcia Teixeira, Escola Politécnica de Saúde Joaquim Venâncio, Fiocruz, Rio de Janeiro, Brasil

Maurice Cassier, Institut National de la Santé et de la Recherche Médicale, Paris, França

Nelly Oudshoorn, University of Twente, Twente, Holanda

Paulo Elias, Universidade de São Paulo, São Paulo, Brasil

Peter Ganea, Max Planck Institute for Intellectual Property, Competition and Tax Law, Munich, Alemanha

Pierre Lévy, University of Ottawa, Ottawa, Canadá

Pierre Tambourin, Institut National de la Santé et de la Recherche Médicale, Paris, França

Reinaldo Guimarães, Secretaria de Ciência e Tecnologia e Insumos Estratégicos, Ministério da Saúde, Brasília, Brasil

Rita Barradas Barata, Santa Casa de São Paulo, São Paulo, Brasil

Sandra Braman, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, Estados Unidos

Sandra Harding, University of California, Los Angeles, California, Estados Unidos

Sarita Albagli, Instituto Brasileiro de Informação em Ciência e Tecnologia, Rio de Janeiro, Brasil

Sergio Pena, Universidade Federal de Minas Gerais, Belo Horizonte, Brasil

Soraya Cortes, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil

Terry Shinn, Centre National de la Recherche Scientifique, Paris, França

Timothy Lenoir, Duke University, Durham, North Carolina, Estados Unidos

Walter Zin, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil

Walter Colli, Universidade de São Paulo, São Paulo, Brasil

## Editores Científicos

Carlos Saldanha, Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Fiocruz, Brasil

Josué Laguardia, Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Fiocruz, Brasil

## Editores Convidados

Frederico Freitas, Universidade Federal de Pernambuco, Recife, Brasil

Stefan Schulz, Centro Médico Universitário de Freiburg, Alemanha

## Editores de Seção

### Artigos Originais

Álvaro Matida, Associação Brasileira de Pós Graduação em Saúde Coletiva, Brasil

Bianca Cortes, Escola Politécnica de Saúde Joaquim Venâncio, Fiocruz, Brasil

Debora Diniz, Instituto de Ciências Humanas, Universidade de Brasília, Brasil

Regina Maria Marteleto, Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Fiocruz, Brasil

Roseni Pinheiro, Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro, Brasil

### Pesquisas em Andamento

Christovam Barcellos, Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Fiocruz, Brasil

Katia Lerner, Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Fiocruz, Brasil

### Avanços Tecnológicos

Frederico Freitas, Centro de Informática, Universidade Federal de Pernambuco, Brasil

Laura Cristina Simões Viana, Vice-Presidência de Pesquisa e Desenvolvimento Tecnológico, Fiocruz, Brasil

Lia Hasenclever, Instituto de Economia, Universidade Federal do Rio de Janeiro, Brasil

### Artigos de Revisão

Julia Guivant, Centro de Filosofia e Ciências Humanas, Universidade Federal de Santa Catarina, Brasil

Maria Conceição da Costa, DPCT, Instituto de Geociências, Universidade Estadual de Campinas, Brasil

### Ensaios

Regina Erthal, Instituto de Pesquisa Clínica Evandro Chagas, Fiocruz, Brasil

Sérgio Carrara, Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro, Brasil

### Resenhas

Ana Filipecski, Escola Politécnica de Saúde Joaquim Venâncio, Fiocruz, Brasil

### Cartas

Rejane Machado, Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Fiocruz, Brasil

## Conselho Consultivo Local

Carlos Vogt, Secretário de Ensino Superior do Estado de São Paulo

Cecília Minayo, Coord. Científica do Claves/Ensp, Fiocruz, Brasil

Cícera Silva, Coord. Adj. do Curso de Especialização do Inst. Com. Inf. Cient. Tecnol. Saúde, Fiocruz, Brasil

João Aprígio, Coord. Banco de Leite/Instituto Fernandes Figueira, Fiocruz, Brasil

José Carneiro, Presidente da Associação Brasileira de Pós-Graduação em Saúde Coletiva, Brasil

Moses Goldbaum, Prof. do Depto. de Medicina Preventiva, Faculdade de Medicina/Universidade de São Paulo, Brasil

Paulo Gadelha, Presidente da Fundação Oswaldo Cruz, Brasil

Ricardo Ceccim, Prof. de Educação na Saúde, Universidade Federal do Rio Grande do Sul, Brasil

Tânia de Araújo Jorge, Diretora do Instituto Oswaldo Cruz, Fiocruz, Brasil

## Produção Editorial

### Editor Administrativo

Luciane Wilcox, Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Fiocruz, Brasil

### Secretária

Gisele Neves, Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Fiocruz, Brasil

### Assistente do Editor Científico

Helena Klein, Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Fiocruz, Brasil

### Normalização

Rejane Machado, Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Fiocruz, Brasil

### Editores de Arte

Mauro Campello, Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Fiocruz, Brasil

Rodrigo Murtinho, Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Fiocruz, Brasil

### Desenvolvedor Web e Suporte

Técnicos  
Marcus Lessa, Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Fiocruz, Brasil

### Diagramação

Letra e Imagens Editora Ltda.



# Sumário

## Apresentação

<b>Ontologias, Web semântica e saúde</b> .....	4
<i>Frederico Freitas e Stefan Schulz</i>	

## Artigos originais

<b>Pesquisa de terminologias e ontologias atuais em biologia e medicina</b> .....	8
<i>Fred Freitas, Stefan Schulz e Eduardo Moraes</i>	

<b>Bases ontológicas e conceituais para um modelo do conhecimento científico em artigos biomédicos</b> .....	21
<i>Carlos Henrique Marcondes, Marília Alvarenga Rocha Mendonça, Luciana Reis Malheiros, Leonardo Cruz da Costa e Tatiana Cristina Paredes Santos</i>	

<b>Vantagens e limitações das ontologias formais na área biomédica</b> .....	33
<i>Stefan Schulz, Holger Stenzhorn, Martin Boeker e Barry Smith</i>	

<b>Uma análise ontológica do eletrocardiograma</b> .....	49
<i>Bernardo Gonçalves, Veruska Zamborlini e Giancarlo Guizzardi</i>	

<b>Aspectos metodológicos no reuso de ontologias: um estudo a partir das anotações genômicas no domínio dos tripanosomatídeos</b> .....	64
<i>Maria Luiza de Almeida Campos, Maria Luiza Machado Campos, Alberto M. R. Dávila, Hagar Espanha Gomes, Linair Maria Campos e Laura Lira</i>	

<b>Extração automática de termos compostos para construção de ontologias: um experimento na área da saúde</b> .....	76
<i>Lucelene Lopes, Renata Vieira, Maria José Finatto, Daniel Martins, Adriano Zanette e Luiz Carlos Ribeiro Jr.</i>	

## Resenhas

<b>Semantic Web services, processes and applications, Jorge Cardoso; Amit P. Sheth (Eds.) / Semantic Web and semantic Web services, Liyang Yu</b> .....	89
<i>Por Laura Cristina Simões Viana</i>	

<b>Semantic Web technologies: trends and research in ontology-based systems, John Davies; Rudi Studer; Paul Warren</b> .....	93
<i>Por Karin Breitman</i>	

## Ontologias, Web semântica e saúde

DOI: 10.3395/reciis.v3i1.238pt



**Frederico Freitas**

Centro de Informática,  
Universidade Federal de  
Pernambuco, Recife, Brasil  
fred@cin.ufpe.br



**Stefan Schulz**

Instituto de Biometria  
Médica e Informática da  
Medicina, Centro Médico  
Universitário de Freiburg,  
Freiburg, Alemanha  
stschulz@uni-freiburg.de

Com uma longa tradição de estudo no ramo da filosofia, que remonta há pelo menos remotos 23 séculos, o termo “Ontologia” transformou-se num dos termos mais em moda no mundo da informática de hoje, sendo aplicado em sistemas de muitos outros campos, incluindo-se, entre os principais, biologia e medicina. Enquanto o termo “Ontologia” outrora denotava um ramo da metafísica, “ontologias” são hoje entendidas como vocabulários formais que descrevem as premissas básicas de um determinado domínio.

Há pelo menos uma razão principal para todo este interesse por parte dos informatas. Segundo Tim Berners-Lee, um dos principais responsáveis pela sedimentação da própria internet, as ontologias constituem o componente e a motivação principais da Web semântica, uma Web em que os programas e sistemas são capazes de “entender” e processar dados das páginas, de acordo com o contexto. Mas o que são ontologias e como elas ajudam os sistemas a conseguir processar os dados com tal profundidade?

Em palavras práticas, ontologias padronizam significado através de identificadores semânticos, os quais podem representar o mundo real e conceitual. Ontologias constituem-se em definições de conceitos, classes, propriedades, relações, restrições e axiomas sobre um determinado domínio (por exemplo, neurologia).

Ao longo dos artigos deste número temático são apresentados vários exemplos de ontologias, que variam em complexidade, expressividade etc. Então, se uma

página da Web referencia uma ou mais ontologias (por exemplo, dizendo que um determinado trecho da página é um nome de professor, referenciando este conceito de uma ontologia sobre Academia), ela está atribuindo significado ao conteúdo da página de forma que um software poderá usar as outras definições da ontologia que definem as relações e restrições *a priori* relacionadas ao conceito referenciado e interpretará outras partes da página dentro do contexto definido pela ontologia. Por exemplo, ao achar uma lista de alunos de doutorado na página do professor, o software poderá interpretar que aqueles alunos são orientados pelo professor, que trabalham em áreas que o professor pesquisa (e que podem também estar na página do professor), e que estão matriculados no programa de pós-graduação ao qual o professor pertence. Digno de nota é o fato de que estas informações podem não estar literalmente presentes na página Web. Portanto, processar informações usando ontologias, que provêem excelente contexto para o entendimento das informações, tanto para usuários humanos quanto para agentes de software, vem se tornando uma tendência em várias áreas e tipos de aplicações. A Web semântica, por exemplo, deve ter um forte impacto sobre os sítios comerciais da Web; ontologias também servem como vocabulário em uma troca de mensagens entre os ditos agentes inteligentes, entidades de software que podem raciocinar sobre conhecimento, permitindo que estes agentes possam negociar pedidos de compras em nome de suas empresas – e já existem protótipos acadêmicos simulando esta situação.

Entre as áreas mais comuns de aplicação de ontologia estão medicina e biologia, existindo, inclusive, alguns centros quase que exclusivamente dedicados ao estudo desta tecnologia. Nos Estados Unidos, o Departamento de Informática Médica da Universidade de Stanford, criou o mais empregado editor de ontologias, o Protégé. Existe também uma organização virtual, o Centro Nacional de Pesquisas em Ontologias Biomédicas (National Center for Biomedical Ontologies), que envolve os grupos de pesquisa das universidades de Stanford, Victoria e Buffalo, além da Mayo Clinic, em Rochester. Na Europa, existem o Grupo de Pesquisa em Ontologias em Medicina e Ciências Biológicas no Instituto de Informática Médica, Estatística e Epidemiologia, em Leipzig, e o Grupo de Pesquisa em Informática Médica no Centro Médico Universitário de Freiburg, ambos na Alemanha, e o Grupo em Informática Biológica e Saúde, na Universidade de Manchester, na Inglaterra, estão entre os que mais pesquisam em ontologias biológicas e Web Semântica. A complexidade do conhecimento médico e biológico torna difícil a confecção de sistemas tradicionais, pois para assistir tarefas médicas, os sistemas precisam de muito conhecimento e capacidade de inferência. Esta talvez seja a principal razão da aplicabilidade e conseqüente sucesso do uso de ontologias nestas áreas. Discorremos um pouco mais acerca destas necessidades no primeiro artigo. Apenas cabe ressaltar que o volume de pesquisas produzido ligando ontologias, Web semântica e saúde justificaram o lançamento deste número temático e aceitamos a tarefa, convencidos de que a RECIIS é, de fato, um veículo em que todos estes tópicos de pesquisa são contemplados. A comunidade que pesquisa nestes tópicos respondeu à chamada de trabalhos satisfatoriamente e os editores se sentem agradecidos com as contribuições que nos foram enviadas.

Das submissões recebidas para este número temático, foram selecionados cinco artigos pelos revisores, que, juntados a este artigo introdutório dos editores convidados, compõem os seis artigos deste número temático. Durante a confecção do número temático, procuramos ordenar os artigos de forma a ir apresentando gradativamente os conceitos e práticas em ontologias biomédicas em grau crescente de complexidade, uma vez que os dois primeiros temas do número temático – Ontologias e Web semântica - ainda não são muito difundidos no Brasil. Abaixo, descreveremos brevemente cada um dos artigos do número temático.

O primeiro artigo, intitulado “Levantamento das atuais terminologias e ontologias em biologia e medicina”, tem por objetivo apresentar conceitos introdutórios sobre ontologias e sistemas terminológicos em biologia e medicina, proporcionando uma visão abrangente do assunto pela descrição de sistemas mais difundidos e/ou emblemáticos. Após uma breve discussão sobre o uso de terminologias em contraste com o uso de ontologias, são apresentados os sistemas CID (Classificação Internacional de Doenças), SNOMED CT (Nomenclatura Sistematizada de Termos médico-clínicos), MeSH (Descritores de Saúde), openGalen (Arquitetura Geral para Linguagens, Enciclopédias e Nomenclaturas in

Medicine, FMA (Modelo Fundacional de Anatomia), bem como as iniciativas UMLS (Sistema Unificado de Linguagem Médica), a OBO Foundry (Fundação de Ontologias Biomédicas Aberta”) em geral e a Ontologia Gênica (GO) em particular.

O artigo “Bases ontológicas e conceituais para um modelo do conhecimento científico em artigos biomédicos”, escrito por Carlos Marcondes e co-autores, propõe uma classificação de artigos científicos da área de Medicina. Cada classe possui um modelo de anotação semântica, anotações estas que devem ser efetuadas usando ontologias, de forma que o conhecimento expresso nos artigos consiga ser processado e “entendido” por sistemas computacionais. Antes da proposição, o artigo traz uma discussão que justifica a existência do modelo a partir de teorias de Metodologia Científica e Filosofia da Ciência e da análise de uma base de 75 artigos médicos. O modelo habilita a recuperação semântica de informações de artigos, permitindo buscá-lo através de conceitos e relações, ao invés de palavras-chave como acontece nos engenhos de busca tradicionais.

O artigo “Vantagens e limitações de ontologias formais no domínio biomédico”, elaborado por Stefan Schulz e co-autores, dentre os quais Barry Smith, filósofo e autor de vários livros e artigos bastante conhecidos sobre bio-ontologias e diretor de várias organizações relacionadas ao assunto, como a própria Fundação OBO, serve como excelente introdução para os formalismos matemáticos usados em ontologias biomédicas, propondo uma série de critérios para delimitar o conceito de ontologia de recursos de representação de conhecimento de uma forma mais abrangente. As vantagens e desvantagens de cada representação são cuidadosamente discutidas, a partir de tesouros, fonte de conhecimento freqüentemente disponível na área biomédica, até a expressiva lógica de descrições - o mais expressivo formalismo de representação de conhecimento padronizado pelo Consórcio da Web (W3C). A discussão gira, primordialmente, em torno da relação custo-benefício em adicionar-se expressividade à forma de representação em uso. Exemplos simples usando lógica de descrição são introduzidos, em complexidade gradativa, desta forma ajudando o leitor a ter uma idéia concreta de como são representados conceitos biomédicos complexos e de como estes conceitos são usados durante o processo de raciocínio automático, bem como possíveis resvalos de modelagem que podem levar a raciocínios automáticos incorretos.

O artigo seguinte, “Uma análise ontológica do eletrocardiograma” por Bernardo Gonçalves e co-autores, traz a descrição de uma ontologia cuidadosamente construída para possibilitar o raciocínio sobre um domínio complexo, específico e desafiador em termos de representação, a execução de um eletrocardiograma. A ontologia foi elaborada empregando-se duas ontologias biomédicas bastante ricas e populares, a OBO e a FMA - já descritas no primeiro artigo - e a ontologia de topo UFO (Ontologia Fundacional Unificada). Esta ontologia de topo empresta bases sólidas para a representação de elementos complexos presentes direta ou indiretamente




em eletrocardiogramas, tais como relações parte-todo universais e relações temporais de produção, caracterização e geração, classificando-as com suas meta-características adequadas, sejam elas simétricas ou anti-simétricas, reflexivas ou irreflexivas e transitivas ou intransitivas. Portanto, a elaboração da ontologia ECG constitui um bom exemplo de como empregar princípios de modelagem bem fundamentados para garantir que o raciocinador automático não derive conclusões errôneas por falta de conhecimentos básicos, como as consequências de um conceito ser parte de outro. Ao fim do artigo, são sugeridas algumas formas de aplicação do conhecimento disponibilizado.

Os últimos dois artigos são dedicados à construção de ontologias, mas não sobre o ponto de modelagem. O artigo “Aspectos metodológicos no reuso de ontologias: um estudo a partir das anotações genômicas no domínio dos tripanosomatídeos”, de Maria Luiza Campos e co-autores, discute em linhas gerais o reuso de ontologias, descrevendo ainda seus experimentos na área biomédica, a descrição do genoma de tripanosomatídeos. Foi empregado um alinhador, cuja função é comparar duas ontologias, mostrando as correspondências entre os elementos destas ontologias, uma tarefa correlata à tarefa de reuso. Os autores findam o artigo com uma lista de conclusões acerca de seu experimento específico, ou seja, em que casos o alinhador correspondeu às expectativas dos autores.

Finalmente, o artigo “Extração automática de termos compostos para construção de ontologias: um experimento na área da saúde”, confeccionado por Lucelene Lopes e Renata Vieira, discorre sobre uma ferramenta de aprendizado automático de ontologias,

elaborada dentro do grupo de pesquisa dos autores. A ferramenta está fundamentada em técnicas específicas de uma subárea específica de estudo de Inteligência Artificial, conhecida como Processamento de Linguagem Natural, que se debruça sobre o processamento de textos do ponto de vista lingüístico, acarretando, portanto, em um processamento muito mais profundo das informações textuais. Estas técnicas são usadas para extrair elementos que irão compor uma ontologia. Este tipo de abordagem é particularmente interessante por automatizar o processo de produção de conhecimento para as ontologias, acelerando o processo de obtê-las. As autoras descrevem experimentos iniciais com *corpus* sobre pediatria. Os experimentos mostraram que de fato abordagens baseadas em aprendizado podem ser úteis na etapa de construção da terminologia para uma ontologia, não garantindo, porém, a cobertura necessária para incluir toda a terminologia contida no *corpus*.

Como últimas palavras, gostaríamos de agradecer aos revisores, Guilherme Ataíde, Werner Ceusters, Ronald Cornet, Marcos Galindo, Rosario Girardi, Giancarlo Guizzardi, Robert Hoehndorf e César Tacla, por realizarem um trabalho sério e de boa qualidade, sem o qual a produção deste número temático provavelmente não seria possível. Os editores convidados dedicam ainda um agradecimento aos editores científicos da RECIIS, Carlos Saldanha Machado e Josué Laguardia, por nos guiarem durante todo o processo de produção da edição, com várias dicas úteis, sendo muito atenciosos e respondendo quase imediatamente a uma longa lista de dúvidas nossas surgidas ao longo do trabalho, além de brindar-nos com a oportunidade de publicá-lo. 

## Sobre os editores

### *Fred Freitas*

É PhD pela Universidade de Santa Catarina, Brasil, e atualmente é afiliado ao Centro de Informática da Universidade Federal de Pernambuco, Brasil (CIn/UFPE). Conduziu pesquisas por quase um ano no Departamento de Informática da Universidade de Karlsruhe, como integrante do projeto Brasil-Alemanha “*A semantic approach to data retrieval*” (Abordagem semântica da recuperação de dados). Publicou diversos artigos em conferências e seminários de renome, como IJCAI e outros patrocinados pela ACM (*Association on Computer Machinery*) e pelo IEEE (*Institute of Electrical and Electronical Engineering*). Co-presidiu duas séries de seminários: O WONTO (*Workshop on Ontologies and their Applications/Seminário de Ontologias e Suas Aplicações*), no Brasil, e o BA-OSW (*Building Applications with Ontologies for the Semantic Web/Construção de Aplicações com Ontologias para a Semantic Web*), em Portugal. Co-editou Edições Especiais sobre temas relacionados do JBCS (*Journal of Brazilian Computer Society*) e do JUCS (*Journal of Universal Computer Science*). Colabora, atualmente, com a Universidade de Paul Cessane em Marselha, e INRIA, Montbonnot, na França, e as Universidades de Karlsruhe, Freiburg e Mannheim, na Alemanha. Suas áreas de interesse incluem ontologias, sistemas multiagentes, representação de conhecimento, mediação, e mineração de texto.

## *Stefan Schulz*

É formado em medicina pela Heidelberg University, Alemanha, e é pesquisador sênior e professor do Instituto de Biometria Médica e Informática da Medicina do Centro Médico Universitário Freiburg, onde chefia o Grupo de Pesquisas em Informática na Medicina. Seu trabalho se concentra em terminologias e ontologias biomédicas, representação do conhecimento biomédico, recuperação de documentos médicos multilíngües, mineração de texto e dados em repositórios de documentos clínicos, aprendizado eletrônico na Medicina, e informática da saúde em países em desenvolvimento.

Após executar trabalhos clínicos em cirurgia e medicina interna, obteve seu diploma de doutorado na área da higiene tropical, onde efetuou um estudo de campo parasitológico em São Luís, Brasil. Após obter qualificação técnica em computação médica, mudou-se para a Universidade de Freiburg, onde participou de projetos de desenvolvimento de software clínico e educacional, e de diversos projetos de pesquisa na área da extração de informações, terminologias biomédicas, engenharia da linguagem médica, e tecnologias semânticas. Tem desempenhado papéis de liderança em diversos projetos financiados pela União Européia. Stefan Schulz é autor de mais de cem publicações revisadas por especialistas, e recebeu vários prêmios. Tem oferecido repetidas contribuições a projetos de pesquisa na área da informática de saúde brasileira desde 2001, como pesquisador convidado da Pontifícia Universidade Católica do Paraná (PUC-PR).



**Artigos originais**

# **Pesquisa de terminologias e ontologias atuais em biologia e medicina**

DOI: 10.3395/reciis.v3i1.239pt



**Fred Freitas**

Centro de Informática,  
Universidade Federal de  
Pernambuco, Recife, Brasil  
fred@cin.ufpe.br



**Stefan Schulz**

Instituto de Biometria  
Médica e Informática da  
Medicina, Centro Médico  
Universitário, Freiburg,  
Alemanha  
stschulz@uni-freiburg.de

## **Eduardo Moraes**

Centro de Informática, Universidade Federal de Pernambuco, Recife, Brasil  
ecm2@cin.ufpe.br

## **Resumo**

Este documento apresenta o estado da arte das terminologias e ontologias aplicadas à Biologia e à Medicina. Sem a intenção de torná-lo inteiramente abrangente, descrevemos alguns dos recursos mais importantes que atualmente atraem interesse da indústria e da área acadêmica. Apresentamos uma estrutura descritiva, e comparamos os sistemas em termos de seus elementos de arquitetura, expressividade e cobertura, e também analisamos a natureza das entidades que eles denotam. Em especial, examinamos a Classificação Internacional de Doenças - CID, *Medical Subject Headings* - MeSH (Cabeçalhos Médicos), *Gene Ontology* - GO (Ontologia Genética), *Systematized Nomenclature of Medicine - Clinical Terms* - SNOMED CT (Nomenclatura Sistematizada de Medicina - Termos Clínicos), *Generalized Architecture for Languages, Encyclopaedias and Nomenclatures* - openGALEN, (Arquitetura Generalizada de Linguagens, Enciclopédias e Nomenclaturas), *Foundational Model of Anatomy* - FMA (Modelo Fundamental de Anatomia), *Unified Medical Language System* - UMLS (Sistema Unificado de Linguagem Médica) e *Open Biomedical Ontologies (OBO) Foundry* (Oficina de Ontologias Biomédicas Abertas).

## **Palavras-chave**

terminologias; ontologias; biologia; medicina

## **Introdução**

### **Panorama geral**

A crescente disponibilidade digital de enormes quantidades de fontes de dados e conhecimentos biomédicos sobrecarregou os pesquisadores e médicos com a

tarefa de gerenciar terabytes de conteúdo semântico, que é, naturalmente, sutilmente interconectado, e precisa ser agregado e manipulado. Uma vasta quantidade de dados utilizados para resolver tarefas complexas exige técnicas cada vez mais sofisticadas de gerenciamento inteligente de informação e conhecimento, aumentando a interope-

rabilidade de conteúdos em grandes repositórios apoiados por diferentes tipos de raciocínio automatizado. Este desafio tem sido cada vez mais enfrentado por biólogos, pesquisadores de saúde pública e clínica, economistas da saúde, e também por médicos. Um resultado prático desses esforços é o surgimento de um conjunto crescente de sistemas de referência semântica, muitas vezes caracterizados como vocabulários, tesouros, terminologias, e ontologias (Rubin 2007).

Os progressos atuais do gerenciamento do conhecimento biomédico têm essencialmente duas causas:

- o estabelecimento de vocabulários e sistemas de classificação indexadores, como a Classificação Internacional de Doenças, e o *Index Medicus*, do século XIX, impulsionados pelos interesses da saúde pública e da epidemiologia, por um lado, e pela biblioteconomia, por outro; e

- a pesquisa sobre sistemas de suporte de decisão e especialistas para medicina, que se iniciou na década de 1970, impulsionada pelo crescente campo de pesquisas em Inteligência Artificial, inspirada pela idéia de criar ferramentas de computador baseadas em conhecimento para auxiliar no complexo processo de tomada de decisões médicas.

O termo “ontologia” tornou-se um dos termos mais em voga da Ciência Computacional, devido à visão da Semantic Web. Defende-se que as ontologias descrevem áreas com precisão, e empregam essas descrições em muitos tipos de aplicações, do processamento natural da linguagem a sistemas de raciocínio lógico e suporte a decisões. Muitas áreas de aplicação atualmente se beneficiam das ontologias, mas o campo das ciências biológicas está ganhando cada vez mais visibilidade neste cenário, já que muito poucas áreas científicas - se é que há alguma - contêm uma quantidade tão impressionante e rapidamente crescente de termos, conceitos e definições.

## Ontologias

O termo “ontologia” tornou-se muito popular desde os meados dos anos 1990, mas, infelizmente, não havia definições universalmente aceitas (Kuzniersky 2006). Desde o século XVII, o termo tem sido utilizado para denominar a disciplina de metafísica geral, dentro da tradição da “primeira Filosofia” de Aristóteles, como sendo a ciência do ser no papel de ser. É, muitas vezes, encarada como um complemento à idéia de Epistemologia (ciência do conhecimento).

Na Ciência Computacional prevalece a definição de Ontologia como sendo a especificação explícita de uma conceitualização (Gruber 1995). O termo “conceitualização”, aqui, significa uma visão abstrata, simplificada, do mundo que desejamos representar com algum propósito: tirar conclusões, executar classificações automáticas, e assim por diante. Uma conceitualização, geralmente, inclui conceitos (também chamados de classes, ou tipos, como *Coração*), indivíduos como sendo ocorrências de conceitos (por exemplo, o indivíduo Fido é uma ocorrência de *Cachorro*), relações binárias

entre conceitos ou indivíduos (por exemplo, *Cachorro é um vertebrado*), restrições com base lógica (todas as ocorrências de *Herbívoros* ingerem apenas vegetais, enquanto todas as ocorrências de *Carnívoros* ingerem algumas ocorrências de *Animais*), e axiomas (sentenças que sempre são verdadeiras dentro de uma área, como, por exemplo, todas as ocorrências de *Indivíduo Vivo* têm alguma ocorrência de *Coração*). As relações ontológicas são, claramente, o fator aglutinante dessas entidades. Elas representarão diferentes aspectos nos quais os conceitos se relacionam uns com os outros. Os tipos mais relevantes e utilizados de relações são as subclasses (*Coração* é uma subclasse de *Órgão*, uma vez que todas as ocorrências do primeiro são ocorrências do último, com algumas características especiais que o distinguem dos outros), e relações partonômicas (toda ocorrência de *Ventrículo Cardíaco* é parte de um *Coração*). Existem, porém, outras definições de ontologia, como as “representações de uma área de discurso, que consistem de uma lista de termos, as relações entre eles e os axiomas que sempre são válidos na área” (Antoniou & Harmelen 2004), ou um “artefato de representação, cujas unidades de representação devem designar classes ou universalidades da realidade e suas inter-relações” (Smith 2005).

A idéia de ontologia é, freqüentemente, restrita ao que se chama “ontologia formal” (Guarino 1998). Isto significa que o conteúdo de uma ontologia é descrito pela utilização de lógica matemática, que pode dotar os sistemas de computador da habilidade de realizar inferência lógica. Pode, também, apoiar a descoberta autônoma a partir de dados registrados, assim como a reutilização e o intercâmbio de conhecimento.

A ascensão das ontologias na comunidade da Ciência Computacional se expandiu para muitos outros ramos de conhecimento: Motivados pela visão da Semantic Web (Berners-Lee 2001), muitos grupos dos meios acadêmico e industrial do mundo inteiro se interessaram pelas ontologias, e o número de ferramentas, padrões e usuários cresceu na mesma proporção. Realmente, foram feitos alguns esforços no sentido de se produzir ontologias padrão em algumas áreas, especialmente na Medicina e na Biologia.

## Terminologias versus ontologias

A medicina, especialmente, é caracterizada por uma vasta gama das chamadas terminologias, melhor descritas como artefatos lingüísticos que unem os diversos sentidos ou significados das entidades lingüísticas. As terminologias geralmente são construídas com fins bem definidos, como recuperação de documentos, apontamento de recursos, registro de estatísticas de mortalidade e morbidade, ou faturamento de serviços de saúde. As terminologias biomédicas não utilizam descrições formais e bem definidas; elas definem os termos (quando isto ocorre) pelas expressões da linguagem humana, e expressam as associações entre os termos por relações informais, próximas das relações da linguagem humana. Termos de uma ou mais palavras são os blocos fundamen-

tais das terminologias, que geralmente os organizam em hierarquias, que relacionam seus significados em termos de sinonímia (mesmo significado), hiperonímia (significado mais amplo), hiponímia (significado mais restrito). Embora as terminologias possam ser empregadas com êxito na representação de significados abstratos, como, por exemplo, no processamento natural da linguagem ou no apontamento de recursos (resumos literários, resultados experimentais), não são suficientemente precisas e expressivas para aplicações com carga de conhecimento mais intensa.

Enquanto um caso de utilização pode exigir conhecimento sobre *como* e *de que forma* alguns termos diferem entre si, outros podem requerer relações mais precisas entre os termos (por exemplo, que toda ocorrência de *Braço* normal tem uma ocorrência de *Antebraço* como sua parte). Um recurso baseado em linguagem não é suficiente para atender essas exigências. Aqui, um recurso baseado na realidade é mais adequado, de forma a poder capturar as sutilezas de quais entidades (objetos, qualidades, processos, etc.) se relacionam com outras; sob que circunstâncias tais relações ocorrem; e como exatamente essas relações devem ser interpretadas (por exemplo, se a relação parte-de entre uma parte do corpo e um corpo ainda se mantém após a remoção da parte, como um rim, por exemplo). É aqui que entram as ontologias. As ontologias são expressas em formalismos baseados em lógica, que fornecem (meta) definições de classes (conceitos), relações, ocorrências e axiomas. Assim, as ontologias podem representar uma área de uma maneira que os computadores possam manusear as definições de acordo com suas semânticas, ao invés de empregar apenas termos de identificadores semânticos. Desta maneira, um sistema pode verificar se determinada interpretação está correta ou não, se determinada sentença é verdadeira de acordo com determinada ontologia, dentre outras tarefas relacionadas. As ontologias podem, ainda, abranger diferentes dimensões que uma área deve incluir: por exemplo, no caso dos organismos, o grau de conformidade com os padrões de um órgão (se um organismo funciona conforme geralmente deveria ou não), o grau de desenvolvimento (por exemplo, um embrião versus um adulto), o local de um organismo ou matéria orgânica na taxonomia biológica (por exemplo, mosca versus rato), ou a granularidade através da qual a estrutura biológica é descrita (por exemplo, macroscópico versus microscópico), para mencionar alguns (Schulz 2004).

Entretanto, existe uma fusão crescente da abordagem terminológica clássica com os princípios do delineamento da ontologia moderna, com as linguagens ontológicas da área de Ciência Computacional, e com a ascendente disciplina da ontologia aplicada embutida no campo da Filosofia Analítica.

O que temos a intenção de descrever neste estudo é a ampla variedade desses artefatos bastante heterogêneos, para os quais uma definição universal ainda não existe (o termo geralmente utilizado, “vocabulários biomédicos”, é capcioso, pois enfatiza demais o aspecto da

linguagem). No restante deste documento utilizaremos a sigla OTBMs para “ontologias e terminologias biomédicas”. O artigo está organizado da seguinte forma: A próxima seção explica as OTBMs detalhadamente. A Seção 3 é dedicada aos fundamentos e esforços empregados em muitos desses sistemas. A Seção 4 discute alguns tópicos importantes de cada OTBM, enquanto a Seção 5 trata de questões abertas e desafios para a integração das OTBMs.

## Exemplos importantes de terminologias e ontologias (OTBMs)

### Esquema descritivo

Diversas contribuições foram feitas na área biomédica para o desenvolvimento de padrões semânticos, como terminologias médicas, ontologias, e sistemas de codificação. Nesta seção analisaremos um conjunto de OTBMs que reflete a ampla variedade deste gênero. Examinaremos a Classificação Internacional de Doenças (CID), *Medical Subject Headings* (MeSH), Gene Ontology (GO), *Systematized Nomenclature of Medicine - Clinical Terms* (SNOMED CT), *Generalized Architecture for Languages, Encyclopaedias and Nomenclatures* (openGALEN), *Foundational Model of Anatomy* (FMA), e iniciativas de abrangência universal, o *Unified Medical Language System* (UMLS) e o *Open Biomedical Ontologies (OBO) Foundry*. Esses sistemas serão descritos e comparados, através da identificação de diferenças e características em comum; discutiremos o que eles representam, e que arquitetura utilizam. Com esta finalidade, apresentamos os elementos de arquitetura que encontramos em todas as OTBMs, conforme abaixo:

- **Nodes** - Identificadores primários do significado
  - **Links** - Conexões entre os *nodes*
  - **Códigos** - Identificadores alfanuméricos de um *node* ou *link*.
  - **Hierarquias** - Rede de *links* que constituem uma ordem parcial, definindo, assim, diagramas em árvore ou gráficos direcionados
  - **Atributos** - Encarados com uma descrição mais profunda dos *nodes* e *links*
  - **Axiomas** - Sentenças expressas em lógica, sempre verdadeiras dentro da área
- Além disso, descrevemos os sistemas em termos de
- **Finalidade** - Por que foram construídos, e onde foram utilizados
  - **Escopo** - A área de conhecimento que representam
  - **Referência** - O que os *nodes* e *links* denotam

### Classificação Internacional de Doenças

A padronização terminológica da medicina tem um longo histórico. Em 1880 foi criada a Classificação Internacional de Doenças (CID) (OMS 2008), baseada na *London Bills of Mortality*, que distinguia aproximada-

mente 200 causas de morte, e fornecia códigos para todas as doenças conhecidas naquela época. Por muitos anos, o CID foi a única fonte de terminologia médica. Sua atual edição (10ª) é mantida pela Organização Mundial da Saúde (OMS), e está traduzida em 42 idiomas. O CID-10 fornece aproximadamente 13.000 classes para a classificação de doenças e formas de contração. O CID, originalmente criado com fins epidemiológicos, atualmente constitui o sistema de codificação de doenças mais amplamente utilizado, sendo empregado no mundo inteiro como base comum para as estatísticas de saúde. Em muitos países, o CID também é empregado como base para os *Diagnosis Related Groups* (Grupos de Diagnósticos Relacionados - DRG), utilizados em faturamento. Os pacientes clinicamente semelhantes do DRG devem, supostamente, utilizar os mesmos recursos de assistência médica.

O CID tem uma arquitetura simples, porém eficiente. Dividido em 22 capítulos (*Infecções, Neoplasmas, Doenças Sangüíneas, Doenças Endócrinas etc.*), seus *nodes* denotam classes de doenças e problemas relacionados. Isto significa que cada doença específica se encaixa em uma categoria com um código único, por exemplo, a miopia do segundo autor deste documento pode ser codificada como H52.1. As classes do CID são hierarquicamente dispostas em até cinco níveis. A relação de construção hierárquica é a relação *é-uma* (subclasse), que expressa que cada membro de uma classe também é um membro de qualquer classe matriz. O CID axiomáticamente supõe que classes irmãs não se sobrepõem. Isto garante que nenhuma classe tenha mais que uma classe matriz, e que haja exatamente uma classe terminal para a classificação de cada entidade, daí sua caracterização como “classificação”. A simples razão para isto é impedir que uma doença seja contada duas vezes. Com o objetivo de evitar lacunas, foram criadas as categorias residuais (“não classificadas em nenhum outro local”). Atributos adicionais das classes de CID são sentenças de inclusão e exclusão e também, em um capítulo, definições livres em texto, semelhantes a um glossário. As sentenças de inclusão relacionam doenças mais específicas que são contidas na mesma classe, enquanto classes com sentenças de exclusão segregam certas condições de uma classe, designando-as, assim, para uma classe diferente.

O escopo do CID ultrapassa o universo das doenças, pois também inclui lesões e causas extrínsecas de problemas de saúde, sinais e sintomas, e qualquer tipo de condição que justifique uma consulta a um profissional de saúde. O quadro 1 demonstra um trecho do CID relacionado a certos tipos de enfermidades oculares, que são subclasses da categoria de três dígitos H52. Observe a exclusão de dentro da H52. 1, e as inclusões na H52. 5. A primeira deve ser codificada numa ramificação diferente, enquanto a última descreve enfermidades mais específicas para as quais não existe código separado. Observe também que a H52.6 constitui o complemento para a H52.0-H52.5, e que a H52.7 corresponde a H52, e expressa que o codificador

não possui detalhes que permitiriam a utilização de um código mais específico.

<b>H52</b>	<b>Disorders of refraction and accommodation</b>
<b>H52.0</b>	<b>Hypermetropia</b>
<b>H52.1</b>	<b>Myopia</b> <i>Excludes: degenerative myopia ( H44.2 )</i>
<b>H52.2</b>	<b>Astigmatism</b>
<b>H52.3</b>	<b>Anisometropia and aniseikonia</b>
<b>H52.4</b>	<b>Presbyopia</b>
<b>H52.5</b>	<b>Disorders of accommodation</b> <b>Internal ophthalmoplegia (complete)(total)</b> <b>Paresis }</b> <b>Spasm } of accommodation</b>
<b>H52.6</b>	<b>Other disorders of refraction</b>
<b>H52.7</b>	<b>Disorder of refraction, unspecified</b>

**Quadro 1** – Extraído da Classificação Internacional de Doenças, 10ª versão (CID-10).

## Medical Subject Headings (MeSH)

O *Medical Subject Headings* (MeSH) (Nelson 2007, MESH 2008), editado e mantido pela *U.S. National Library of Medicine* (NLM), consiste em um vocabulário controlado, utilizado na indexação de conteúdo de documentos da área de saúde, principalmente resumos literários da base de dados de literatura de ciências biológicas MEDLINE, com mais de 10 milhões de citações (Nelson 2007, PubMed). O MeSH está disponível em 41 idiomas.

O MeSH é dividido, em seu nível mais elevado, em 16 ramificações (*Anatomia, Organismos e Doenças*, dentre outras). Os *nodes* do Mesh são chamados de “cabeçalhos”, e denotam um significado padronizado de um grupo de termos médicos. Em contraste com a hierarquia em forma de árvore do CID, os cabeçalhos do MeSH são dispostos em hierarquias múltiplas. A ordem hierárquica baseia-se no princípio de que todos os documentos indexados por determinado cabeçalho são também relevantes para qualquer descritor matriz. Esses links informais também são caracterizados pelos termos “mais abrangente/mais restrito”. Assim, o cabeçalho MeSH *Leishmaniose* é parte da hierarquia *Doenças Parasitárias*, e também da hierarquia *Doenças da Pele e do Tecido Conjuntivo*, conforme mostrado no quadro 2. Assim, documentos sobre a leishmaniose são encontrados numa busca no MEDLINE por doenças parasitárias, bem como numa busca por doenças de pele. Os cabeçalhos do MeSH têm, além de seu identificador único, um “número de árvore” para cada contexto hierárquico.

Os cabeçalhos são mais detalhadamente definidos por uma definição textual, chamada de nota de escopo. Atributos adicionais são termos de registro (sinônimos ou termos mais específicos) e qualificadores admissíveis, como prevenção, terapia e outros, no caso das doenças, e patogenicidade, no caso de organismos.



<b>MeSH Heading</b>	<b>Leishmaniasis</b>
<b>Tree Number</b>	<b>C03.752.700.500.508</b>
<b>Tree Number</b>	<b>C03.858.560</b>
<b>Tree Number</b>	<b>C17.800.838.775.560</b>
<b>Annotation</b>	<b>protozoan infect; GEN or unspecified; prefer specifics; American leishmaniasis is LEISHMANIASIS, AMERICAN see LEISHMANIASIS, CUTANEOUS; tegumentary leishmaniasis = LEISHMANIASIS, CUTANEOUS</b>
<b>Scope Note</b>	<b>A disease caused by any of a number of species of protozoa in the genus LEISHMANIA. There are four major clinical types of this infection: cutaneous (Old and New World) ( LEISHMANIASIS, CUTANEOUS), diffuse cutaneous ( LEISHMANIASIS, DIFFUSE CUTANEOUS), mucocutaneous ( LEISHMANIASIS, MUCOCUTANEOUS), and visceral ( LEISHMANIASIS, VISCERAL).</b>
<b>Allowable Qualifiers</b>	<b>BL CF CI CL CN CO DH DI DT EC EH EM EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX RA RH RI RT SU TH TM UR US VE VI</b>
<b>Date of Entry</b>	<b>19990101</b>
<b>Unique ID</b>	<b>D007896</b>

<b>Parasitic Diseases [C03]</b>	<b>Skin and Connective Tissue Diseases [C17]</b>
<b>Protozoan Infections [C03.752]</b>	<b>Skin Diseases [C17.800]</b>
<b>Sarcomastigophora Infections [C03.752.700]</b>	<b>Skin Diseases, Infectious [C17.800.838]</b>
<b>Mastigophora Infections [C03.752.700.500]</b>	<b>Skin Diseases, Parasitic [C17.800.838.775]</b>
<b>Leishmaniasis [C03.752.700.500.508]</b>	<b>Leishmaniasis [C17.800.838.775.560]</b>

**Quadro 2** – Registro MeSH para “Leishmaniose”. A tabela fornece definição e atributos. Duas das “árvores” nas quais esse cabeçalho está inserido são mostradas na parte inferior.

## Gene Ontology

O *Gene Ontology* (GO) (GO 2008) é mantido pelo *Gene Ontology Consortium*, que originalmente a criou para dar suporte a apontamentos compartilhados de dados genômicos nas bases de dados de três modelos de organismos (Drosófila, Levedo, Rato). Desde então, seu escopo foi ampliado de forma que atualmente abrange toda a biologia, independentemente das características de organismos específicos. Ao contrário do que o nome indica o GO não é uma ontologia de genes; fornece identificadores semânticos que padronizam a descrição de dados sobre genes ou produtos genéticos (proteínas, por exemplo) em três dimensões: (i) em que compartimento celular o gene é expresso (por exemplo, a mitocôndria); (ii) com que funções uma proteína é associada (por exemplo, sinalização); e (iii) de quais processos biológicos uma proteína participa (por exemplo, mitose). Assim, o GO é capaz de dar suporte a pesquisas nas bases de dados que os membros do consórcio mantêm, facilitando o acesso ao conhecimento descoberto por eles.

Assim como o MeSH, o *Gene Ontology* é dividido em ramificações desarticuladas em seu nível superior. As três ramificações *Componente Celular*, *Processo Biológico*

e *Função Molecular* esboçam seu escopo. Cada ramificação consiste de uma hierarquia múltipla, de um total de 24.500 *nodes*, chamados *termos* de GO. Por mais que a arquitetura do GO possa se assemelhar à do MeSH à primeira vista, há diferenças cruciais que podem justificar sua qualificação como ontologia. Primeiramente, todos os seus *nodes* são mais que descritores semânticos. Ao contrário dos cabeçalhos do MeSH, os termos GO representam classes de entidades reais. Por exemplo, a classe (abstrata) *Núcleo Celular* tem por membros todos os núcleos celulares (materiais) do mundo. Os termos GO são caracterizados por identificadores, os chamados números de inclusão, e têm por atributos adicionais sinônimos e definições. Outra diferença, em comparação ao MeSH, é a clareza semântica dos links. Em vez de “mais abrangente/mais restrito”, o GO fornece duas relações precisamente identificadas: *é-um* e *parte-de*. A primeira significa que toda entidade que é membro de uma classe também é membro de todas as classes matrizes *é-um*, assim como no CID. *Parte-de* deve ser interpretada no sentido de que toda entidade que é membro de uma classe é parte de uma entidade que é membro de todas as suas classes *parte-de*. O quadro 3 apresenta um registro do GO referente à classe *Célula*.

(I) GO:0005623 : cell  
 (P)GO:0044464 : cell part  
 (I) GO:0009334 : 3-phenylpropionate dioxygenase complex  
 (I) GO:0020007 : apical complex  
   (P) GO:0020032 : basal ring of apical complex  
   (P) GO:0020010 : conoid  
   (P) GO:0033289 : intraconoid microtubule  
   (P) GO:0020009 : microneme  
   (P) GO:0070074 : mononeme  
   (P) GO:0020031 : polar ring of apical complex  
   (P) GO:0020008 : rhoptry  
   (P) GO:0020025 : subpellicular microtubule

Cell

#### Term Information

**Accession:** GO:0005623

**Ontology:** cellular component

**Synonyms:** None

**Definition:** The basic structural and functional unit of all organisms. Includes the plasma membrane and any external encapsulating structures such as the cell wall and cell envelope.

[source: GOC:go\_curators]

**Quadro 3** – Registro da classe Célula no Gene Ontology (GO). (I) representa hierarquias é-um, (P) representa hierarquias parte-de.

## SNOMED-CT

O *Systematized Nomenclature of Medicine-Clinical Terms* (SNOMED-CT) (Spackman 2004, IHTSDO 2008) é uma terminologia abrangente, criada para cobrir o registro do paciente por inteiro. Também aborda estruturas corporais, procedimentos e aspectos relevantes relacionados à saúde, incluindo também contexto social. SNOMED CT é o resultado da fusão da versão 3 do *UK Clinical Terms* (também chamado *Read Codes*) e do SNOMED RT (*Reference Terminology*) (Spackman 1997), sendo o último construído a partir de diversas gerações de versões anteriores (Cornet 2008). Desde abril de 2007 o SNOMED CT é de propriedade, mantido e distribuído pela *International Health Terminology Standards Development Organization* (IHTSDO), uma organização sem fins lucrativos baseada na Dinamarca. Os produtos e serviços do SNOMED CT estão abertos para pesquisadores, mas sua utilização para codificação clínica ou outros fins comerciais é restrito aos licenciados (atualmente dez países e algumas empresas). O SNOMED CT está oficialmente disponível em inglês e espanhol, e traduções para outros idiomas (ex. Holandês, Dinamarquês, Sueco) estão sendo feitas.

Do ponto de vista estrutural, o SNOMED CT oferece múltiplas hierarquias *é-um*, contendo mais de 310.000 *nodes*. Alguns dos *nodes* do SNOMED CT, chamados de *conceitos*, denotam, em sua maior parte, classes de entidades individuais (como doenças, procedimentos, resultados laboratoriais, medicamentos etc., mas também particularidades, como entidades geográficas), embora ainda haja certa controvérsia sobre a que se refere, por exemplo, o conceito *Dor no Peito*: se aos objetos em si (por exemplo,

a dor no peito de determinado paciente), ou se à sua menção no registro de saúde (por exemplo, o registro “dor no peito”). Os conceitos do SNOMED CT são exclusivamente identificados por chaves numéricas, juntamente com seus nomes especificados por completo. A maioria dos conceitos SNOMED CT inclui diversos sinônimos (chamados de “descrições”) e, em apenas alguns casos, também definições em texto livre. Atributos adicionais são qualificadores SNOMED, que oferecem refinamentos opcionais para conceitos como, por exemplo: *Lateralidade* para anatomia, ou *Gravidade* para doenças.

O SNOMED CT oferece ainda 50 tipos de link, chamados *conceitos de ligação*. São utilizados no que pode ser considerado o critério distintivo mais importante do SNOMED CT, que é a utilização de uma linguagem rica de representação ontológica, compatível com o padrão Semantic Web OWL-DL (lógica descritiva) (Bechhofer et al. 2004). A lógica descritiva permite a definição de novas classes através da utilização de classes e relações existentes. Conforme mostra o quadro 4, a *Colecistectomia* é inteiramente definida como uma nova classe, utilizando as classes existentes *Extirpação* e *Vesícula Biliar*, juntamente com os links (relações) *Método* e *Local do Procedimento*. Isso significa que cada procedimento de extirpação de uma vesícula biliar é uma colecistectomia, e vice versa.

A criação de expressões complexas baseadas nos conceitos SNOMED que obedece sintaxe e semântica formais é chamada de coordenação. Isto pode ser feito no momento da codificação (pré-coordenação) ou antecipadamente, através da introdução de novos conceitos na terminologia (pós-coordenação) (Chen 2005).

<b>Current Concept:</b>	<b>Fully Specified Name:</b> Cholecystectomy (procedure)
	<b>ConceptId:</b> 38102005
<b>Defining Relationships:</b>	
	<i>Is a</i> Biliary tract excision (procedure)
	<i>Is a</i> Operation on gallbladder (procedure)
<b>Group 1:</b>	
	<i>Method (attribute):</i> Excision - action (qualifier value)
	<i>Procedure site - Direct (attribute):</i> Gallbladder structure (body structure)
	<b>This concept is fully defined.</b>
<b>Qualifiers:</b>	
	<i>Access (attribute):</i> Surgical access values (qualifier value)
	<i>Priority (attribute):</i> Priorities (qualifier value)
<b>Descriptions (Synonyms):</b>	
	<i>Preferred:</i> Cholecystectomy
	<i>Synonyms:</i> Excision of gallbladder, Gallbladder excision, Removal of gallbladder
<b>Parents:</b>	
	Biliary tract excision (procedure)
	Operation on gallbladder (procedure)
<b>Children:</b>	
	Cholecystectomy and exploration of bile duct (procedure)
	Cholecystectomy and operative cholangiogram (procedure)
	Excision of lesion of gallbladder (procedure)
	Laparoscopic cholecystectomy (procedure)
	Partial cholecystectomy (procedure)
	Total cholecystectomy and excision of surrounding tissue (procedure)

Quadro 4 – Definição do SNOMED CT para Colecistectomia. Observe que este conceito é completamente definido, isto é, a combinação de Método – Ação de Extirpação com Local do Procedimento – Estrutura da vesícula biliar.

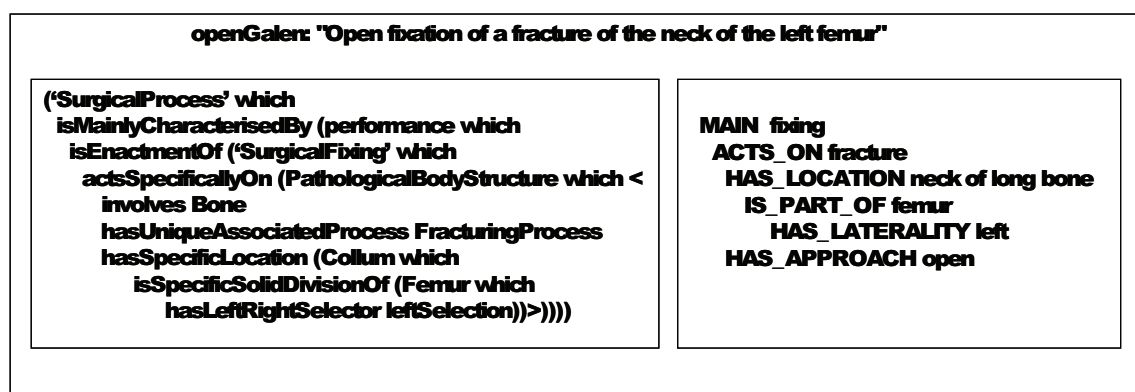
## openGALEN

O *Generalized Architecture for Languages, Encyclopaedias and Nomenclatures* (openGALEN) fornece uma ontologia clínica de fonte aberta que foi desenvolvida nos anos 1990, como resultado de uma série de projetos europeus (GALEN) (Rector 2003). Tem foco nas aplicações clínicas, e contém aproximadamente 25.000 *nodes* (conceitos) e 26 tipos de *link* (relações). Os conceitos openGALEN são também dispostos em múltiplas hierarquias é-um. Utiliza uma linguagem de lógica descritiva chamada GRAIL (GALEN Representation and Integration Language), que permite a definição de classes de forma semelhante à feita pelo SNOMED CT, mas fornece uma sintaxe mais rica, como pode ser visto no exemplo do quadro 5, que descreve a consolidação da fratura do pescoço do fêmur

esquerdo. O modelo GALEN é dividido nos seguintes componentes:

- uma ontologia de alto nível, que fornece uma estrutura geral de categorização;
- o modelo de referência comum (CORE), que contém definições reutilizáveis da anatomia, doenças, procedimentos cirúrgicos, sintomas, etc.;
- extensões detalhadas de sub-domínios específicos, como a cirurgia.

Seu propósito é, assim, semelhante ao do SNOMED CT, mas jamais alcançou seu escopo e granularidade. O openGALEN, no entanto, pode ser considerado pioneiro na utilização da lógica formal nas terminologias biomédicas. Seu exemplo mais importante de utilização foi o desenvolvimento da classificação de procedimentos médicos CCAM (Trombert-Paviot 2000).



Quadro 5 – Registro detalhado do openGALEN, definindo um tipo de consolidação de fratura. Esquerda: Representação do tipo lógica descritiva (sintaxe GRAIL). Direita: sintaxe próxima ao usuário, desenvolvida para facilitar a definição de conceitos de cirurgia.

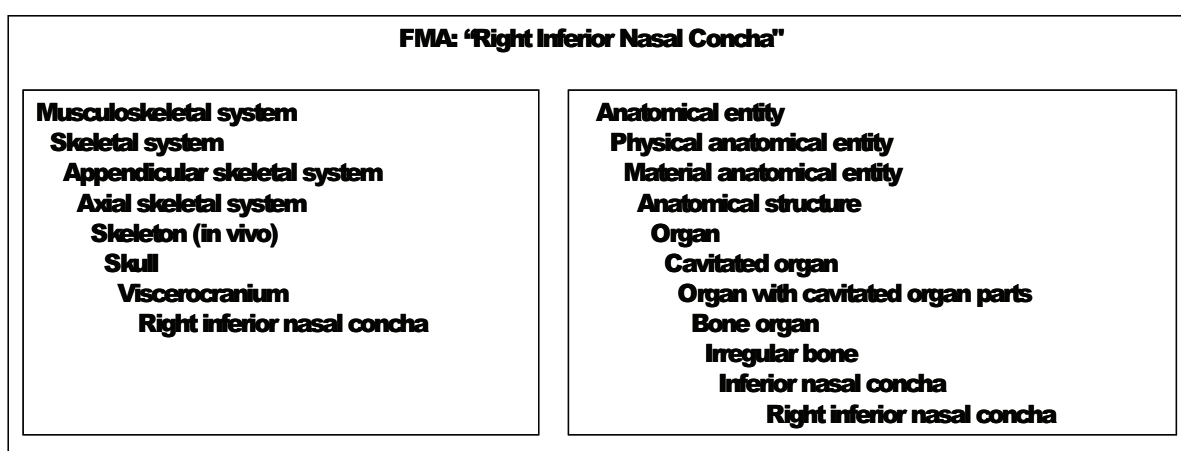
## Foundational Model of Anatomy

O *Foundational Model of Anatomy* (FMA) (FMA 2008) é uma ontologia biomédica que fornece conhecimento declaratório sobre a estrutura microscópica do corpo humano. Foi originalmente desenvolvido para descrever imagens anatômicas para fins didáticos. Assim como o GO, os *nodes* são dispostos em duas hierarquias, a *Anatomy Taxonomy*, que é uma monohierarquia *é-um*, e a multihierárquica *Part-Whole Network*, que emprega *parte-de* como uma relação de hierarquização. Atributos adicionais são identificadores, sinônimos, e relações adicionais (por exemplo, *tem-dimensão*, *tem-massa*, *adjacente\_a* etc.). O FMA é representado no formalismo de estrutura, que faz suposições ontológicas menos rígidas e, portanto, só pode ser traduzido de forma incompleta para a Lógica Descritiva.

Os *nodes* no FMA são denominados de *classes* ou *tipos*, o que ampara seu comprometimento com entida-

des do mundo real, ao invés de com os significados de termos. Entretanto, o FMA explicitamente declara que suas classes abrangem entidades anatômicas *padrão*, como num atlas anatômico, o que resulta na descrição de um corpo humano ideal, sem nenhuma deficiência, alteração anatômica ou malformação. Isto causa, algumas vezes, inconsistências, como aquela com o axioma da FMA, que declara que “*O trato gastrointestinal inferior tem-parte Apêndice*”. Há, claramente, um conflito com situações clínicas frequentes.

O quadro 6 mostra a classe *Concha Nasal Inferior Direita*, e declara que ela é parte do *Crânio*, que é, por sua vez, parte do *Esqueleto*, e assim por diante. Outro registro a define como um subtipo de *Concha Nasal Inferior*, que é um *Órgão Ósseo*, que é um subtipo de muitas outras classes, incluindo a classe mais geral, *Entidade Anatômica*.



Quadro 6 – Definição do *Foundational Model of Anatomy* para Concha Nasal Inferior Direita.

## Esforços para reunir diferentes fontes de conhecimento biomédico

### Fundamentos

Esforços consideráveis foram investidos, por um lado, no alinhamento das terminologias e ontologias biomédicas, - numerosas e, em grande parte, sobrepostas – e, por outro, na prevenção da proliferação desordenada das OTBMs, através do estabelecimento de princípios para o desenvolvimento coordenado de recursos interoperáveis. Descreveremos o *Unified Medical Language System* (UMLS) e o OBO (*Open Biological Ontologies Foundry*). Enquanto o UMLS é um exemplo da primeira estratégia, o OBO adota a segunda abordagem.

### Metatesauro do Unified Medical Language System UMLS

O Metatesauro do *Unified Medical Language System* (UMLS) constitui a mais rica fonte de terminologias, te-

sauros, sistemas de classificação, e ontologias biomédicas (Nelson 2006, UMLS 2008). Foi criado em 1986 pela *U.S. National Library of Medicine* (NLM), com o propósito de integrar informações de diversas fontes terminológicas incompatíveis. O UMLS atualmente cobre 2 milhões de nomes para aproximadamente 1 milhão de conceitos biomédicos para mais de 120 OTBMs, bem como 12 milhões de relações entre esses conceitos (Bodenreider 2004). Exceto pelo openGalen, todos os sistemas mencionados acima estão incluídos no Metatesauro UMLS, assim como muitos outros, cobrindo organismos, medicamentos, substâncias químicas, dispositivos, procedimentos, etc.

Além de facilitar o acesso transparente às fontes (através do fornecimento de arquivos não-processados e serviços *online*), a principal realização do Metatesauro UMLS baseia-se, essencialmente, em dois recursos:

- cada *node* da fonte OTBM é mapeado em retrospecto em um conceito de Metatesauro, cada um com seu identificador único, denominado CUI (*Concept Unique Identifier*). Tais mapeamentos são periodicamente atuali-



zados manualmente. Permitem que seja feita uma ponte entre OTBMs de diferentes fontes. Conseqüentemente, os *links* entre *nodes* das fontes são mapeados para *links* entre CUIs, denominados relações semânticas. Os aplicativos que os utilizam podem, assim, beneficiar-se das ligações entre conceitos de ambos os lados;

- cada conceito do Metatesouro é categorizado por, no mínimo, um tipo semântico da *UMLS Semantic Network*, um conceito global de toda a área biomédica (McCray 2003). Uma árvore de 135 tipos semânticos, ligados por relações *é-um*, forma o suporte principal desta *Semantic Network*. Além disso, a rede inclui uma hierarquia de 53 relações associativas (por exemplo, *localização-de*, *trata*), que são utilizadas para formar 612 trios (por exemplo, *Tecido*, *Procedimento de Diagnóstico*, etc.), dos quais 6.252 trios adicionais podem ser inferidos. Esses trios são interpretados como restrições área/abrangência das relações.

## Open Biomedical Ontologies (OBO) Foundry

Criada em 2003, a plataforma OBO, *Open Biomedical Ontologies* (OBO 2008), evoluiu como uma biblioteca de ontologias biomédicas online, de domínio público. A partir disso, a iniciativa *OBO Foundry* desenvolveu um conjunto de princípios compartilhados que regulam o desenvolvimento de ontologias biomédicas (Smith 2007). A cobertura da *OBO Foundry* compreende diversas ontologias anatômicas (incluindo o FMA), o *Gene Ontology*, bem como ontologias especializadas de bioquímica (ChEBI), fenótipos (PATO), seqüências (SO), e técnicas de investigação (OBI). Atualmente, mais de 50 ontologias estão na lista de candidatas à *OBO Foundry*.

A *OBO Foundry* dissemina duas linguagens representativas. Além do OWL-DL, há um formato patenteado (OBO-EDIT 2009), onde a maior parte das ontologias OBO está codificada.

Assim como na *Gene Ontology*, os *nodes* das ontologias OBO denotam classes de entidades do mundo real. Os *links* entre essas classes são interpretados como *links* quantificados existencialmente. Por exemplo, *A parte de B* significa que toda ocorrência de *A* é parte de alguma ocorrência de *B* (mas não vice-versa). As principais relações da OBO (*é-um*, *parte-de*, *parte-integral-de*, *parte-específica-de*, *localizado-em*, *contido-em*, *adjacente-a*, *transformação-de*, *deriva-de*, *precedido-por*, *tem-participante*, *tem-agente*, *ocorrência-de*) foram dotadas de definições formais consistentes e inequívocas.

## Discussão

Descrevemos uma amostra de OTBMs, que parcialmente representa a variedade de padrões semânticos da Biologia e Medicina. Nosso propósito foi dar aos leitores uma visão geral dos significativos esforços que vêm sendo feitos para descrever termos e as entidades que denotam, de forma a dar apoio a buscas e ao processamento inteligente de dados e conhecimento, em aplicações gerais e específicas. Além disso, apresentamos tais esforços em ordem crescente de expressividade. Dois aspectos diretamente ligados à expressividade são escalonamento e cobertura, uma vez que OTBMs codificadas em forma-

lismos expressivos devem ser empregadas em áreas mais restritas, enquanto que tal restrição não é relevante no caso das terminologias informais.

Embora teoricamente pareça simples distinguirmos terminologias das ontologias formais, na prática a distinção é menos clara. A idéia central é que as terminologias são muito mais relacionadas à organização de termos das áreas (já que uma enorme quantidade de termos forma a base de qualquer subárea da Biomedicina) – enquanto as ontologias dão uma descrição mais precisa, baseada em lógica formal, e tão independente quanto possível da linguagem humana. Um exemplo típico disto é SNOMED CT. Seus predecessores têm raízes em uma nomenclatura composicional padronizada (SNOMED Int.), e em um sistema de codificação clínica (*NHS Clinical Terms*, versão 3), mas sua atual reestruturação está sendo cada vez mais guiada por princípios ontológicos. Por outro lado, OTBMs como o CID e MeSH podem ser considerados mais estabelecidos, já que casos importantes e globalmente bem-sucedidos existem há décadas. O CID tem, ainda, um histórico mais antigo e uma disseminação maior, devido à sua arquitetura simples e à necessidade precoce de estatísticas de saúde ou doença. Endossado pela OMS e por entidades nacionais, seus objetivos tem incluído cada vez mais a epidemiologia clínica, a administração da saúde, a garantia de qualidade e o faturamento em diversos países, incluindo o Brasil. O MeSH, por outro lado, tem uma estrutura multi-hierárquica complexa, especificamente projetada para buscas dentro de coleções de textos biomédicos.

Pode-se observar uma tendência clara, que é a adoção crescente das linguagens e formalismos da *Semantic Web*, especialmente a linguagem ontológica OWL e seu subgrupo OWL-DL, sendo a última adaptada às necessidades do raciocínio eletrônico. As principais vantagens de se utilizar maquinário de inferência como aquele disponível para a lógica descritiva são poder verificar os vínculos dos axiomas contidos na ontologia, dar suporte a buscas que demandam conhecimento, calcular as equivalências semânticas de expressões semanticamente diferentes, e desambiguar as expressões da linguagem natural. Embora os classificadores atualmente disponíveis enfrentem problemas de escalabilidade com formalismos mais expressivos (e, desta forma, mais interessantes), o fato de que padrões como a lógica descritiva e o OWL existem compensa as aplicações que exigem conhecimento profundo de um pequeno número de sub-campos. Como pôde ser visto na seção anterior, muitas das OTBMs apresentadas evidaram esforços no sentido de mudar de seu formato original para a lógica descritiva. O SNOMED era uma terminologia pura, no passado; o FMA já mudou parcialmente de *frames* para OWL, e há uma tendência de que as ontologias OBO adotem OWL-DL, embora um formato patenteado tenha sido desenvolvido no passado, e ainda seja utilizado em grande escala. Curiosamente o openGALEN foi concebido, desde o início, para utilizar uma linguagem baseada em lógica, semelhante a DL. Assim, ele pode se orgulhar de ter sido o primeiro a axiomatizar uma quantidade significativa de termos médicos, e as lições aprendidas são de grande valor para a engenharia da ontologia biomédica até hoje.

A enorme quantidade de OTBMs que descrevem áreas parcialmente sobrepostas para casos de utilização semelhantes ou diferentes baseados em formalismos, filosofias e suposições (tácitas) diferentes foi identificada como sendo um problema já nos anos 1980. Desde então, grandes esforços foram investidos no Metatesouro UMLS, através do qual um número cada vez maior de fontes heterogêneas é anualmente intermapeado e categorizado. Devemos, entretanto, chamar a atenção para duas restrições. Primeiro, o mapeamento não pode ser mais expressivo que a OTBM de fonte menos expressiva, e, segundo, a serventia do UMLS para aplicações práticas

é obstruída pelo fato de que muitas das suas fontes estão sujeitas a licenciamento individual.

Em contrapartida, as fontes OBO são completamente de domínio público, e podem ser acessadas por todos. Isto, ao menos parcialmente, explica seu sucesso e o alto nível de conhecimento biológico sendo investido em sua construção e manutenção.

O quadro 7 resume algumas características principais das OTBMs descritas e dos esforços nesse sentido, demonstrando seu escopo, cobertura, volume, formalismo e utilizações.

Nome	Escopo	Formalismo	Número de Nodes	Aplicação	URL
CID	Doenças	Classificação, estritamente é um	Aproximadamente 13.000 classes	Saúde, Estatística, Epidemiologia Relatórios da Saúde Faturamento	<a href="http://www.who.int/classifications/apps/icd/">www.who.int/classifications/apps/icd/</a>
MESH	Medicina, Enfermagem, Odontologia Medicina Veterinária, Sistemas de Assistência Médica Ciências pré-clínicas	Terminology Semantic Networks (Redes de Semântica da Terminologia)	24,767 (2008) termos	Indexação, artigos de 4.800 das principais publicações biomédicas do mundo para a base de dados MEDLINE/PubMED®	<a href="http://www.pubmed.gov">www.pubmed.gov</a>
SNOMED	Tudo codificado no registro eletrônico de saúde	Lógica Descritiva	311.000 conceitos (2008)	Informação sobre o histórico médico de um paciente, doenças, e resultados laboratoriais.	<a href="http://www.ihtsdo.org">www.ihtsdo.org</a>
GO	Componentes celulares, funções moleculares, processos biológicos	OBO/OWL	24,500 termos (2008)	Pesquisa de genes, proteínas	<a href="http://www.geneontology.org">www.geneontology.org</a>
GALEN	Anatomia, ações cirúrgicas, doenças, assistência médica	Linguagem do tipo Lógica Descritiva GRAIL	Mais de 10.000	Registros eletrônicos de assistência médica, interfaces de usuário, sistemas de suporte a decisão, sistemas de acesso ao conhecimento, processamento de linguagem natural	<a href="http://www.opengalen.org">www.opengalen.org</a>
FMA	Conteúdo anatômico	Frames e (parcialmente) OWL	75.000 classes	Educação, pesquisa biomédica	<a href="http://sig.biostr.washington.edu/projects/fm/AboutFM.html">http://sig.biostr.washington.edu/projects/fm/AboutFM.html</a>
OBO	Bioinformática e Biologia Molecular	OBO/ OWL / OBO_XML / RDF	60 ontologias	Utilizado como repositório e esquema unificado para interoperar projetos biomédicos	<a href="http://www.obofoundry.org">www.obofoundry.org</a>
UMLS	Conceitos biomédicos e relacionados à saúde	Semantic Networks	Mais de 1 milhão de conceitos	Literatura científica, diretrizes, e dados de saúde pública, processamento de linguagem natural	<a href="http://www.nlm.nih.gov/research/umls/">http://www.nlm.nih.gov/research/umls/</a>

**Quadro 7 – OTBMs, OBO, UMLS e algumas de suas características principais.**

## Desafios e questões em aberto

A informática biomédica está entrando em uma nova era. Além dos algoritmos empregados na pesquisa genética, as ontologias são consideradas, cada vez mais, o assunto do momento. Já existe uma comunidade ativa pesquisando e se beneficiando da interoperabilidade semântica através das ontologias, uma vez que estas são cada vez mais utilizadas para o apontamento de dados de pesquisas sobre Biologia Molecular e Genômica. Os vocabulários reutilizáveis emergentes demonstram ser úteis na descrição de dados biomédicos de um número cada vez maior de tipos de aplicação. A captura precisa de conhecimento biológico num meio computacional permite a criação de sistemas capazes de cumprir exigências severas, como as de biólogos, pesquisadores da área médica, e médicos: fácil acesso a textos e bases de dados que contêm dados, informação e sentenças detalhados; raciocínio estável e completo; rápido desenvolvimento de sistemas de suporte de decisão para diversos tipos de utilização, etc. Entretanto, alguns desafios têm de ser superados para que o campo atinja sua maturidade.

O primeiro é relacionado à modelagem. Os aspectos sutis que têm de ser descritos em ontologias biomédicas geralmente exigem a utilização de ontologias e técnicas de avaliação de ontologias de primeira linha (Guarino 2000). Do contrário, o raciocínio resultante pode ser falho. Um exemplo simbólico pode ser percebido nas relações entre as classes principais Objeto físico e Quantidade de matéria. A famosa ontologia WordNet (Miller 1995), utilizada pelos pesquisadores da informática, especialmente da área de Processamento de Linguagem Natural, diz que *Objeto Físico é uma Quantidade de Matéria*. Por outro lado, a *Pangloss*, uma extensa ontologia utilizada principalmente para tradução entre linguagens, descreve duas classes de forma contrária, sendo a *Quantidade de Matéria* uma superclasse de *Objeto Físico*. De fato, (Guarino & Welty 2000) afirmam que ambas as interpretações estão erradas: Toda ocorrência de *Objeto Físico* é constituída de uma ou mais ocorrências de *Quantidade de Matéria*. Não existe, entretanto, relação de superclasse, o que pode ser facilmente percebido pela análise de meta-propriedades, como unidade, rigidez, ou identidade. Essa inexactidão também ocorre na área Biomédica: uma versão anterior do *Gene Ontology* incluía o axioma *Célula tem-parte Axônio*. Num exame mais próximo, esta definição levou a ambigüidades e especificações deficientes, uma vez que há células sem axônios, e axônios sem células no mínimo desempenham funções laboratoriais (Schulz 2004). Esses dois exemplos enfatizam a necessidade de existir maior formalidade e riqueza semântica nas ontologias biomédicas.

Outra questão fundamental, que também pode ser percebida no primeiro exemplo, é a da integração. Conforme cresce o número de ontologias biomédicas, muitas aplicações precisam empregar mais de uma ontologia, o que leva a uma série de conseqüências significativas. Inegavelmente, este não é um problema apenas da Biomedicina; os principais obstáculos para a reutilização

de conhecimento na Ciência Computacional vêm da heterogeneidade do conhecimento. O conhecimento é, naturalmente, diverso em suas muitas características: forma, expressão, formalismos da representação, linguagem, sintaxe, conteúdo, significado, princípios de modelagem, práticas e padrões, pontos de vista, perspectivas, utilização, granularidade, terminologia, premissas; isto para não mencionar que as fusões de alguns deles podem ser de difícil raciocínio, do ponto de vista dos recursos computacionais. Embora as ontologias (no sentido mais estrito, ou seja, sentenças a respeito do que é sempre verdadeiro e inequivocamente aceito) só cubram um segmento bem definido do que é normalmente compreendido como representação de conhecimento, essas variedades sempre terão impacto sobre decisões cruciais a respeito de estruturação, e gerarão questões sutis para as aplicações ontológicas. Lidar com a heterogeneidade tornou-se um problema recorrente e desafiador da pesquisa no campo da ontologia, e, por outro lado, também uma boa fonte de utilização ontológica; por exemplo, em problemas como a integração de informações de ontologias heterogêneas, como, por exemplo, em buscas por hotéis, cuja descrição é feita de forma diferente em cada um dos muitos sistemas.

Granularidade é um problema específico que também tem grande impacto sobre a integração de ontologias biomédicas (Schulz 2009). Há esperança de que as pesquisas médicas e biológicas unam as ontologias nos níveis celular, anatômico, medicamentoso, etc. Tais comunidades podem necessitar de granularidades diferentes, ou mesmo de visões diferentes da mesma ontologia. Outro desafio relacionado à integração é como lidar com ontologias biomédicas já existentes que contêm informações sobrepostas, e oferecem pontos de vista diferentes a respeito de certa subárea, ou abrangem diferentes áreas.

Várias pesquisas estão sendo feitas no sentido de possibilitar a integração de ontologias. Há um breve resumo da descrição dessas pesquisas em (Freitas et al. 2007), e uma cobertura mais profunda em (Stuckenschmidt et al. 2000).

O processamento de texto é certamente uma das principais aplicações reais das ontologias biomédicas. Um caso muito popular é a designação automática de termos MeSH para as consultas de usuários no PubMed. Outro é a extração automatizada de informações relacionadas a genes individuais ou proteínas dos textos científicos. O registro eletrônico de saúde e a plataforma do consumidor também constituem um vasto campo para o processamento de texto e conhecimento. Para lidar com esse assunto, os sistemas podem basear-se em sistemas de extração de informações e mineração de texto (Muslea 1999, Ananiadou 2006). Muitas questões, no entanto, permanecem sem resposta, e a combinação de metodologias de análise de texto de alta qualidade com ontologias altamente expressivas e bem padronizadas constitui um desafio permanente para a pesquisa.

## Referências bibliográficas

- Ananiadou S, McNaught J. Text Mining for Biology and Biomedicine, chapter Introduction. Norwood, MA: Artech House Publishers; 2006.
- Antoniou G, van Harmelen F. A Semantic Web Primer. MIT Press, Cambridge; 2004.
- Bechhofer S, Harmelen F, Hendler J, Horrocks I. OWL Web Ontology Language Reference. W3C Recommendation; 2004 . <http://www.w3.org/TR/2003/PR-owl-ref-20031215/>. Last accessed February 3, 2009.
- Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology, Oxford University. 2004 January 1; 32(1) Suppl.1: D267-D270.
- Chen H, Fuller SS, Friedman C, Hersh W. Knowledge Management and Data Mining in Biomedicine Series: Integrated Series in Information Systems , New York: Springer; 2005. Vol. 8.
- Cornet R. and de Keizer N. Forty years of SNOMED: a literature review. BMC Medical Informatics and Decision Making. 2008; 8(Suppl 1): S2.
- FMA - Foundational Model of Anatomy [sig.biostr.washington.edu/projects/fm](http://sig.biostr.washington.edu/projects/fm) Accessed in April 2008. Berners-Lee T, Hendler J, Lassila O, editors. The Semantic Web, Scientific American. 2001; 28-37.
- Freitas F, Stuckenschmidt H, Noy N. Ontology Issues and Applications: Guest Editors' Introduction. Journal of the Brazilian Computer Society. 2005; 11(2).
- GO - The Gene Ontology <http://amigo.geneontology.org/cgi-bin/amigo/go.cgi>. Last accessed February 3, 2009.
- Gruber T. A translation approach to portable ontologies. Knowledge Acquisition. 1995; 5(2):199-220.
- Guarino N. Formal ontology in information systems. Proc FOIS'98. 1998; 3-15.
- Guarino N, Welty C. A formal ontology of properties. In: Knowledge Engineering and Knowledge Management - Proceedings of 12th International Conference EKAW 2000. France: Springer; 2000.
- IHTSDO - International Healthcare Terminology Standards Development Organisation. <http://www.ihtsdo.de>. Last accessed February 3, 2009.
- Kunierczyk W. Nontological Engineering. Formal Ontology In Information Systems. In: Proceedings of the 4th International Conference FOIS 2006, Amsterdam, The Netherlands: IOS Press; 2006. 39-50.
- MESH - Medical Subject Headings, <http://www.nlm.nih.gov/mesh/>. Last accessed February 3, 2009.
- Miller G. WordNet: a lexical database for English. Communications of the ACM; 1995.
- Muslea I. Extraction patterns for information extraction tasks: A survey. American Association for Artificial Intelligence ([www.aaai.org](http://www.aaai.org)) he AAAI-99 Workshop on Machine Learning for Information (1999).
- Nelson SJ, Powell T, Humphreys LB. The Unified Medical Language System (UMLS) of the National Library of Medicine. Journal of American Medical Record Association. 2006; 61: 40-42.
- Nelson SJ, Schulman J. A Multilingual Vocabulary Project - Managing the Maintenance Environment. MeSH Section, National Library of Medicine, Bethesda, Maryland; 2007.
- OBO - Open Biomedical Ontologies. <http://www.obo-foundry.org>. Last accessed February 3, 2009.
- OBO-EDIT. An Introduction to OBO Ontologies [http://oboedit.org/docs/html/An\\_Introduction\\_to\\_OBO\\_Ontologies.htm](http://oboedit.org/docs/html/An_Introduction_to_OBO_Ontologies.htm). Last accessed February 3, 2009.
- OpenGalen Foundation. <http://www.opengalen.org>. Last accessed February 3, 2009.
- PubMed. <http://www.ncbi.nlm.nih.gov/pubmed/>. National Library of Medicine. Last accessed February 3, 2009.
- Rector A, Rogers JE, Zanstra PE, Haring E. OpenGALEN: Open Source Medical Terminology and Tools. AMIA Annual Symposium Proceedings. 2003; 982.
- Rector A. Clinical Terminology: Why is it so hard? Methods of Information in Medicine. 2000; 38(4): 239-52.
- Rubin DL, Shah NH, Noy N. Biomedical Ontologies: a functional perspective. Briefing in Bioinformatics. 2008 Jan; 9(1): 75-90.
- Schulz S, Hahn U. Mereotopological Reasoning about Parts and (W) holes in Bio-Ontologies, In: C. Welty and B. Smith, editors, Formal Ontology in Information Systems. Collected Papers from the 2nd International FOIS Conference, New York, NY: ACM Press, 2001; 210-21.
- Schulz S, Hahn U. Towards the ontological foundations of symbolic biological theories. Artificial Intelligence in Medicine. 2007 Mar; 39(3): 237-50.
- Schulz S, Boeker M, Stenzhorn H, Niggemann J. Granularity Issues in the Alignment of Upper Ontologies. Methods of Information in Medicine. 2009. Accepted for Publication.
- Smith B, Ashburner M, Rosse C, Bard C, Bug W, Ceusters W, Goldberg L J, Eilbeck K, Ireland A, Mungall CJ, The OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone S-A, Scheuermann R H, Shah N, Whetzel PL and Lewis S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, Nature Biotechnology. 2007; 25: 1251-5.
- Smith B, Mejino JLV, Schulz S, Rosse C. Anatomical Information Science. In: COSIT 2005: Spatial Information Theory. Foundations of Geographic Information Science, New York: Springer. 2005; 149-64



Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C., Relations in Biomedical Ontologies. *Genome Biology*. 2005; 6(5).

Spackman KA, Campbell KE, Côté RA. SNOMED RT: A reference terminology for health care. In Masys DR (Ed.), *The Emergence of Internetable Health Care: Systems that Really Work*. Proceedings of the 1997 AMIA Annual Symposium, 640-644. Philadelphia: Hanley & Belfus, Inc. 1997 .

Spackman KA. SNOMED CT milestones: endorsements are added to already-impressive standards credentials. *Healthcare Informatics*. 2004; 21: 54-6.

Stuckenschmidt H, Wache H, Vogele T, Visser U. Enabling technologies for interoperability. In Visser, U. and Pundt, H. editors, *Workshop on the 14th International*

*Symposium of Computer Science for Environmental Protection*, Bonn, Germany. TZI, University of Bremen. 2000; 35-46.

Trombert-Paviot B, Rodrigues JM, Rogers JE, Baud R, van der Haring E, Rassinoux AM, Abrial V, Clavel L, Idir H. GALEN: a third generation terminology tool to support a multipurpose national coding system for surgical procedures. *Intern J Med Informatics*. 2000 Sep; 58-59: 71-85.

UMLS - Unified Medical Language System <http://www.nlm.nih.gov/research/umls/>. Last accessed February 3, 2009.

WHO - International Classification of Diseases, 10th Edition. World Health Organization. <http://www.who.int/classifications/apps/icd/icd10online/> . Last accessed February 3, 2009.



## Sobre os autores

### *Fred Freitas*

É PhD pela Universidade de Santa Catarina, Brasil, e atualmente é afiliado ao Centro de Informática da Universidade Federal de Pernambuco, Brasil (CIn/UFPE). Conduziu pesquisas por quase um ano no Departamento de Informática da Universidade de Karlsruhe, como integrante do projeto Brasil-Alemanha “*A semantic approach to data retrieval*” (Abordagem semântica da recuperação de dados). Publicou diversos artigos em conferências e seminários de renome, como IJCAI e outros patrocinados pela ACM (*Association on Computer Machinery*) e pelo IEEE (*Institute of Electrical and Electronical Engineering*). Co-presidiu duas séries de seminários: O WONTO (*Workshop on Ontologies and their Applications/Seminário de Ontologias e Suas Aplicações*), no Brasil, e o BA-OSW (*Building Applications with Ontologies for the Semantic Web/Construção de Aplicações com Ontologias para a Semantic Web*), em Portugal. Co-editou Edições Especiais sobre temas relacionados do JBCS (*Journal of Brazilian Computer Society*) e do JUCS (*Journal of Universal Computer Science*). Colabora, atualmente, com a Universidade de Paul Cessane em Marselha, e INRIA, Montbonnot, na França, e as Universidades de Karlsruhe, Freiburg e Mannheim, na Alemanha. Suas áreas de interesse incluem ontologias, sistemas multiagentes, representação de conhecimento, mediação, e mineração de texto.

### *Stefan Schulz*

É formado em medicina pela Heidelberg University, Alemanha, e é pesquisador sênior e professor do Instituto de Biometria Médica e Informática da Medicina do Centro Médico Universitário Freiburg, onde chefia o Grupo de Pesquisas em Informática na Medicina. Seu trabalho se concentra em terminologias e ontologias biomédicas, representação do conhecimento biomédico, recuperação de documentos médicos multilíngües, mineração de texto e dados em repositórios de documentos clínicos, aprendizado eletrônico na Medicina, e informática da saúde em países em desenvolvimento.

Após executar trabalhos clínicos em cirurgia e medicina interna, obteve seu diploma de doutorado na área da higiene tropical, onde efetuou um estudo de campo parasitológico em São Luís, Brasil. Após obter qualificação técnica em computação médica, mudou-se para a Universidade de Freiburg, onde participou de projetos de desenvolvimento de software clínico e educacional, e de diversos projetos de pesquisa na área da extração de informações, terminologias biomédicas, engenharia da linguagem médica, e tecnologias semânticas. Tem desempenhado papéis de liderança em diversos projetos financiados pela União Européia. Stefan Schulz é autor de mais de cem publicações revisadas por especialistas, e recebeu vários prêmios. Tem oferecido repetidas contribuições a projetos de pesquisa na área da informática de saúde brasileira desde 2001, como pesquisador convidado da Pontifícia Universidade Católica do Paraná (PUC-PR).



**RECIIS**

Revista Eletrônica de Comunicação  
Informação & Inovação em Saúde

[[www.reciis.cict.fiocruz.br](http://www.reciis.cict.fiocruz.br)]

ISSN 1981-6278

**Artigos originais**

# **Bases ontológicas e conceituais para um modelo do conhecimento científico em artigos biomédicos**

DOI: 10.3395/receis.v3i1.240pt



**Carlos Henrique Marcondes**

Departamento de Ciência da Informação, Universidade Federal Fluminense, Niterói, Brasil  
[marcon@vm.uff.br](mailto:marcon@vm.uff.br)



**Marília Alvarenga Rocha Mendonça**

Departamento de Ciência da Informação, Universidade Federal Fluminense, Niterói, Brasil  
[marilaalvarenga@terra.com.br](mailto:marilaalvarenga@terra.com.br)

**Luciana Reis Malheiros**

Departamento de Fisiologia e Farmacologia, Universidade Federal Fluminense, Niterói, Brasil  
[malheiro@vm.uff.br](mailto:malheiro@vm.uff.br)

**Leonardo Cruz da Costa**

Departamento de Computação, Universidade Federal Fluminense, Niterói, Brasil  
[leo@dcc.ic.uff.br](mailto:leo@dcc.ic.uff.br)

**Tatiana Cristina Paredes Santos**

Instituto Biomédico, Universidade Federal Fluminense, Niterói, Brasil  
[tatianacps@biof.ufrj.br](mailto:tatianacps@biof.ufrj.br)

## **Resumo**

Artigos científicos publicados em formato digital se constituem em bases de conhecimento científico, em especial na Medicina. Um obstáculo para o processamento semântico desse conhecimento por computadores é o fato de que, apesar do formato digital, estas bases de conhecimento são voltadas para leitura e processamento do seu conteúdo por pessoas. Propõe-se um modelo de publicação e registro para representar o conhecimento contido em artigos científicos em Medicina em formato “inteligível” por programas. Segundo o modelo, artigos científicos seriam publicados não só em formato textual, legível por pessoas, mas também como ontologias, representando o conhecimento específico contido em cada artigo. O modelo inicial foi obtido a partir de aportes teóricos de Metodologia Científica e Filosofia da Ciência e da análise de 75 artigos científicos em Medicina. O conteúdo de conhecimento científico de um artigo é associado à proposições que caracterizam fenômenos ou que estabelecem relações entre fenômenos. O modelo permite que programas “agentes de software” processem o conteúdo de cada artigo, viabilizando a recuperação semântica de informações, a avaliação da coerência, a identificação de lacunas no conhecimento científico e de novas descobertas.

## **Palavras-chave**

conhecimento médico; representação do conhecimento; ontologias; publicações eletrônicas; comunicação científica

Desde que foi ultrapassado o período de cultura oral com a invenção da escrita, por muitos séculos o conhecimento humano, em especial, o conhecimento científico, esteve registrado em documentos.

Estamos no limiar de uma transformação profunda nos meios de que dispõe a humanidade para registro, guarda e disseminação do conhecimento. Hoje não só podemos registrar este conhecimento em meio digital, armazenado nos dispositivos de memórias de massa, como também podemos disseminá-lo em larga escala através de redes de computadores. Um dispositivo como um *pen drive* consegue armazenar mais de 4 GB de dados, permitindo ter uma biblioteca gigantesca carregada no bolso. Mais significativo no entanto do que estas questões, é que esse conhecimento não é mais somente codificado em formato textual, legível por pessoas, não somente em formato legível por programas, que por sua vez o tornam legível por pessoas, como no caso de documentos textuais em formato WORD ou PDF, mas, o que se constitui numa real novidade, também em formatos “inteligíveis” por programas, permitindo a estes graus crescente de capacidade de realizar “inferência”, “decisões” e “raciocínios” sobre o conteúdo desses documentos. Esta é a proposta do projeto Web Semântica (Berners-Lee 2001).

Uma das bases do projeto Web Semântica são as ontologias. Uma ontologia é um modelo informacional descrevendo e representando um domínio de conhecimento específico, através dos conceitos correspondendo aos objetos relevantes nesse domínio, de sua estrutura e seus inter-relacionamentos; esse modelo deve ser de entendimento compartilhado por uma comunidade de usuários. Os conceitos são organizados em hierarquias de classes em níveis crescentes de generalidade e possuem atributos e relações entre si. Uma ontologia é representada em linguagem “inteligível” por programas “agentes de software”, e usada por estes para fazer inferências sobre os conceitos desse domínio. Quando às classes de um domínio de conhecimento específico que constituem uma ontologia são agregados representações de objetos individuais desse domínio, têm-se uma base de conhecimento. Existem muitas comunidades de usuários desenvolvendo ou utilizando ontologias, em especial na área Biomédica.

A passagem do registro de conhecimentos do formato textual para formatos suscetíveis de serem

“inteligíveis” por programas pressupõe novas formas de encará-lo. Questões óbvias e imediatas que se colocam são: o quê é o conhecimento? de quê é constituído? como pode ser sistematizado? como pode ser registrado num formato alternativa ao textual? Embora a maioria destas questões tenham sido colocadas há séculos pela Filosofia, a última questão é um problema totalmente novo e não era praticamente colocada enquanto a humanidade só conhecia a forma textual, mesmo que digital.

A pesquisa científica, em especial na área biomédica, usa de forma crescente o computador como ferramenta. Quantidades crescentes de dados sobre sequenciamento genético, proteômica etc. são mantidos em bancos de dados computacionais (Stein 2008).

O conhecimento codificado em formato “inteligível” por programas permite agenciá-los em tarefas nas quais computadores/programas são claramente mais eficientes que nós. Para o uso em larga escala da tecnologia dos “agentes de software”, deve-se extrair o conhecimento científico hoje já registrado em meio digital mas ainda em formato textual e representá-lo também num formato “inteligível” por programas.

A forma tradicional e institucionalizada através da qual a sociedade contemporânea registra e dissemina o conhecimento científico é através da publicação de artigos em periódicos. Artigos científicos se constituem em bases de conhecimento, mas para leitura e processamento por cientistas, dado ao seu formato textual. O processamento desse conhecimento, através da leitura desses artigos, sua crítica, citação, reprodução dos experimentos aí relatados, inclusão do seu conteúdo em aulas, textos didáticos, manuais e tratados, se constitui num lento processo social. Trabalhamos há anos no projeto de registrar o conteúdo de artigos científicos publicados eletronicamente em formato “inteligível” por programas. Propusemos um ambiente Web de “software” que permita a publicação eletrônica dos artigos simultaneamente de forma convencional, como texto, e em formato de ontologia, conforme ilustrado na Figura 1 a seguir. Este ambiente Web de “software” vai interagir com o autor através de um diálogo estruturado e da análise do texto do artigo, extraindo e representando o conhecimento aí contido no formato de uma ontologia.

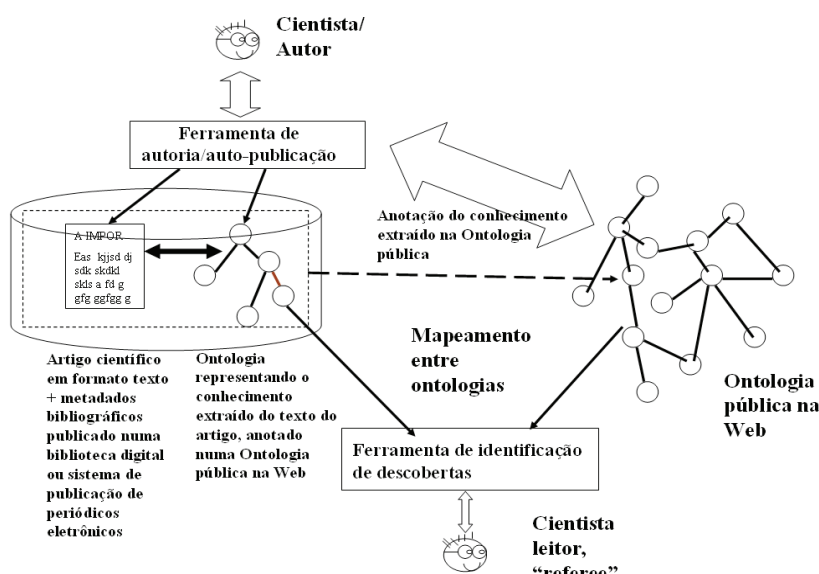


Figura 1 – Ambiente Web de autoria/autopublicação.

Criamos um modelo para os elementos semânticos que constituem o conteúdo de conhecimento de um artigo baseado nos elementos do Método Científico, na forma como eles aparecem em artigos científicos. O objetivo desse modelo é servir de base para a uma nova forma de publicar artigos em formato de ontologias, de modo a tornar seu conteúdo “inteligível” por programas, permitindo assim que esses conteúdos sejam processados de uma forma muito mais sofisticada e útil que em programas convencionais de recuperação de informações, de banco de dados, de *data mining* ou de estatística, auxiliando pesquisadores na recuperação semântica de informações, na avaliação da coerência, na identificação de lacunas no conhecimento científico e de novas descobertas.

A proposta inicial do modelo se baseava na concepção de que o conhecimento científico consiste em propor e provar a existência de *relações entre fenômenos*, até então desconhecidas. É assim que Miller (1947) define o conhecimento científico: *The above remarks imply that science is a search after internal relations between phenomena*. Um fenômeno pode ser definido como ... *an event or a process such as it appears to some human subject: it is a perceptible fact, a sensible occurrence* (Bunge 2004, p. 173).

A construção desse modelo se deu a partir de um forte referencial teórico de Filosofia da Ciência e Metodologia Científica. Esse foi complementado e validado pela análise de um conjunto de 75 artigos científicos em Ciências da Saúde, material riquíssimo que permitiu aperfeiçoar o modelo e verificar na prática as nuances de como o conhecimento científico nesta área é registrado e comunicado em artigos científicos. Este modelo foi formalizado numa Ontologia do Conteúdo de Conhecimento em artigos<sup>1</sup>. Futuros desenvolvimentos da pesquisa vão consistir em usar o modelo como base para a construção de um ambiente Web de autopublicação (Costa 2006) de artigos científicos e para o desenvolvimento de programas que comparem o conteúdo de artigos científicos registrados segundo o modelo com o conhecimento registrado em ontologias médicas, como por exemplo a UMLS – *Unified Medical Language System*<sup>2</sup> ou a *Gene Ontology*<sup>3</sup>, com vistas a identificar novas descobertas científicas (Malheiros 2005). A versão mais acabada desta proposta inicial esta descrita em Marcondes (2007).

O modelo inicial no entanto teve que ser aperfeiçoado a luz das conclusões da análise do último conjunto de artigos analisado. Trata-se de um conjunto de 15 artigos que, ao contrário dos conjuntos anteriores, guarda toda uma coerência interna e visíveis inter-relações entre si. São as chamadas *key publications*, do grupo de três pesquisadores, Elizabeth Blackburn, Carol Greider e Jack Szostak, agraciados com o Prêmio Albert Lasker de Medicina de 2006 (<http://www.laskerfoundation.org/index.html>). Os artigos cobrem o período de 1978 a 1999 e, em seu conjunto, relatam a sequência de marcos de conhecimento da descoberta da enzima telomerase (Blackburn 2006), de seu papel fundamental na reprodução celular, sua influência em processos como

de envelhecimento celular e surgimento de câncer. Os artigos guardam forte relação uns com os outros, mostrando claramente o processo de *identificação inicial de um novo fenômeno, a agregação de novos conhecimentos de modo a caracterizá-lo cientificamente, até sua completa identificação e a investigação de suas possíveis relações com outros fenômenos*. Os resultados da análise desses artigos mostraram fatos com os quais não tínhamos nos deparado até então com a análise do material anterior e que o nosso modelo não dava conta. A análise desse conjunto de artigos impôs portanto a revisão e aperfeiçoamento do modelo inicial, de modo a torná-lo mais completo, abrangente e robusto. Este é o objetivo do presente artigo.

Ele esta estruturado assim: após esta introdução a seção 2 discute o quadro conceitual que embasa o modelo proposto; a seguir, a seção 3 discute os aspectos teóricos que embasam o modelo; a seguir, a seção 4 discute as novas características do conjunto de artigos referentes ao Prêmio Lasker 2006 que fizeram rever o modelo; a seção 5 a seguir apresenta o modelo revisto; finalmente a seção 6 apresenta conclusões e futuras direções de pesquisa.

## Material e métodos

Para a proposição do modelo foram buscados aportes teóricos de disciplinas como Ciência da Informação, em especial de Comunicação Científica, Metodologia da Ciência, Filosofia da Ciência e Ciência da Computação. Foram analisados 75 artigos em Medicina, subdivididos nos seguintes grupos: 20 artigos do periódico Memórias do Instituto Oswaldo Cruz, 20 artigos do periódico Brazilian Journal of Medical em Biological Research, ambos disponibilizados através do portal SciELO e escolhidos a partir da lista dos artigos mais consultados de ambas as publicações, disponível no site de cada uma das publicações; foram analisados ainda 20 artigos sobre células-tronco, escolhidos a partir de três importantes artigos de revisão sobre o tema.

Por fim, foram analisados 15 artigos entre as chamadas *key publications* do grupo de pesquisadores agraciados com o Prêmio Albert Lasker de Medicina do ano de 2006. O texto de Charlton (2006) indica que premiações científicas podem ajudar a identificar a elite da “ciência revolucionária”, usando as palavras de Kuhn, o que era de especial interesse para essa pesquisa.

Para este último grupo, um procedimento metodológico foi ordenar os artigos cronologicamente (ver Anexo 1) e utilizar o artigo de 2006 em que os três agraciados com o Prêmio Lasker (Blackburn 2006) comentam a trajetória das pesquisas que culminaram com a descoberta da enzima telomerase, como um guia para orientação da leitura dos 15 artigos que constituem as *key publications*.

A análise procurava identificar fenômenos descritos no artigo ou o estabelecimento de relações entre fenômenos. Após a identificação de fenômenos e relações no texto de cada artigo, verificava-se se os conceitos correspondentes aos fenômenos e relações existiam na UMLS. Os resultados da análise eram registrados em formulário específico.



A área de Medicina foi escolhida devido ao fato de que artigos científicos da área seguem um rígido padrão formal em seus textos, com seções definidas segundo o chamado padrão IMRAD – Introduction, Method, Results and Discussion –, recomendados pelo *International Committee of Medical Journals Editors*<sup>4</sup> para artigos científicos em periódicos biomédicos, facilitando assim a análise.

## Quadro conceitual

O conhecimento científico conforme veiculado através de artigos de periódicos, consistem em formular “afirmações científicas”, sobre fenômenos (Bunge 2004, p. 173) ou sobre relações entre fenômenos (Miller 1947). As afirmações científicas, conforme será visto nos exemplos da seção 4, refletem um grau crescente de certeza, apropriação e enquadramento, do fenômeno científico e de suas inter-relações, no quadro conceitual ou sistema de conceitos que compõe o conhecimento num determinado domínio científico. Este grau crescente de certeza avança na direção do que Bunge (2004, p. 3) considera ser o objetivo da Ciência, ou seja, responder as *why questions*, buscar explicações para os fenômenos.

As visões clássicas de como se faz ciência apresentadas pelos manuais de metodologia científica e pela Filosofia da Ciência em Popper (2001) e Hempel (1995), separam os procedimentos e o raciocínio empregados nas descobertas científicas da explicação metodológica dos fatos científicos. Existe toda uma ênfase nos aspectos lingüísticos, lógicos e formais da Ciência, que tem origem no positivismo lógico ainda no século XIX (Marcondes 2004).

As visões situadas dentro da “lógica da justificativa”, sejam elas dos positivistas lógicos ou de seus críticos como Popper, são por demais formais e os exemplos analisados mostram que, numa área de pesquisa de ponta atual como a biologia celular, fornecem poucos subsídios para se compreender as práticas de pesquisa e descoberta que levaram à construções científicas que resultaram na descoberta da telomerase. Assim, o material por nós analisado, conforme será mostrado na seção 4, apresenta bastante semelhança com as colocações de Aliseda (2004), Klahr e Simon (1999) e Thagard (1993).

A distinção entre “lógica da descoberta”, em oposição à “lógica da justificativa” é enfatizado em Atocha (2004), numa crítica às visões positivistas que afirmam que: “The context of discovery is taken to be purely psychological” (Aliseda 2004, p. 340).

Ou então:

*Reichenbach's distinction between the contexts of justification and of discovery has left out of its analysis – especially from a formal point of view – a very important part of scientific practice, that which includes issues related to the generation of new theories and scientific explanations, concept formation as well as aspects of progress and discovery in science* (Aliseda 2004, p. 341).

Thagard, ao criticar tanto o método indutivo, privilegiado pelos positivistas lógicos, quanto o método

hipotético-dedutivo de Popper, chamando a ambos de mitos (Thagard 1993, p.176), afirma ao contrário que:

*In well-trod areas of investigation, it may be possible to form a Sharp hypothesis and then test it. But when novel topics are being pursued, researchers in psychology and other fields cannot always start with hypotheses sharp enough to be tested. Often some vague ideas will lead to the collection of some data, which then suggested a refinement of an existing hypothesis. Or results are very different from what was expected may spur abductive formation of a new hypothesis that can then be subject to further test* (Thagard 1993, p. 177).

Klahr e Simon (1999, p.8) também criticam Popper e dão especial ênfase ao contexto da descoberta:

In science there is an important, and extremely common, form of experiment, at times referred to somewhat dismissively as “exploratory,” that is guided by no specific hypothesis to be tested, and no clear control condition, but only a vague and general direction of inquiry. The goal of exploratory experiments is to permit phenomena to appear that will invite exploration or suggest whole new forms of representation or generate new hypotheses.

*The contemporary literature on research methodology is dominated by the notion, promulgated by Popper (1959) among others, that the purpose of observation in general, and experiment in particular, is to test hypotheses in order either to falsify or validate them. In contrast to this position, we have argued that much of the important empirical work in science is undertaken – to use Reichenbach's phrase – in the context of discovery rather than the context of verification (see Simon 1973). That is, a major goal of empirical work in science is to discover new phenomena and generate hypotheses for describing and explaining them, and not simply to test hypotheses that have already been generated. Indeed, theories cannot be tested until they have been created, and creation takes place in the context of discovery, not verification.*

Estamos interessados em buscar um modelo para representar o conhecimento contido no texto de artigos científicos em formato “inteligível” por programas, permitindo a estes programas graus crescente de “inferência”, “decisões” e “raciocínios” sobre o conteúdo desses artigos.

Historicamente a intervenção da humanidade no seu ambiente é cada vez mais indireta, através de diferentes ferramentas. Em especial nas ciências as ferramentas usadas são ferramentas cognitivas. Representar formalmente o conhecimento científico em geral e nas Ciências da Saúde em particular em formato “inteligível” por programas significa em termos práticos desenvolver ontologias. Essas ontologias serão tão mais úteis quanto mais acuradamente refletirem as descobertas científica nesta área, ou seja, quanto mais corresponderem ou forem análogas à realidade desta área, poderem ser usadas como modelos científicos desta realidade, modelos esses passíveis de serem processáveis por computadores para testarem hipóteses, fazerem comparações, diagnósticos, identificarem inconsistências etc., se constituindo em ferramentas cognitivas e instrumentos para o avanço da pesquisa e do conhecimento científico.

O termo inferência em Lógica significa o processo (também chamado de “raciocínio”) de derivar conse-

quências verdadeiras de premissas verdadeiras ou tidas como verdadeiras. Simular processos de inferência num ambiente computacional seria o processo em que, a partir de informações fornecidas a um sistema, este retornaria outras informações que estejam de alguma maneira relacionadas às informações fornecidas.

O conhecimento científico conforme veiculado através de artigos de periódicos, consiste em formular, através da linguagem, proposições contendo *afirmações científicas* sobre fenômenos (Bunge 2004, p. 173), ou relacionando fenômenos entre si (Miller 1947) ou relacionando um fenômeno a suas características.

Buscamos conceituações de “fenômeno” que possam servir para representar formalmente o conhecimento científico em Medicina, objeto dessa pesquisa. Uma definição de fenômeno usada em textos de Filosofia e Metodologia da Ciência seria: *... an event or a process such as it appears to some human subject: it is a perceptible fact, a sensible occurrence* (Bunge 2004, p. 173).

Afirmarções científicas refletem um grau crescente de certeza, apropriação do fenômeno científico e de suas inter-relações, no quadro conceitual ou sistema de conceitos que compõe o conhecimento num determinado domínio científico.

Foram identificadas na análise da literatura que constituiu o material empírico duas formas de conhecimento científico enquanto relações:

A primeira forma seria a apropriação de um fenômeno através da progressiva caracterização do mesmo através coleta sistemática de “afirmações científicas” (Bunge 2004, p. 173) sob a forma de proposições *relacionando o fenômeno à suas características* (Dahlberg 1977, p. 16). Mais especificamente, Dahlberg chama de “características essenciais” aquelas que caracterizam ou dão identidade a um determinado fenômeno e que sem as quais esse fenômeno perderia sua identidade (Guarino 1997). São características que constituem o que Aristóteles chama de Essência ou atributos essenciais da substância (Chauí 2005). Através da coleta sistemática de suas características, testadas cientificamente, um fenômeno é progressivamente identificado e integrado ao sistema de conceitos de um domínio científico.

A segunda forma de conhecimento científico seria *a identificação e o estabelecimento de relações entre fenômenos distintos*, até então desconhecidas.

Relações, na forma proposta acima, também refletiriam graus crescentes de certeza das proposições científicas, desde uma Questão ou Problema: qual o mecanismo que determina a síntese das extremidade dos telômeros? Onde um dos *relata* é desconhecido, passando por uma Hipótese, onde a relação entre os *relata* é hipotética, uma atividade enzimática determina a síntese das extremidade dos telômeros? Até uma Conclusão: a enzima telomerase determina a síntese das extremidade dos telômeros, onde a relação entre os *relata* é provada por um experimento.

No entanto a palavra fenômeno está eivada de conotações subjetivas, necessariamente incompatíveis com o conhecimento científico, por sua vinculação à

Fenomenologia (Chauí 2005). Esta disciplina da filosofia estuda os fenômenos como percebidos por um observador individual, em oposição a real natureza das coisas, o ser real, ou seja, as aparências em oposição à realidade.

Se a observação de um fenômeno pode ser distorcida pelo observador, como já têm sido largamente discutido, se este está condicionado socialmente e historicamente e, em termos científicos, como já mostrou Kuhn, paradigmaticamente, se o conhecimento é construído pelo indivíduo progressivamente, como diz Piaget (1978), o conhecimento científico é uma construção eminentemente social (Ziman 1979). A Ciência enquanto instituição tem mecanismos que asseguram um alto grau de consenso para um determinado estágio de conhecimento num determinado momento histórico. É claro que este estágio de conhecimento provisório, superável, limitado foi (socialmente) construído. No entanto, corresponde ao que é consensualmente identificado como *correspondendo* à realidade, ao estágio de conhecimento sobre a realidade. Na medida em que este conhecimento é partilhado e consensado socialmente, na medida que, com base nele, podemos intervir na realidade, usá-lo como ferramenta cognitiva fazendo previsões sobre ela, este conhecimento – uma representação mental ou um registro, inscrição ou documento capaz de ser apropriada intersubjectivamente – *corresponde à realidade*. Assim, a pesquisa científica, ao observar e estudar os fenômenos, fornece os elementos - o conhecimento científico – para a construção de uma sempre provisória, sempre em construção, inacabada, Ontologia. Esta deve corresponder, o máximo possível num determinado estágio de conhecimento de uma dada ciência, à realidade mesmo.

Barry Smith (2002, p. 2), discutindo a relação entre Ciência e Ontologia (enquanto domínio de conhecimento preocupado com a natureza dos seres), afirma que Ontologia não pretende “explicar” a natureza como a Ciência, seu papel seria vir a seguir das explicações para descrever, organizar e sistematizar o conhecimento obtido pelas descobertas científicas. Este parece ser um lugar a ser ocupado também pela Ciência da Informação.

## Resultados

O processo de crescente caracterização e apropriação científica de um fenômeno e a posterior identificação de relações entre esse fenômeno e outros é ilustrado nos seguintes exemplos e pode ser acompanhado pelos títulos dos artigos pertencentes ao grupo da premiação Lasker 2006, ordenados cronologicamente no Anexo 1:

- No artigo mais antigo analisado do grupo referente ao Prêmio Lasker 2006 (Blackburn & Gall 1978), vê-se esse aspecto da gradual caracterização de um novo fenômeno científico, não só no título, mas também nos objetivos do artigo, colocados no seu abstract: *The extrachromosomal genes coding for the ribosomal RNA in the ciliated protozoan Tetrahymena thermophila we studied with respect to sequences occurring at their termini* (Blackburn 1998, p. 33).

- Num artigo de revisão que mostra um quadro atual da pesquisa sobre telomerase, Cech (2004) afirma que

o objetivo da pesquisa que conduziu a descoberta da telomerase seria identificar a *entidade* (termo por sinal, bastante usado na modelagem de ontologias) responsável pela replicação das extremidades dos cromossomos, até então desconhecida, não caracterizada, não integrada ao quadro do conhecimento existente até então: *Carol Greider, a graduate student in Liz Blackburn's group at the University of California, Berkeley, had chosen an ambitious PhD thesis project: identify the molecular entity responsible for replicating chromosome ends* (Cech 2004, p. 273).

- No mesmo artigo Cech (2004) esclarece que o objetivo das pesquisas de Greider e Blackburn no artigo que marca a descoberta da telomerase seria: *The identification and characterization of this new enzymatic activity was the subject of Greider and Blackburn (1985).*

- Outro exemplo retirado de um artigo do mesmo grupo, mostra que a telomerase ainda não estava identificada como uma enzima (*a terminal transferase-like activity*) nem estava clara a sua relação com o processo de complementação dos telômeros (*which adds the host cell telomeric sequence repeats onto recognizable telomeric ends*): *Based on all these considerations, the proposal was made that telomere replication involves a terminal transferase-like activity which adds the host cell telomeric sequence repeats onto recognizable telomeric ends.* (Shampay et al. 1984), citado em Greider (1985, p. 405).

- Num artigo subsequente do mesmo conjunto, verifica-se uma crescente apropriação científica do mesmo fenômeno. É identificada uma atividade enzimática de transferase, explicitada no título do artigo *Identification of a specific telomere terminal transferase activity in Tetrahymena extracts* (Greider CW, Blackburn EH. Cell. 1985; 43: 405-413).

- No mesmo artigo, as autoras propõem, ainda com pouco grau de certeza, o enquadramento da atividade da telomerase dentro do quadro conceitual já conhecido: *The authors made the reasonable proposal that the activity might be related to known terminal transferases, such as the enzyme that adds CCA to the 3' ends of transfer RNAs* (Cech 2004, p. 273).

- Posteriormente esta atividade enzimática é identificada, ou seja, enquadrada no sistema de conceitos desse domínio científico específico, ou seja, numa classificação de substâncias, e essa enzima é finalmente batizada de telomerase com este nome:

Greider CW, Blackburn EH. *The telomere terminal transferase of Tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity.* Cell. 1987; 51: 887-898.

- Posteriormente a atividade, função ou papel da telomerase como catalisador e molde no processo de sínteses e complementação das extremidades dos telômeros é identificada:

Greider CW, Blackburn EH *A telomeric sequence in the RNA of Tetrahymena telomerase required for telomere repeat synthesis.* Nature. 1989; 337: 331-337.

- Ou então: *Our results indicate the involvement of such sequence-specific telomeric DNA-protein interaction in cell or nuclear division* (Yu 1990, p. 131).

Uma vez caracterizado o fenômeno da complementação das extremidades dos telômeros pela ação da enzima telomerase, começam a surgir propostas estabelecendo relações desse fenômeno com dois outros: a senescência celular, ou seja, o número finito de vezes que uma célula é capaz de se reproduzir, consequência do encurtamento progressivo dos telômeros a cada duplicação e da incapacidade da célula em complementá-los através da ação da telomerase, que conduz a morte celular; e a relação da telomerase com o câncer, identificado como um processo descontrolado de duplicação celular. Esses casos são descritos a seguir:

- *These mutations also lead to nuclear and cell division defects, and senescence, establishing an essential role for telomerase in vivo* (Yu 1990, p. 126).

Ou então, no artigo

- Allsopp RC, Vaziri H, Patterson C, Goldstein S, Younglai EV, Futcher CW, Greider CW, Harley CB. *Telomere length predicts the replicative capacity of human fibroblasts.* Proc. Natl. Acad. Sci. USA. 1992; 89: 10114-10118.

Ou nos trechos, e no artigo seguinte: *This shortening has been proposed to play a role in signaling the cell cycle exit characteristics of senescent cells (14, 15), although a causal role has not been demonstrated* (Prowse 1993, p. 1493). Nesse artigo os autores propõem a existência do que é chamado aqui de uma relação “fraca” ao afirmar que uma relação causal ainda não foi demonstrada. No outro trecho e no artigo é proposta uma relação entre a atividade da telomerase e o câncer: *It has been proposed that the finite cell division capacity of human somatic cells is limited by telomere length (10). This is consistent with reports that telomerase activity is often high in cancer and immortalized tissue culture cells* (Mceachern 1995, p. 403).

E ainda no artigo:

- Rudolph KL, Chang S, Lee HW, Blasco M, Gottlieb G, Greider CW, DePinho RA. *Longevity, stress response, and cancer in aging telomerase deficient mice.* Cell. 1999; 96: 701-716.

No artigo mais antigo analisado do grupo referente ao Prêmio Lasker 2006 (Blackburn & Gall 1978), vemos esse aspecto da gradual caracterização de um novo fenômeno científico, não só no título, mas também nos objetivos do artigo, colocados no seu abstract: *The extrachromosomal genes coding for the ribosomal RNA in the ciliated protozoan Tetrahymena thermophila we studied with respect to sequences occurring at their termini* (Blackburn 1998, p. 33).

## Um modelo para conhecimento médico em artigos científicos

A seguir um modelo para representar, em formato “inteligível” por programas o conhecimento contido em artigos científicos em Medicina é proposto. São mostradas as bases do modelo, a partir do conteúdo dos artigos que constituem o campo empírico e de sua análise dentro do quadro conceitual descrito anteriormente. Como já foi mencionado, a versão anterior do modelo (Marcondes



2007) enfatizava o papel das hipóteses enquanto relações entre fenômenos, segundo uma visão mais convencional da Ciência, baseada no Método Hipotético-dedutivo; a versão atual do modelo, apresentada aqui, incorpora elementos que representam o conhecimento científico sendo progressivamente construído, através da caracterização e incorporação de um novo fenômeno.

Artigos científicos distinguem-se pelo tipo de raciocínio que empregam ao conduzirem a argumentação acerca dos fenômenos discutidos. Existem artigos teóricos e artigos experimentais. Esta classificação é baseado em Hutchins (1997) e Gross (1990) e em textos a partir da visão de abdução em Pierce (1977), como processo de descoberta de novos *insights* em Ciência (Hoffman 1997, Magnani 2001, Paavola 2004, Aliseda 2004).

**Artigos teórico-abdutivos** se caracterizam por discutirem questões de maior abrangência. Analisam criticamente diversas hipóteses anteriores, mostrando suas fragilidades. Estes artigos são os que têm mais potencial de apresentarem contribuições para a Ciência, já que discutem ou questionam o paradigma vigente (Kuhn 2003). Sua contribuição é uma nova hipótese, indicando um novo caminho de pesquisa. O tipo de raciocínio empregado é o abduutivo ou seja, o *insight* sobre a solução de questões não explicadas na Ciência e a formulação de novas hipóteses de solucioná-las.

**Artigos experimentais** constam necessariamente de um experimento empírico; dividem-se em exploratórios, dedutivos e indutivos. Caracterizam-se por discutirem questões num escopo de abrangência limitado. Não discutem os rumos de uma teoria científica, mas se limitam a confirmá-la ou aperfeiçoá-la. Sempre trazem resultados experimentais.

**Artigos experimentais-exploratórios** têm um caráter exploratório ao desvendar e buscar caracterizar um fenômeno, trabalhando na direção proposta por Dahlberg de formular e provar proposições que caracterizam um fenômeno,

**Artigos experimentais-dedutivos** trabalham a partir de relações entre fenômenos já formuladas anteriormente, cujas referências vêm citadas, aplicando-as a testando-as e validando-as um contexto específico. Os **artigos experimentais-indutivos** se caracterizam por proporem e testarem novas relações entre fenômenos.

A estrutura textual dos artigos em Ciências da Saúde segue o padrão IMRAD, como já foi mencionado. Essa estrutura corresponde à “surface structure” de Chomsky (1981) e a microestrutura de Kintsh e Van Dijk (1972). Já os componentes semânticos num artigo, que compõe o modelo proposto, correspondem à *deep structure* de Chomsky e à macroestrutura de Kintsh e Van Dijk; são descritos a seguir, identificados em maiúsculas.

Um PROBLEMA expressa uma carência, insatisfação ou deficiência conceitual com o atual estado de conhecimento num domínio. Um PROBLEMA pode se desdobrar em OBJETIVOS de pesquisa e, eventualmente, na formulação mais precisa de uma QUESTÃO que endereça a deficiência conceitual; esta QUESTÃO pode ser referir a um FENÔMENO (nos artigos EX-

PLORATÓRIOS), ou a dois ou mais FENÔMENOS envolvidos numa RELAÇÃO\_ENTRE\_FENÔMENOS ou HIPÓTESE. Uma HIPÓTESE relaciona dois ou mais FENÔMENOS através de um TIPO-DE-RELAÇÃO.

Um autor num artigo pode formular uma hipótese original – HIPÓTESE(o) ou tomar a hipótese prévia – HIPÓTESE(p) - de outros autores; neste caso uma ou mais citações referentes à HIPÓTESE(p) – CITAÇÕES(h) - são feitas. Um autor também pode analisar várias HIPÓTESES(p) para mostrar que elas são insatisfatórias como soluções para o PROBLEMA e formular sua HIPÓTESE(o). Um artigo teórico se justifica simplesmente por propor uma nova HIPÓTESE(o).

Da hipótese, num artigo experimental, deve ser derivado um EXPERIMENTO capaz de ser observável empiricamente. Em um artigo científico EXPERIMENTAL, significa ter RESULTADOS observados segundo determinada MEDIDA, em determinado CONTEXTO segundo determinada METODOLOGIA. Este CONTEXTO onde os FENÔMENO(s) relacionados na HIPÓTESE são observados pode ser desdobrado em AMBIENTE – comunidade ou instituição onde o fenômeno ocorre -, ESPAÇO - o lugar onde o fenômeno ocorre -, TEMPO ou época em que o fenômeno ocorre e GRUPO de indivíduos onde o fenômeno ocorre. Todo artigo também traz uma CONCLUSÃO, na forma de uma proposição sobre um fenômeno ou sobre RELAÇÕES\_ENTRE\_FENÔMENOS.

O desenvolvimento do raciocínio num **artigo teórico abduutivo** segue o seguinte padrão:

- dado um PROBLEMA, com os seguintes aspectos e dados
- os seguintes Autores/HIPÓTESES anteriores para sua solução não são satisfatórias,
- *diante disso, propomos a seguinte HIPÓTESE original.*

O desenvolvimento do raciocínio num **artigo experimental dedutivo** segue o seguinte padrão:

- dado um PROBLEMA, com os seguintes aspectos e dados,
- os seguinte Autores formularam HIPÓTESES anteriores para sua solução,
- diante disso, escolhemos a seguinte (uma das HIPÓTESE anteriores).

Ampliamos e re-contextualizamos esta HIPÓTESE anterior; desenvolvemos o seguinte EXPERIMENTO para testar esta HIPÓTESE anterior;

- o EXPERIMENTO apresentou os seguintes RESULTADOS.

O desenvolvimento do raciocínio num **artigo experimental indutivo** segue o seguinte padrão:

- dado um PROBLEMA, com os seguintes aspectos e dados,
- uma solução para este PROBLEMA pode se basear na seguinte HIPÓTESE,
- desenvolvemos o seguinte EXPERIMENTO para estar esta HIPÓTESE,



- estes testes apresentaram os seguintes RESULTADOS.

O desenvolvimento do raciocínio num artigo **experimental exploratório** segue o seguinte padrão:

- dado um PROBLEMA ou FENÔMENO ainda não bem caracterizado,

- desenvolvemos o seguinte EXPERIMENTO que permite identificar a(s) seguinte(s) CARACTERÍSTICA(s) desse FENÔMENO.

Estes esquemas resultaram no atual modelo ou **Ontologia do Conteúdo de Conhecimento em Artigos Científicos – OCCAC**, também ilustrado na Figura 2:

Classes: artigos TEÓRICOS

têm componentes

PROBLEMA

HIPÓTESE(a)

HIPÓTESE(o)

CONCLUSÃO(ões) e

artigos EXPERIMENTAIS

Subclasses: artigos EXPLORATÓRIOS

têm componentes

PROBLEMA

FENÔMENO

EXPERIMENTO

CONCLUSÃO(ões)

artigos INDUTIVOS

têm componente

PROBLEMA

HIPÓTESE(o)

EXPERIMENTO

CONCLUSÃO(ões) e

artigos DEDUTIVOS

têm componente

PROBLEMA

HIPÓTESE(a)

EXPERIMENTO

CONCLUSÃO(ões)

COMPONENTES semânticos de artigos

PROBLEMA

Subcomponentes: OBJETIVOS

QUESTÃO de pesquisa

HIPÓTESE (prévia ou nova)

Subcomponentes: FENÔMENO(S)

TIPO-DE-RELAÇÃO

REFERENCIAS (somente nas HIPÓTESES

prévias)

FENÔMENO (um, nos artigos EXPLORATÓRIO(s))

Subcomponentes: CARACTERÍSTICAS(s)

EXPERIMENTO

Subcomponentes: RESULTADOS (dados quantitativos)

MEDIDA

CONTEXTO

Subcomponentes:

ESPAÇO

TEMPO

GRUPO social

CONCLUSÃO(ões)

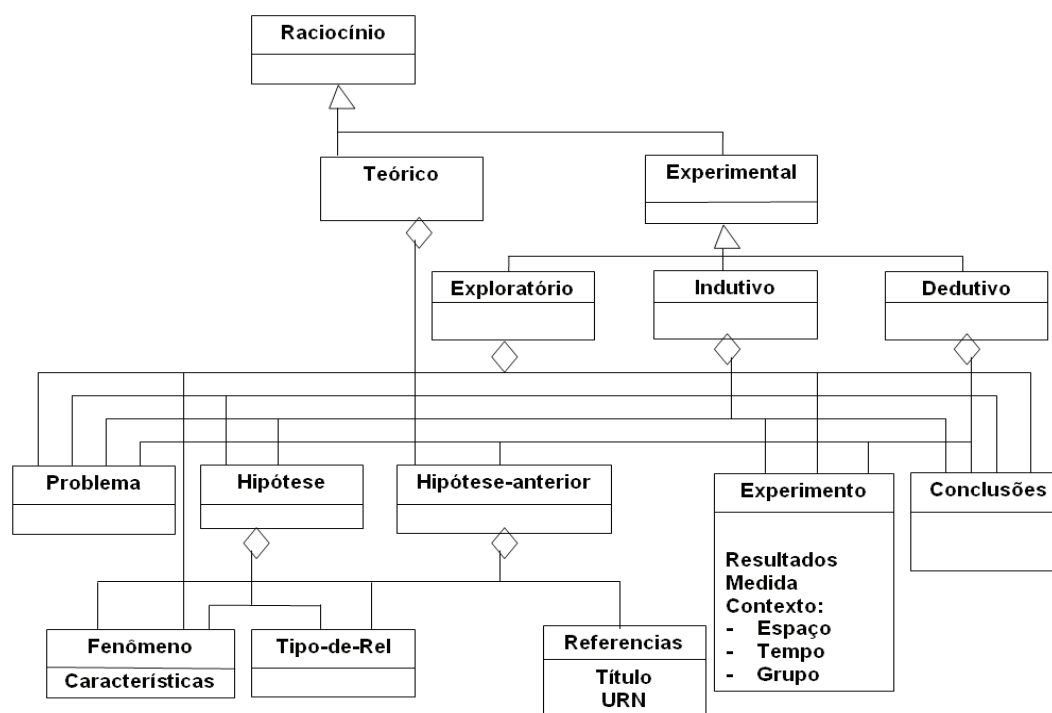


Figura 2 - Ontologia do Conteúdo de Conhecimento em Artigos Científicos Digitais.

A análise do seguinte artigo mostra como componentes semânticos de um artigo (neste caso a hipótese) seriam identificados e registrados segundo o modelo.

- Camara GNI, Cerqueira DM, Oliveira APG et al. *Prevalence of human papillomavirus types in women with pre-neoplastic and neoplastic cervical lesions in the Federal District of Brazil*. Mem. Inst. Oswaldo Cruz. [online]. 2003 Oct.; 98:7.

3 passos:

Passo 1 – Tipo de raciocínio identificado: experimental-dedutivo, ou seja, o artigo realiza um experimento para mostrar a prevalência de HPV, hipótese já formulada anteriormente por outro autor.

Passo 2 – Elementos semânticos de conhecimento são identificados no texto, como a hipótese formulada pelo autor:

Hipótese (anterior)

Antecedente: HPV

Tipo de Relação: causa

Consequente: lesões cervicais pre-neoplásticas e neoplásticas

Passo 3 – Cada um desses elementos é mapeado em termos ou relações da base de conhecimento pública, a UMLS, UMLS Semantic Network

Papillomavirus, Human

“Causes”, UMLS Semantic network relação R147

Colonic Neoplasms

## Conclusões

A Ciência da Informação vêm de uma longa tradição teórica, metodológica e prática que converge para muitas das questões atuais colocadas pela proposta da Web semântica e para a construção de ontologias. Uma área de pesquisa recorrente na Ciência da Informação é o processamento semântico de informações por computadores, com contribuições já históricas como as de Shera (1957), Luhn (1960) e mais recentes como as de Gardin (2001) e Kajikawa (2006).

A Ciência da Informação pode e deve ir além de prover técnicas e metodologias para permitir acesso ao texto completo de artigos científicos em bibliotecas digitais e repositórios científicos. As metodologias atuais de indexação consistem em assinalar a registros bibliográficos palavras-chave ou termos de um vocabulário controlado isolados, sem nenhuma relação ou papel semântico entre si. Mas as pesquisas em CI relativas à importância das relações e seu papel semântico também já são históricas, como atesta a recente revisão de Khoo (2007).

O modelo proposto, ao identificar tipos de relações (com sua semântica própria) e características dos fenômenos descritos num artigo, permite realizar inferências automáticas e, por exemplo, resolver consultas sofisticadas como:

- que (outros) artigos (também) têm hipóteses relacionando HPV como causa de lesões pré-neoplásticas e neoplásticas em mulheres?

- que artigos têm hipóteses relacionando outros fatores que não HPV como causa de lesões pré-neoplásticas e neoplásticas em mulheres?

- que artigos identificam outras características relacionadas com a estrutura das extremidades das moléculas lineares de rDNA?

- que artigos identificam características do fenômeno de replicação de telômeros que podem ser associadas à atividade enzimática?

A implementação prática do modelo de registro do conteúdo de artigos científicos conforme descrito pressupõe o desenvolvimento de todo um conjunto de ferramentas de software que processem conteúdos estruturados segundo o proposto. Trata-se na verdade de um programa de pesquisa. Acreditamos que um ponto de partida sólido para isto é estabelecer o modelo proposto aqui. Duas destas aplicações inicialmente visualizadas estão descritas em Malheiros (2005) e Costa (2006), e delineadas na Figura 1.

O modelo mostra os benefícios de um formato semanticamente rico para o registro do conteúdo de artigos científicos, baseado em relações, viabilizando que, através delas, programas agentes de “software” realizem “inferências”. Forsythe (1989), falando sobre as primeiras experiências de construção de sistemas especialistas nas décadas de 1970 e 1980, chama a aquisição de conhecimento de *o gargalo da construção de sistemas especialistas*. Realizar a *aquisição de conhecimento* diretamente a partir dos autores/pesquisadores a partir de seus artigos científicos, que já trazem um alto grau de formalização desse conhecimento, como é a presente proposta, pode se mostrar uma alternativa promissora. Com o apoio de ferramentas de software adequadas, o conteúdo de artigos científicos pode ser extraído como um subproduto do processo de autoria/autopublicação de artigos científicos pelos próprios autores, hoje já bastante comum quando estes submetem seus trabalhos a repositórios, publicações eletrônicas ou bibliotecas digitais. A visão delineada aqui (Figura 1) resultará em publicações eletrônicas com um potencial de tratamento semântico de seu conteúdo por programas agentes de “software” muito mais rico do que é possível nas publicações eletrônicas textuais atuais.

## Notas

1. Disponível em [www.professores.uff.br/marcondes/Scientific\\_reasoning.owl](http://www.professores.uff.br/marcondes/Scientific_reasoning.owl)

2. Disponível em [www.nlm.nih.gov/pubs/factsheet/umls.html](http://www.nlm.nih.gov/pubs/factsheet/umls.html)

3. Disponível em <http://www.geneontology.org/>

4. Disponível em [www.icmje.org](http://www.icmje.org)

## Referências bibliográficas

Aliseda A. Logics in scientific discovery. Foundations of Science. 2004; 9: 339-63.

Berners-Lee T, Hendler J, Lassila O. The semantic web. Sci Am. 2001 May. Disponível em <<http://www.scian>

com/2001/0501issue/0501berniers-lee.html>. Acesso em 24 maio 2001.

Blackburn EH, Greider CW, Szostak J. Telomeres and telomerase: the path from maize, Tetrahymena and yeast to human cancer and aging. *Nature*. 2006 October; 12(10).

Bunge M. *Philosophy of Science*. New Brunswick, London: Transaction Publishers; 1998.

Cech TR. Beginning to understand the end of the chromosome. *Cell*. 2004 Jan.; 116:273-9.

Charlton BG. Editorial. Scientometric identification of the elite 'revolutionary science' research institutions by analysis of trends in Nobel prizes 1947-2006. *Medical Hypotheses*. 2007; 68:931-4.

Chauí M. *Convite à Filosofia*. São Paulo: Érica; 2005.

Chomsky N. *Regras e representações: a inteligência humana e seus produtos*. Rio de Janeiro: Ed. Zahar; 1981.

Costa LC da. Uma ferramenta para edição, extração e representação do conhecimento contido em artigos científicos publicados na Web. Projeto de Tese de Doutorado para ingresso no PPGCI UFF/IBICT. Niterói; 2006.

Dahlberg I. *Ontical structures and universal classification*. Bangalore: Sarada Ranganathan Endowment for Library Science; 1978.

Forsythe DE, Buchanan BG. Knowledge acquisition for expert systems: some pitfalls and suggestions. *IEEE Transactions on Systems, Man and Cybernetics*. 1989 May/Jun.; 19(3):435-42. Disponível em: <<http://ieeexplore.ieee.org/iel1/21/1336/00031050.pdf?tp=&isnumber=&arnumber=31050>>. Acesso em: 21 abr. 2008.

Gardin J-C. Vers un remodelage des publications savantes: ses rapports avec sciences de l'information. In: *Colloques Isko-France. Filtrage et résumé automatique de l'information sur les réseaux*, Conference invitee, Univesité de Nanterre – Paris X; 2001. (Conference proceedings).

Gross AG. *The Rhetoric of Science*. Cambridge, EUA; London, Inglaterra: Harvard University Press; 1990.

Guarino N. Some organizing principles for a unified top-level ontology. New version of paper presented at AAAI Spring Symposium on Ontological Engineering, Stanford University; March 1997.

Hempel K. *Aspects of scientific explanation: and other essays in the philosophy of science*. New York:Free Press; 1965.

Hutchins J. On the structure of scientific texts. In: *UEA Papers in Linguistics*, 5 th., 1977, Norwich. **Proceedings**. Norwich, UK: University of East Anglia, 1977. p. 18-39. Disponível em: <<http://ourworld.compuserve.com/homepages/wjhutchins/UEAP/L-1977.pdf>>. Acesso em: 30 mar. 2006.

Kajikawa Y, Abe K, Noda S. Filling the gap between researchers studying different materials and different methods: a proposal for structured keywords. *J Inform Sci*. 2006; 32: 511-24.

Khoo C, Na JC. Semantic Relations in Information Science. *Ann Rev Inform Sci Technol*. 2007; 157-228.

Kintsh W, Van Dijk TA. Towards a model of text comprehension and production. *Psychol Rev*. 1972; 84(5):363-93.

Klahr D, Simon HA. *Studies of scientific Discovery: complementary approaches and convergent findings*. *Psychol Bull*. 1999; 125(5): 524-43.

Kuhn TS. *A estrutura das revoluções científicas*. São Paulo:Perspectiva; 2003. (Série Debates Ciência).

Luhn H. Keyword in Context Index for Technical Literature. *American Documentation*. 1960; 11(4): 288-95.

Malheiros L. A identificação de novas descobertas científicas através da análise do conhecimento contido em artigos científicos. Projeto de Tese de Doutorado para ingresso no PPGCI UFF/IBICT. Niterói; 2005.

Marcondes D. *Filosofia analítica*. Rio de Janeiro:Jorge Zahar; 2004. (Coleção Passo a passo).

Marcondes CH, Mendonça MAR, Malheiros CLC da, Santos TCP, Pereira LG. Representing and coding the knowledge embedded in texts of Health Science Web published articles. In: Chan L, Marten B. Ed. *ICCC EIPub - International Conference on Electronic Publishing*, Viena, Austria; 2007. Disponível em <<http://elpub.scix.net>>.

Miller DL. Explanation versus description. *Philosoph Rev*. 1947; 56(3): 306-12.

Piaget J. *Psicologia e Epistemologia: por uma teoria do conhecimento*. Rio de Janeiro: Forense; 1978.

Popper K. *A lógica da pesquisa científica*. São Paulo: Ed. Cultrix, Ed. USP; 2001.

Shera JH, Kent A, Perry JW, editors. *Information systems in documentation*. New York: Interscience Publishers; 1957. (Advances in Documentation and Library Science, v. 1).

Smith B. Beyond concepts: ontology as reality representation. In: FOIS - International Conference on Formal Ontology and Information Systems, Turin; nov. 2004. Disponível em: <http://ontology.buffalo.edu/bfo/BeyondConcepts.pdf>. Acesso em: 16 mar. 2007.

Smith B. *Ontology and information systems*. 2002. Disponível em: <[http://ontology.buffalo.edu/ontology\(PIC\).pdf](http://ontology.buffalo.edu/ontology(PIC).pdf)>. Acesso em: 26 maio 2008.

Stein LD. Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nature Rev Gen*. 2008 Sept.; 9:678-88.

Thagard P. *Computational Philosophy of Science*. Cambridge: The MIT Press; 1993.

Ziman J. *Conhecimento público*. Belo Horizonte: Itatiaia, São Paulo: Ed. da Universidade de São Paulo; 1979.



## Anexo 1 – “Key publications” do Prêmio Lasker/2006, ordenados cronologicamente

Ano	Artigo
1978	Blackburn, E.H. and Gall, J.G. (1978) A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in <i>Tetrahymena</i> . <i>J. Mol. Biol.</i> 120: 33-53.
1987	Szostak, J.W. and Blackburn, E.H. (1982) Cloning yeast telomeres on linear plasmid vectors. <i>Cell</i> 29: 245-255.
1983	Murray, A.W. and Szostak, J.W. (1983) Construction of artificial chromosomes in yeast. <i>Nature</i> 305: 189-193.
1984 JAN	Shampay, J., Szostak, J.W., and Blackburn, E.H. (1984) DNA sequences of telomeres maintained in yeast. <i>Nature</i> 310: 154-157.
1984 MAIO	Dunn, B.L., Szauter, P., Pardue, M.-L., Szostak, J.W. (1984) Transfer of telomere-adjacent sequences to linear plasmids by recombination. <i>Cell</i> 39: 191-201.
1985	Greider, C.W. and Blackburn, E.H. (1985) Identification of a specific telomere terminal transferase activity in <i>Tetrahymena</i> extracts. <i>Cell</i> 43: 405-413.
1987	Greider, C.W. and Blackburn, E.H. (1987) The telomere terminal transferase of <i>Tetrahymena</i> is a ribonucleoprotein enzyme with two kinds of primer specificity. <i>Cell</i> 51: 887-898.
1988 NOV	Greider, C.W. and Blackburn, E.H. (1987) A telomeric sequence in the RNA of <i>Tetrahymena</i> telomerase required for telomere repeat synthesis. <i>Nature</i> 337: 331-337.
1989 JAN	Lundblad V. and Szostak, J.W. (1989) A mutant with a defect in telomere maintenance leads to senescence in yeast. <i>Cell</i> 57: 633-643.
1990	Yu, G.L., Bradley, J.D., Attardi, L.D. and Blackburn, E.H. (1990) In vivo alteration of telomere sequences and senescence caused by mutated <i>Tetrahymena</i> telomerase RNAs. <i>Nature</i> 344: 126-132.
1992	Allsopp, R.C., Vaziri, H., Patterson, C., Goldstein, S., Younglai, E.V., Fletcher, C.W., Greider, C.W., and Harley, C.B. (1992) Telomere length predicts the replicative capacity of human fibroblasts. <i>Proc. Natl. Acad. Sci. USA</i> 89: 10114-10118.
1993	Prowse, K.R., Avilion, A.A., and Greider, C.W. (1993) Identification of a nonprocessive telomerase activity from mouse cells. <i>Proc. Natl. Acad. Sci. USA</i> 90: 1493-1497.
1995	McEachern, M.J. and Blackburn, E.H. (1995) Runaway telomere elongation caused by telomerase RNA mutations. <i>Nature</i> 376: 403-409.
1999	Rudolph, K.L., Chang, S., Lee, H.W., Blasco, M., Gottlieb, G., Greider, C.W., and DePinho, R.A. (1999) Longevity, stress response, and cancer in aging telomerase deficient mice. <i>Cell</i> 96: 701-716
2001	Kim, M.M., Rivera, M.A., Botchkina, I.L., Shalaby, R., Thor, A.D., and Blackburn, E.H. (2001) A low threshold level of expression of mutant-template telomerase RNA is sufficient to inhibit tumor cell growth. <i>Proc. Natl. Acad. Sci. USA</i> 98: 7982-7987



## Sobre os autores

### *Carlos Henrique Marcondes*

<http://www.professores.uff.br/marcondes>.

Professor do Departamento de Ciência da Informação da UFF, onde atua nos cursos de graduação em Arquivologia e Biblioteconomia, professor e atual coordenador do PPGCI/UFF – Mestrado – Programa de Pós-graduação em Ciência da Informação -, professor do Curso de Especialização em Informação Científica e Tecnológica em Saúde, do Ict/Fiocruz, pesquisador do CNPq, desenvolvendo pesquisa em modelização de artigos científicos como ontologias, consultor “ad hoc” da Capes e CNPq, membro do conselho editorial e “referee” de diversos periódicos científicos, membro da Câmara Técnica de Documentos Digitais do Conarq/Arquivo Nacional, atua na área de tecnologia da informação aplicadas ao tratamento de informações, é autor de diversos artigos científicos nessa área; atuou como consultor em projetos como a Biblioteca Digital de Teses e Dissertações – BDTD - do Ibict e no desenvolvimento do servidor SciELO/Open Archives.

### *Marília Alvarenga Rocha Mendonça*

Graduada em Biblioteconomia pela Escola de Biblioteconomia da Universidade Federal de Minas Gerais (1996), especialista em Planejamento, Organização e Direção de Arquivos (1985) e Organização e Administração de Bibliotecas Universitárias (1986) pela Universidade Federal Fluminense, Mestre em Administração pela Universidade Federal Fluminense (2001). Atuou como bibliotecária na gerência da Biblioteca da Faculdade de Educação/UFG (1972/79), na implantação e gerenciamento do Centro de Microfilmagem/UFG (1979/82), no Núcleo de Documentação/UFF como gerente do Arquivo Geral e da Biblioteca da Faculdade de Farmácia (1982/94). Atualmente é professora assistente do Departamento de Ciência da Informação da Universidade Federal Fluminense. Pesquisadora do Grupo de Pesquisa “Informação, Conhecimento e Tecnologia da Informação”, tendo como linha de pesquisa “gestão de conteúdos digitais e tecnologia da informação”. Participa, desde 2003, de projetos de pesquisa CNPq/UFF, cujos resultados têm sido apresentados em eventos nacionais e internacionais e publicados em Anais e na forma de artigos científicos em periódicos nacionais e internacionais. Possui um livro publicado e três capítulos de livros.

**Artigos originais**

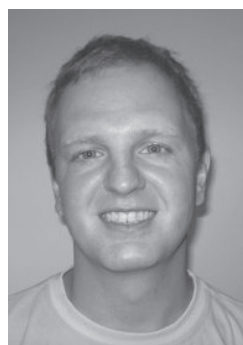
## **Vantagens e limitações das ontologias formais na área biomédica**

DOI: 10.3395/reciis.v3i1.241pt



### ***Stefan Schulz***

Instituto de Biometria  
Médica e Informática da  
Medicina, Centro Médico  
Universitário Freiburg, Frei-  
burg, Alemanha  
steschulz@uni-freiburg.de



### ***Holger Stenzhorn***

Instituto de Biometria  
Médica e Informática da  
Medicina, Centro Médico  
Universitário Freiburg, Frei-  
burg, Alemanha  
holger.stenzhorn@unikli-  
nik-freiburg.de

### ***Martin Boeker***

Instituto de Biometria Médica e Informática da  
Medicina, Centro Médico Universitário Freiburg,  
Freiburg, Alemanha  
martin.boeker@uniklinik-freiburg.de

### ***Barry Smith***

Departamento de Filosofia e Centro de Excelência  
em Bioinformática e Ciências Biológicas e Centro  
Nacional de Ontologia Biomédica, Universidade de  
Buffalo, Buffalo, EUA  
phismith@buffalo.edu

## **Resumo**

Propomos uma tipologia dos artefatos de representação para as áreas de saúde e ciências biológicas, e a associação dessa tipologia com diferentes tipos de ontologia formal e lógica, chegando a conclusões quanto aos pontos fortes e limitações da ontologia de diferentes tipos de recursos lógicos, enquanto mantemos o foco na lógica descritiva.

Consideramos quatro tipos de representação de área: (i) representação léxico-semântica, (ii) representação de tipos de entidades, (iii) representação de conhecimento prévio, e (iv) representação de indivíduos.

Defendemos uma clara distinção entre os quatro tipos de representação, de forma a oferecer uma base mais racional para o uso das ontologias e artefatos relacionados no avanço da integração de dados e interoperabilidade de sistemas de raciocínio associados.

Destacamos que apenas uma pequena porção de fatos cientificamente relevantes em áreas como a biomedicina pode ser adequadamente representada por ontologias formais, quando estas últimas são concebidas como representações de tipos de entidades. Particularmente, a tentativa de codificar conhecimento padrão ou probabilístico pela utilização de ontologias assim concebidas é fadada à produção de modelos não intencionais e errôneos.

## **Palavras-chave**

ontologia biomédica; lógica descritiva; ontologia formal; representação do conhecimento

## Introdução

É cada vez mais reconhecido o fato de que a complexidade das áreas de assistência à saúde e ciências biológicas necessita de um consenso a respeito dos termos e linguagem utilizados em documentos e na comunicação. Tal necessidade é impulsionada pelo crescimento exponencial de dados gerados nos contextos de assistência ao paciente e de pesquisas biológicas. Atualmente, tais dados não podem ser completamente explorados em termos de integração, recuperação ou

interoperabilidade, porque os sistemas básicos de terminologia e classificação (frequentemente classificados sob o tópico “terminologia biomédica” – ver Tabela 1) são inadequados, de diversas formas. Sua heterogeneidade reflete as diferentes experiências, tarefas e necessidades de diferentes comunidades – incluindo aquelas à parte da tecnologia da informação – e cria um grave obstáculo à interoperabilidade e agregação consistentes de dados, conforme exigido pela pesquisa biomédica, assistência à saúde, e medicina translacional.

**Tabela 1 – Exemplos de terminologia biomédica.** A maior parte dos termos abaixo é disponibilizada através do Metatesouro UMLS – Sistema Unificado de Terminologia Médica (*Unified Medical Language System (UMLS) Metathesaurus*), um sistema geral que abrange uma ampla variedade de sistemas de terminologia biomédica (NLMb 2008, McCray et al. 1995)

Termo	Propósito
ICD-9-CM/ICD-10 (OMS, 2008)	Classificação de doenças, estatísticas de saúde, faturamento hospitalar
Dicionário de Medicamentos da OMS ((UMC, 2008)	
ATC (WHOCC, 2008),	
RxNorm (NLMa, 2008)	Classificação de medicamentos
DM+D (NHS, 2008)	
Tesouro e Metatesouro da NCI (NCI, 2008)	Pesquisa do câncer
LOINC (REGENSTRIEF INSTITUTE, 2008)	Comunicação interlaboratorial
MedDRA (NORHTROP GRUMMAN, 2008)	Atividades regulatórias relacionadas à medicina
DICOM (MITA, 2008)	Descrições de imageamento médico e processos de imageamento
MeSH (NLM, 2008)	Indexação da literatura médica
SNOMED CT (IHTSDO, 2008)	Documentação clínica

## Conhecimento ontológico e biomédico

O que era anteriormente denominado de “sistemas de terminologia” ou “terminologia” é atualmente vagamente chamado de “ontologia”. O termo, inicialmente, tornou-se comum nas esferas da biologia através do sucesso da Ontologia Genética (OG), e sua utilização está se tornando cada vez mais popular também na área médica. As chamadas disciplinas “ômicas” caracterizam mais um incentivo para seu desenvolvimento e adoção. Dentro deste contexto, a iniciativa Oficina de Ontologias Biomédicas Abertas (*Open Biomedical Ontologies (OBO) Foundry*) conta com mais de 60 ontologias atualmente e, beneficiando-se do sucesso da OG, está se tornando um recurso padrão (Smith et al. 2007).

O próprio termo “ontologia”, porém, é claramente afetado por múltiplas interpretações inconsistentes (Kusnierczyk 2006) e, assim, os usuários tendem a ter expectativas irreais a respeito do que as ontologias podem alcançar (Stenzhorn et al. 2009). Sendo assim, a utilização deste termo deve, preferencialmente, ser precedida da explanação de seu significado pretendido. A título de ilustrar os tipos de problemas que podem

surgir, podemos mencionar o absoluto contraste entre as definições desenvolvidas pelos profissionais da área da informática, e aquelas inspiradas por filósofos:

– Ontologia (Ciência da Computação): Uma ontologia define (ou especifica) os conceitos, relações, e outras distinções relevantes para a modelagem de um domínio. A especificação assume a forma das definições de terminologia representacional (classes, relações, e assim por diante), que dão significado ao termo e restrições formais para sua utilização coerente (Gruber 1992).

– Ontologia (Filosofia): A ontologia é o estudo daquilo que existe (Quine 1948). Ontologias formais são teorias que tentam dar as fórmulas matemáticas precisas das propriedades e relações entre certas entidades (Hofweber 2004).

Embora haja grande diferença entre estas duas famílias de definição, as ontologias são consideradas, em ambos os casos, sistemas formais que aplicam princípios fundamentais e formalismos, baseando-se em lógica matemática, para representar determinados tipos de entidades, seja no âmbito da mente e linguagem (“conceitos”), ou no âmbito da realidade (“propriedades”,

“tipos” e “classes”). A principal função da ontologia é, em ambos os casos, fornecer um sistema de distinções independentes de áreas, para estruturar teorias específicas para cada área, com o objetivo de integrar e recuperar dados, e promover a interoperabilidade. Aqui, estamos interessados apenas nas ontologias nas quais uma abordagem formal é utilizada para apoiar um objetivo deste tipo. Para destacar esta característica, utilizaremos o termo “ontologia formal” neste documento. Acreditamos que o foco na formalidade distingue muito claramente a nova geração de ontologias biomédicas – incluindo o SNOMED CT, e versões recentes da Ontologia Genética (OG) – de seus antecessores, semelhantes a termos, que ainda trazem indícios de suas origens na área da biblioteconomia e classificação literária.

Este artigo enfoca o papel que a ontologia formal pode representar na solução dos problemas causados pela heterogeneidade de sistemas de terminologia e classificação utilizados na área biomédica. Queremos esclarecer como a representação de entidades estudadas pelas ciências biológicas pode se beneficiar das ontologias formais, de forma a auxiliar a captura do conhecimento da área de forma mais adequada. Abordamos dois importantes aspectos raramente mencionados explicitamente: (i) a representação do meta-conhecimento - ou conhecimento por experiência; e (ii) a relação das ontologias com a linguagem humana. Buscamos destacar o papel desempenhado por estes fatores no desenvolvimento e utilização das ontologias formais. Procuramos ainda esclarecer as situações em que o conhecimento da área não pode ser adequadamente explicado pelas ontologias formais, especialmente devido a imprecisão e incerteza. Duas questões surgem neste ponto:

– Que critérios podem ser utilizados para se delinear os tipos de conhecimento que podem ser razoavelmente expressos pelas ontologias formais?

– Como os demais tipos de conhecimento podem ser codificados de forma a satisfazer as exigências de integração, recuperação e interoperabilidade?

Procuramos responder essas questões através do enfoque dos padrões de representação desenvolvidos pela comunidade Semantic Web. Fornecemos exemplos da utilização desse formalismo na representação de entidades biomédicas. Destacamos, ainda, algumas concepções errôneas e erros comuns no desenvolvimento da ontologia, e mostramos como podem ser retificados.

## Representações informais

Um esquema simples de representação universal que serve aos propósitos da representação de uma ampla gama de entidades e relações entre elas é fornecido pelo chamado trio Objeto – Atributo – Valor (OAV). Este esquema de codificação já era popular em sistemas pioneiros (Shortlife et al. 1975), e atualmente tem um papel importante na iniciativa Semantic Web (W3C 2008), onde é conhecido como o trio Sujeito – Predicado – Objeto (SPO) dentro do Formato de Descrição de Recurso (Resource Description Format - RDF) (Klyne et al. 2004). Esta representação é ainda muito semelhante à forma pela qual o Metatesouro Sistema Unificado de Terminologia Médica (UMLS) e outras fontes de terminologia ligam pares de conceitos de diferentes sistemas de terminologia, através de relações como *mais amplo que*, *mais restrito que*, *parte de*, *mapeado para*, *é uma*, e assim por diante. A Tabela 2 mostra alguns exemplos deste tipo de representação.

**Tabela 2 – Exemplos de representações OAV**

Conceito/Termo 1 (Objeto, Sujeito)	Relação (Atributo, Predicado)	Conceito/Termo 2 (Valor/Objeto)
Aspirina	previne	Infarto_do_Miocárdio
Aspirina	é_um	salicilato
Aspirina	tem_parte	Anel_aromático
Plasma_sanguíneo	Mais_restrito_que	Sangue
Câncer	causa	Perda_de_peso
Célula	tem_parte	Membrana_celular
Medida_contraceptiva	previne	Gravidez
Diabetes_Mellitus	é_uma	Doença_frequente
Diabetes_Mellitus	Tem_prevalência	2.8%
Diclofenaco	Tem_efeito_colateral	Sangramento_gastrointestinal
Difteria	é_uma	Doença_rara
ELM-2	Interage_com	LMO-2
ELM-2	é_uma	proteína
Febre	Sintoma_de	Malaria_Tropica

Cont.



Mão	tem_parte	polegar
Hepatite	Tem_localização	Fígado
Hepatite	Tem_tradução	Hepatitis
Hipertensão	é_um	Fator_de_risco_cardiovascular
Hipertermia	Tem_sinônimo	Febre
Fígado	é_um	Órgão_do_corpo
Solução_NaCl	tem_parte	Cl-Ion
Faringite	Tem_sintoma	Hipertermia
Fumar	causa	Câncer
THC	é_um	Medicamento_Controlado_Schedule_III
polegar	tem_parte	Unha
OMS	Localizado_em	Genebra

Uma das vantagens deste formato triplo torna-se evidente quando examinamos a tabela. Afirmacões simples são representadas de maneira fácil, próxima às expressões da linguagem humana. Uma das desvantagens é que promove confusão na utilização e menção (por exemplo, ao afirmar que a *Febre* é tanto um *sinônimo de Hipertermia* quanto um *sintoma de Inflamação*). O formato triplo também enfrenta dificuldades referentes à formulação de afirmações mais complexas, como em “Em 2008, o diabetes mellitus teve prevalência de 18,3% nos cidadãos americanos com idade igual ou superior a 60 anos”, que precisa ser dividida em grupos de afirmações mais simples para que se encaixem no formato. A Tabela 3 mostra uma possível representação OAV de uma afirmação em que as linhas sucessivas são agrupadas em uma sentença conjuntiva composta. Um inconveniente é que muitos modelos concorrentes deste tipo podem alegar que representam a sentença em questão igualmente bem, o que cria bifurcações. Diferentes grupos executam as traduções necessárias de formas diferentes, o que resulta em sistemas de informação desprovidos de interoperabilidade. Para evitar este efeito restritivo, é necessário um modelo único de representação uniforme.

**Tabela 3 – Representação OAV tripla da sentença complexa: “Em 2008, o diabetes mellitus teve prevalência de 18,3% nos cidadãos americanos com idade igual ou superior a 60 anos”**

Prevalência_1	Ocorrência_de	Prevalência
Prevalência_1	Tem_data	2008
Prevalência_1	Tem_valor	0.183
Prevalência_1	Tem_população	População_1
Prevalência_1	Tem_doença	Diabetes_Mellitus
População_1	Ocorrência_de	População
População_1	Tem_idade_mínima	60
População_1	Tem_habitat	EUA

Outro inconveniente do esquema de representação OAV é que em nenhum caso fica explícito como as afirmações devem ser interpretadas. A afirmação *Fumar causa Câncer*, por exemplo, poderia ser interpretada de tal forma que seu autor acredite que fumar *sempre* (isto é, sem exceção) causa câncer. Porém, também poderia ser interpretada com o sentido de que fumar freqüente, geral ou tipicamente causa câncer, ou ainda, conforme a Rede Semântica UMLS, que a expressão “Fumar causa câncer” tem significado semântico. Sem conhecimento adicional sobre como interpretar as *causas* da relação, não podemos decidir qual a alternativa pretendida em nenhum dos casos. É claro que, em muitas situações cotidianas, os seres humanos se comunicam perfeitamente bem ao utilizar afirmações ambíguas. Isto acontece, porém, porque os seres humanos têm a capacidade de espontaneamente posicionar tais afirmações dentro de um contexto relevante de suposições básicas implícitas. No caso do processamento eletrônico, entretanto, tal conhecimento implícito não existe. É por este motivo que definições lógicas e axiomas expressos em linguagem formal apropriada são necessários para impedir, ou no mínimo restringir, as interpretações ambíguas. Infelizmente, conforme demonstrado pelos exemplos abaixo, a aplicação do rigor da lógica não é apenas muito cara em termos de recursos humanos; é também de tal natureza que não permite, em princípio, a expressão formal de tudo que conhecemos. Ainda podemos, entretanto, capturar uma parcela importante do nosso conhecimento de forma que acreditamos ser indispensável à lógica computacional e à resolução de nossos três problemas: integração, recuperação, e interoperabilidade.

## Representações formais

Com o objetivo de ilustrar como as afirmações ontológicas básicas relacionadas a entidades de determinada área podem ser formuladas utilizando-se recursos lógicos, apresentamos a família de Lógicas Descritivas (LDs) (Baader et al. 2007). LDs são subgrupos de lógica de primeira ordem (LPO). Embora as LDs estejam longe

de expressar tudo aquilo que desejamos de um registro lógico abrangente de ontologias (o que exigiria toda a extensão de LPO), utilizamos esta abordagem pelas seguintes razões:

- As LDs recentemente atingiram um padrão para a representação de conhecimento de áreas no contexto da Semantic Web, com OWL LD, o subtipo LD da Linguagem Ontológica da Web (Web Ontology Language) (OWL) (Patel-Schneider et al. 2004), desenvolvido e padronizado pelo Consórcio da Internet (World Wide Web Consortium) (W3C).

- LDs possuem maior base de usuários, e são apoiadas por diversas ferramentas em software, como o editor Protégé (Bmir 2008). OWL LD também suporta a utilização de dispositivos de raciocínio tais como o Pellet (Sirin et al. 2007) e FaCT++ (Tsarkov et al. 2006), que permitem a verificação algorítmica da consistência de determinados dados, e a dedução de novas afirmações.

- LDs possuem certas propriedades computacionais favoráveis. Por exemplo, muitas são decidíveis, o que significa que existem algoritmos para os quais é garantido que elas informarão algum resultado. Este é o fator que confere às LDs preferência sobre as (indecidíveis) LPO, que são consideravelmente mais expressivas que as primeiras, porém desprovidas de decidibilidade.

- As LDs têm sido cada vez mais empregadas na terminologia biomédica. Após o projeto GALEN, nos anos 1990 (Rector 1997), que pode ser considerado um pioneiro na utilização em larga escala de um formalismo baseado em lógica para a representação e interpretação da área médica, o exemplo atual mais significativo é a terminologia clínica SNOMED CT (IHTSDO 2009), que contém mais de 300.000 classes. LD OWL também é cada vez mais utilizada como linguagem representativa para as ontologias da OBO Foundry (Smith et al. 2007).

A utilização adequada das LDs requer a compreensão de seus blocos fundamentais, representados por termos como "classe", "relação" e "indivíduo", e também o entendimento de como seus símbolos e expressões lógicos constituintes são interpretados. Por exemplo, todas as mãos individuais passadas, presentes e futuras do mundo são ocorrências da classe *Mão*. Relações binárias ("propriedades de objeto", em LD OWL) têm pares de indivíduos por extensão (Patel-Schneider et al. 2004). Por exemplo, o par constituído pelo polegar direito e mão direita do primeiro autor. As classes em LD são sempre distintas dos indivíduos, e classes de classes não são permitidas. As propriedades de objeto LD OWL expressam relações binárias sem nenhuma referência direta ao tempo. Do ponto de vista ontológico – e biológico – este é um grande inconveniente<sup>1</sup>, pois é frequentemente necessário anexar índices temporais a afirmativas a respeito de indivíduos; por exemplo, no sentido de que determinado indivíduo pertence à classe *Embrião* em  $t_1$ , e à classe *Feto* em  $t_2$ . Deve-se ter a preocupação de reconhecer que a mesma expressão pode ser interpretada de diferentes formas em diferentes áreas. Por exemplo, uma afirmação com o sentido de que todas

as mãos têm polegares é limitada à esfera da anatomia humana normal (ou padrão). Claramente não indica se abrange indivíduos lesionados ou mal-formados, ou em estágios embrionários iniciais (Neuhaus et al. 2007, Schulz et al. 2008).

A seguir, ilustramos a sintaxe e semântica da LD através de um conjunto de exemplos de complexidade crescente. Para começar, vamos examinar a classe *Fígado*. Ao introduzirmos esta classe, definimos sua extensão como sendo o conjunto de todos os fígados de todos os organismos, em todas as ocasiões. Na mesma linha, a classe *Órgão\_do\_Corpo* tem como extensão todos os órgãos individuais do corpo, em todas as ocasiões. Para ligar as duas classes, podemos introduzir o conceito fundamental da classificação taxonômica: A classe *Fígado* é uma sub-classe (subtipo) da classe *Órgão\_do\_Corpo*. Na simbologia de LD, isto é expresso pelo operador  $\sqsubseteq$ :

$$Fígado \sqsubseteq Órgão\_do\_Corpo$$

E a relação em questão é comumente descrita como a relação *é um*.

Em contraste, a relação de ocorrência *ocorrência\_de* ( $\in$ ) liga os indivíduos às classes das quais são ocorrências.

Por exemplo, cada fígado individual é uma ocorrência da classe *Fígado*. Assim, o fígado (individual) do primeiro autor deste documento é uma específica *ocorrência\_de Fígado*. É importante destacar que as LDs não permitem que seja expressa, por um lado, a distinção entre a inclusão de um indivíduo em uma classe definida de determinada forma, e, por outro lado, a exemplificação individual de um universo ou tipo. Ambas são representadas através da relação *ocorrência\_de* ( $\in$ ).

Afirmações mais complexas podem ser obtidas pelo uso de operadores e quantificadores. No exemplo a seguir, utilizamos o operador  $\sqcap$  ("e"), e adicionamos uma função quantificada, utilizando o quantificador existencial  $\exists$  ("existe"). A expressão

$$Doença\_Inflamatória \sqcap \exists tem\_localização.Fígado$$

denota a classe de todas as ocorrências que pertencem à classe *Doença\_Inflamatória*, e são posteriormente ligados através da relação *tem\_localização* a alguma ocorrência da classe *Fígado*.

Este exemplo, na verdade, nos dá condições tanto necessárias quanto suficientes para completamente definir a classe *Hepatite*:

$$Hepatite \equiv Doença\_Inflamatória \sqcap \exists tem\_localização.Fígado$$

O operador de equivalência  $\equiv$  nesta fórmula nos diz que: (i) cada ocorrência específica de hepatite é uma ocorrência de doença inflamatória localizada em algum fígado, e também (ii) que todas as ocorrências de doença inflamatória localizadas em algum fígado são

ocorrências de hepatite. Assim, em qualquer situação, o termo à esquerda pode ser substituído pela expressão à direita, sem qualquer perda de significado.

Observe que, quando expressamos uma afirmação de equivalência como esta, a afirmação deve ser verdadeira em todas as ocasiões, sem exceção. Sendo assim, não podemos utilizar este tipo de afirmação para informar, por exemplo, que a hepatite tem o sintoma *febre na maioria (mas não em todos) dos casos*. Logicamente, poderíamos formar a expressão

$$\text{Inflamação} \sqcap \exists \text{tem\_localização.Fígado} \sqcap \exists \text{normalmente\_tem\_sintoma.Febre}$$

e afirmar uma equivalência com Hepatite. Em virtude da interpretação LD do quantificador existencial, entretanto, esta afirmativa implica que para todas as ocorrências da classe *Hepatite* (sem exceção) também existe alguma ocorrência de *Febre*. A palavra *normalmente* no nome da propriedade *normalmente\_tem\_sintoma* pode ser interpretada por seres humanos, mas não tem nenhuma função lógica. Isto claramente não está de acordo com o sentido pretendido.

Tais efeitos lógicos são importantes, já que erros ocorrem quando não são levados em consideração pelos usuários dos formalismos LD. Exemplos abundantes desses erros podem ser encontrados na versão atual de SNOMED CT. Seu conceito *Biópsia\_Planejada* (ID:183993008), por exemplo, é relacionado ao conceito *Biópsia*, conforme abaixo:

$$\text{Biópsia\_Planejada} \sqsubseteq \text{Situação} \sqcap \exists \text{procedimento\_associado. Biópsia} \sqcap \dots$$

Esta expressão afirma que, para cada *biópsia planejada* (supondo que este seja o significado de *Biópsia\_Planejada*), sempre existe pelo menos uma ocorrência efetiva de uma *biópsia*, o que certamente pode não ser a intenção, já que nem todos os planos de *biópsia* se realizam. SNOMED CT inclui também a classe *Prevenção\_ao\_Abuso\_de\_drogas* (ID: 408941008):

$$\text{Prevenção\_do\_Abuso\_de\_drogas} \sqsubseteq \text{Procedimento} \sqcap \exists \text{tem\_foco.Abuso\_de\_Drogas}$$

Esta expressão afirma, de forma absurda, que sempre que se executa uma ação de prevenção ao abuso de drogas existe ocorrência de abuso de drogas.

Estes dois exemplos ilustram a facilidade de se criar afirmações com significados não intencionais ao se utilizar até mesmo as LDs mais simples. A razão pela qual estes exemplos são tão comuns nas terminologias biomédicas atuais é que os desenvolvedores da ontologia são, muitas vezes, especialistas da área sem familiaridade com as complexidades da lógica formal, e dão pouca importância aos princípios do desenvolvimento correto da ontologia. Tais profissionais tendem a guiar-se pela

simplicidade dessas afirmações e, assim, não percebem que sua interpretação lógica contradiz o significado pretendido. As afirmações inválidas resultantes geram deduções inválidas quando utilizadas em raciocínio automatizado.

É claro, entretanto, que alguns usuários de ontologia precisarão utilizar em seus trabalhos para definir classes como *Plano\_de\_Biópsia* ou *Prevenção\_do\_Abuso\_de\_Drogas*. Uma vez que qualquer utilização não-negada de funções existencialmente quantificadas em um formalismo LD corresponde a uma afirmação do tipo “para todo... existe algum...”, devemos recorrer às chamadas restrições de valor, caso desejemos causar o efeito necessário. Isto significa que o quantificador  $\forall$  utilizado em uma função quantificada é utilizado para especificar a variação permitida para determinada relação. Poderíamos, assim, (corretamente) afirmar o seguinte:

$$\text{Plano\_de\_Biópsia} \sqsubseteq \text{Plano} \sqcap \forall \text{tem\_realização.Biópsia}$$

Em linguagem simples, esta expressão afirma que um plano de *biópsia* é um plano que – se realizado – pode ser realizado apenas por alguma ocorrência de *Biópsia*. Em contraste com as afirmações existenciais simples, isto não afirma que uma *Biópsia* deve existir para cada *Plano\_de\_Biópsia*. Construções similares são necessárias para outras entidades realizáveis, tais como funções, posições, ou disposições (Grenon 2003).

Ao utilizarmos o quantificador universal  $\forall$ , entretanto, passamos de dialetos LD simples, porém escalonáveis, como *EL* (Baader et al. 2007), para LDs com uma complexidade computacional que oferece graves problemas para ontologias de grande porte, como a SNOMED CT. É ainda mais complicado definir classes como *Prevenção\_do\_Abuso\_de\_Drogas* com o rigor local adequado. Aqui precisamos dizer que, se tal procedimento for aplicado, isso causa um estado no organismo que impede que este participe de *Abuso\_de\_Drogas*. Assim, para expressar a informação adequadamente precisamos introduzir o operador de negação  $\neg$  conforme abaixo:

$$\text{Prevenção\_ao\_Abuso\_de\_Drogas} \sqsubseteq \text{Procedimento} \sqcap \exists \text{tem\_participante.Pessoa} \sqcap \exists \text{causas. (Estado} \sqcap \exists \text{tem\_participante. (Pessoa} \sqcap \exists \text{participa\_de.} \neg \text{Abuso\_de\_Drogas))}$$

Nesta definição a classe *Pessoa* ocorre duas vezes, mas não fica claro se essas duas ocorrências são idênticas – como deveriam ser. Não há LD capaz de expressar o fato de que elas são idênticas, o que exigiria todos os poderes de expressão da LPO, ultrapassando a esfera da decidibilidade.

Outros casos de termos médicos que excedem a capacidade de expressão da lógica descritiva decidível incluem expressões que envolvem “sem”, como em “concussão cerebral sem perda de consciência”, conforme discutido em (Bodenreider et al. 2004, Ceusters et al. 2007, Schulz et al. 2008). São altamente importantes e relevantes na medicina. Sua representação, no entanto,

é complexa, não somente devido às suas exigências de construtores lógicos expressivos, mas também devido à dificuldade de se chegar a uma conclusão unânime sobre seu significado, levando-se em consideração suposições tácitas (novamente relacionadas ao tempo).

Os exemplos acima claramente demonstram o dilema das representações baseadas em lógica: Se o objetivo é logicamente codificar e classificar grandes sistemas terminológicos como o SNOMED CT (Baader et al. 2006), então o conjunto de construtores permitidos deve ser limitado, já que restrições e negações de valor levam à intratabilidade computacional. Alguns (Rector et al. 2008), entretanto, enfatizam que é importante incluir construções computacionalmente mais amplas, de forma a não impedir representações adequadas da área. Uma estratégia alternativa é distinguir as construções contidas dentro da terminologia de sua utilização em contextos específicos, onde a negação e outros termos (como “após\_exame”) sejam adequadamente utilizados.

## Categorias de representação de áreas

Conforme já deve estar claro, muitas vezes não é possível representar fielmente aspectos importantes do conhecimento biomédico através dos formalismos da representação computável, lógica, das áreas. Muitos tipos de afirmação exigem outras formas de representação. Propomos, assim, a distinção entre diferentes categorias de representações de áreas, que exigem tipos diferenciados de tratamento, mesmo que sejam muitas vezes tratados como semelhantes dentro das ontologias formais. Nosso interesse em manter essas categorias em separado é destacar o fato de que cada representação exige seus formalismos próprios, com semântica própria, e que o uso inadequado de formalismos de representação não diferenciados leva a resultados indesejados. Como resultado de nossa discussão, esperamos contribuir para um entendimento mais claro do que as ontologias formais podem ou não realizar na área biomédica.

## Representação léxico-semântica

Utilizamos “representação léxico-semântica” para nos referirmos a tesouros, dicionários semânticos e artefatos similares, que enfocam os significados das expressões encontradas na linguagem natural. Tipicamente, abordam tanto o fato de que um verbo pode ter dois ou mais significados (como ilustrado, por exemplo, pela polissemia de termos como “fratura” ou “envenenamento”), como o fato de que um significado pode ser expresso por um ou mais verbetes (por exemplo, a sinonímia entre “hipertermia” e “febre”). Podem, também, conter traduções de palavras ou termos. Tesouros e léxicos semânticos podem, ainda, conter relações semânticas entre os verbetes individuais, como *mais\_amplo\_que* ou *mais\_restrito\_que*. WordNet (Fellbaum 1998), MeSH e grande parte do Metatesouro UMLS (NLMB 2008) são exemplos de tais sistemas de representação, que têm ampla tradição na biblioteconomia, com recuperação de literatura como caso de uso amplamente aceito.

A questão de como as relações léxico-semânticas como a sinonímia devem ser corretamente expressas não é, na realidade, um assunto que deva ser tratado pelas ontologias. As ontologias se relacionam com entidades reais de forma independente da linguagem. Descrevem tais entidades e as relações entre elas, mas não as descrevem na linguagem humana, isto é, em seus termos e expressões relacionados. Assim, como até mesmo a linguagem humana pode ser utilizada para descrever as entidades na realidade (além da definição lógica formal), o objetivo de tais descrições não é descrever a linguagem em si. Desta forma, relações como *mais\_abrangente\_que* ou *mais\_restrito\_que*, que são relações de subclassificação semanticamente arbitrárias (OBRST 2006) que caracterizam o tesouro MeSH, são substancialmente diferentes da relação de subclasse (*é\_um*) que define a estrutura taxonômica de uma ontologia adequadamente construída. Por exemplo, no MeSH encontramos tanto *Plasma mais\_restrito\_que Sangue* e *Sangue\_Fetal mais\_restrito\_que Sangue*, embora, de um ponto de vista ontológico, as relações aqui envolvidas sejam fundamentalmente diferentes. No primeiro caso, estamos lidando com uma relação de parcialidade (*parte\_de*), mas, no segundo caso, a relação é do subtipo (*é\_um*). A diferença pode não importar no contexto relevante, já que a relação *mais\_restrito\_que*, mesmo sendo semanticamente mal definida, se encaixa perfeitamente bem às necessidades atuais da classificação e recuperação literárias. Os artigos sobre plasma sanguíneo são tão relevantes para uma pesquisa sobre “sangue” quanto artigos sobre sangue fetal.

Os problemas surgem no presente estágio da recuperação de informação, quando é proposta a “ontologização” do MeSH através do simples mapeamento de todas as relações *mais\_restrito\_que* para relações de classificação taxonômica (Soualmia et al. 2004) como em *Plasma*  $\sqsubseteq$  *Sangue* e *Sangue\_Fetal*  $\sqsubseteq$  *Sangue*. Se, por um lado, o resultado é um gráfico de subclassificação aparentemente perfeito que pode ser facilmente processado pelas ferramentas LDs padrão, este exercício, mais uma vez, demonstra o típico caso da criação não intencional de modelos, já que ignora o verdadeiro significado da classificação. O resultado traduz-se em erros como classificar o plasma como um tipo de sangue.

Enquanto as relações léxico-semânticas têm determinadas características em comum com as relações ontológicas entre as entidades da realidade, a construção de uma ontologia a partir de um tesouro requer diversas suposições adicionais, como as relacionadas à quantificação, por exemplo. Portanto, qualquer processo automatizado de conversão não consegue oferecer nada além de um esboço rudimentar, que exige cuidadosa elaboração manual e seleção antes que possa ser seriamente levado em consideração para fins de inferência (Schulz et al. 2001).

Embora encaremos os léxicos ou listas terminológicas como excluídos do reino da ontologia formal, devemos enfatizar que, virtualmente, todas as formas de aplicação da ontologia requerem uma ligação entre as classes ontológicas e os componentes léxicos. Entretan-



to, defendemos que essas duas questões sejam tratadas pelos dois artefatos separados das ontologias formais, por um lado, e pelas representações léxico-semânticas, por outro.

## Representações de tipos de entidades

O realismo científico postula a existência de uma realidade objetiva que pode ser estudada pela ciência, e sobre a qual podemos descobrir verdades (Boyd 2002). Uma teoria científica adequada e, portanto, uma ontologia adequada, contém, por exemplo, afirmações no sentido de que entidades exemplificando determinada classe equivalem, em determinadas relações, a entidades exemplificando outra classe. É importante ressaltar que essa descrição envolve reconhecimento explícito de que todas as afirmativas científicas podem se basear em erros, e devem, poder ser revisadas em cada estágio. Diferentes teorias da realidade foram propostas – por exemplo, teorias baseadas em abordagens tri- e quadri-dimensionalistas, mas o realismo científico assim descrito é compatível com uma ampla gama de tais teorias. Se, por um lado, a visão realista ainda é controversa e não compartilhada por todos os desenvolvedores da ontologia (Smith et al. 2006), possui, por outro lado, diversas vantagens práticas. Assim, por exemplo, permite que se tenha uma visão de que as ontologias oferecem um princípio fundamental para as afirmações axiomáticas acerca de relações simples entre os tipos de entidade mais básicos em termos científicos, que podem, então, ser considerados como certos em trabalhos maiores e mais complexos. Exemplos de tais afirmações são “células têm membranas”, “corações têm câmaras”, “todo caso de hepatite localiza-se em um fígado”, “todo comprimido de aspirina contém salicilato”, e assim por diante.

É útil produzir artefatos que ofereçam raciocínio automático computacionalmente receptivos baseados em tais afirmativas, conforme demonstrado acima. Entretanto, não é assim que funciona no caso da tentativa de se produzir teorias formais que tenham por objetivo caracterizar uma área da realidade. Na engenharia ontológica prática, esses dois objetivos têm de ser conciliados. O histórico de utilização da Ontologia Genética apóia a tese de que características da realidade podem, muitas vezes, ser suficientemente bem representadas, mesmo por meio de uma lógica relativamente simples. Entretanto, como fica claro após as discussões a respeito de LDs acima, devemos sempre nos lembrar que, em muitos casos, tais formalismos não possuem a riqueza necessária à criação de definições completas. A expressividade necessária entra em conflito com a necessidade de se construir modelos que possam ser manipulados computacionalmente. Deve-se aceitar, portanto, que as ontologias (assim como as teorias científicas) oferecem representações apenas parciais da realidade. Elas afirmam o que é considerado como verdade sobre todas as ocorrências de determinadas classes: “Não há hepatite fora do fígado”; “não há solução de NaCl sem íons de cloreto”; “não há célula sem membrana celular”. Porém, é muito claro que tais declarações constituem apenas uma pequena parte do

conhecimento que pode ser necessário para a abrangência adequada de determinada área. Conforme Rector (2008) afirma, “Há muito poucos componentes interessantes do conhecimento que sejam verdadeiramente ontológicos neste sentido mais restrito”. Entretanto, é evidente que tais componentes têm importância crucial, pois formam a base de todo raciocínio, tanto de seres humanos quanto de aplicativos de computador.

Além disso, até agora foi amplamente ignorado que este tipo de representação de área (declarações sobre o que é verdadeiro de todas as ocorrências de uma classe) também está presente em diversos artefatos raramente identificados como ontologias. A UniProt, um grande repositório (base de dados) central de dados de proteínas (UniProt 2008), é um exemplo típico. Sob análise ontológica, a maior parte de seu conteúdo descreve tipos de proteínas (e não indivíduos), em termos do que é universalmente verdadeiro para absolutamente cada uma das moléculas de proteína deste tipo. Sendo assim, consideramos este tipo de representação, também, como sendo de natureza essencialmente ontológica.

## Representação de conhecimento prévio

O termo “conhecimento prévio”, conforme utilizado por Rector (2008), abrange o conhecimento padrão, presuntivo, e probabilístico. Refere-se a todos os tipos de sentenças que supostamente sejam ao menos geralmente (mas não necessariamente universalmente) verdadeiras em alguma área e contexto. Esse conhecimento é, tradicionalmente, transmitido por livros científicos de forma altamente dependente do contexto, muitas vezes fazendo uso de declarações prototipais; por exemplo, referindo-se às relações entre as doenças, sinais e sintomas, ou entre efeitos colaterais e medicamentos, que são expressas em termos de probabilidades qualitativas.

É a familiaridade com esse conhecimento prévio, e não a familiaridade com o conhecimento que pode ser transmitido pela utilização de ontologias formais, que distingue um especialista de um novato, assim como marca a distinção em contexto entre um livro comum e um dicionário. Os exemplos abaixo demonstram como as abordagens da ontologia formal e os formalismos da representação lógica atingem seus limites quando se trata de representar esse tipo de conhecimento. A utilização de formalismos baseados em LDs, mesmo em descrições simplificadas de conhecimento prototipal, levaria a resultados falhos. Existem outros formalismos lógicos que são capazes de expressar esse tipo de conhecimento, mas, novamente, tais formalismos são computacionalmente caros, se não indecifráveis.

## Conhecimento padrão

Um exemplo de conhecimento prévio é o conhecimento padrão (Rector 2004, Hoehndorf et al. 2007), que é o conhecimento relacionado àquilo que pode ser considerado geralmente verdadeiro na ausência de provas contrárias. A LD não nos oferece meios de afirmar o que é geralmente verdadeiro. Especificamente em relação à anatomia geral versus a anatomia clínica, (Smith et al.,

2005) poderíamos querer dizer que, por exemplo, as mãos normalmente têm polegares. Uma afirmação do tipo

$$\text{Mão} \sqsubseteq \exists \text{tem\_parte\_própria.Polegar}$$

não descreveria isto de forma adequada. Ela afirma que todas as mãos têm um polegar, e exclui a possibilidade de haver mãos sem polegar; isto é, exclui mãos não-prototípicas (por exemplo, após terem sofrido um acidente).

## Meta classes

Outras sentenças de conhecimento prévio são meta-sentenças a respeito de classes. São verdadeiras quando vistas como afirmações a respeito de classes como um todo, mas tornam-se falsas quando encaradas como afirmações a respeito de ocorrências. O ponto de vista da LD é que todas as sentenças a respeito de classes *são* sentenças a respeito dos conjuntos de ocorrências correspondentes. Ao ignorar isto, sentenças de classificação aparentemente óbvias, como:

$$\begin{aligned} \text{Diabetes\_Mellitus} &\sqsubseteq \text{Doença\_Frequente} \\ \text{Diabetes\_Mellitus\_Relacionada\_a\_má\_nutrição} &\sqsubseteq \text{Diabetes\_Mellitus} \end{aligned}$$

levariam à falsa conclusão que

$$\text{Diabetes\_Mellitus\_Relacionada\_a\_má\_nutrição} \sqsubseteq \text{Doença\_Frequente}$$

O problema aqui é está em erroneamente considerar as propriedades de determinado tipo relacionadas à população, tais como *frequência*, como sendo propriedades inerentes a subtipos desse tipo. O símbolo  $\sqsubseteq$  (*é uma*) acima é utilizado em dois sentidos logicamente distintos, sendo que apenas um deles é ratificado pelas LDs, e o resultado *é uma sobrecarga* foi identificado como erro típico que ocorre ao se construir ontologias de forma desprovida de embasamento (Guarino 1999, Welty & Guarino 2001, Smith et al. 2004).

## Aptidões

Codificar fatos não triviais em ontologias formais pode exigir construções adicionais complicadas, tais como a adição de representações de aptidões para passar informação a respeito de potencialidades. É importante observar que aptidões podem existir sem nem jamais serem percebidas, e mesmo que não consigamos apontar as condições precisas nas quais tal disposição é realizada (Jansen 2007). Um medicamento analgésico, por exemplo, é uma substância que tem aptidão para tratar dor. Porém, irá realizar tal aptidão apenas quando administrado de determinada forma, para certo tipo de paciente. Podemos representar a classe de processos de tratar (um paciente com) dor através de:

$$\text{Tratar} \sqcap \exists \text{tem\_participante.Dor}$$

Podemos, então, representar a classe de aptidões realizadas quando a dor é tratada:

$$\text{Aptidão} \sqcap \forall \text{tem\_realização.}(\text{Tratar} \sqcap \exists \text{tem\_participante.Dor})$$

A definição abaixo declara que um Medicamento\_Analgésico é uma substância à qual esta aptidão é inerente:

$$\begin{aligned} \text{Medicamento\_Analgésico} &\sqsubseteq \text{Substância} \sqcap \exists \text{portadora\_} \\ &\text{de.}(\text{Aptidão} \sqcap \forall \text{tem\_realização.}(\text{Tratar} \sqcap \exists \text{tem\_} \\ &\text{participante.Dor})) \end{aligned}$$

Tais construções podem fortemente afetar a escalabilidade de uma implementação ontológica, uma vez que um maior conjunto de tais expressões - como, por exemplo, na representação da farmacodinâmica das substâncias - não pode ser eficientemente manipulada pelos algoritmos de raciocínio atuais.

## Dados no contexto

O grupo de afirmações científicas e clínicas não é restrito à expressão de afirmações padrão e características distribucionais. Inclui, também, afirmações incertas como, por exemplo, as referentes ao efeito de um medicamento no tratamento de determinada doença, ou à existência de um suposto fator de risco para determinada condição. Pelas razões mencionadas, a codificação de tais afirmativas nas ontologias formais pode ser extremamente complicada, e é, acima de tudo, realmente questionável se tais afirmações deveriam ser incluídas numa ontologia formal.

Por exemplo, uma ontologia está sendo criada no contexto do projeto @neurIST, da União Européia, como base para a mediação semântica e integração de dados na área de aneurismas cerebrais e sangramentos sub-aracnoidais (Boeker et al. 2007). Os dados dentro do projeto têm origem em diversas fontes, e demonstram um alto grau de fragmentação e heterogeneidade, tanto em formato quanto em escala. A ontologia precisa representar todos os tipos relevantes de entidades, e também respeitar diversos pontos-de-vista a respeito dessas entidades, da parte de disciplinas como a medicina ou epidemiologia, engajadas em estudá-las. Para fazer justiça a todos esses aspectos, a ontologia aplica sentenças relacionadas a aptidões na formulação de definições de classe, e divide-se em duas partes: (i) a ontologia, no sentido literal da palavra, e (ii) um conjunto de artefatos representacionais que capturam conhecimento específico do contexto acerca de determinados fatos, por exemplo, fatores de risco em contextos clínicos. (Uma abordagem semelhante também é o objetivo da Ontologia de Investigações Biomédicas (Ontology of Biomedical Investigations-OBi) (OBi 2008)). Na ontologia @neurIST, a classe *Doença\_Hipertensiva* é uma subclasse de *Processo\_ou\_Estado\_Biológico*, que é associada a *Pressão\_Sanguínea\_Elevada* e causa alguma *Aptidão\_para\_Ruptura*, isto é, uma tendência de que o aneurisma se rompa. Esta aptidão é, então,

conectada à classe (e, ao fazê-lo, identificada como um) *Fator\_de\_Risco para Ruptura\_de\_Aneurisma*, no sentido de que esta última classe também é definida de tal forma que suas ocorrências causam algumas ocorrências de *Aptidão para Ruptura*.

$$\text{Aptidão\_Para\_Ruptura} \equiv \text{Predisposição\_à\_Doença} \sqcap \\ \forall \text{ tem\_realização.Ruptura\_do\_Aneurisma}$$

$$\text{Fator\_de\_Risco\_Para\_Ruptura\_do\_aneurisma} \sqsubseteq \text{Fator\_de\_} \\ \text{Risco} \sqcap \exists \text{ causa.Aptidão\_para\_Ruptura}$$

A seguinte afirmação é crucial para o estudo de aneurisma, mas transgride os limites da ontologia formal. É incompleta, no sentido de que as restrições contextualmente definidas, e que tornam esta afirmação válida, estão ausentes:

$$\text{Doença\_Hipertensiva} \sqsubseteq \text{Fator\_de\_Risco\_para\_Ruptura\_de\_} \\ \text{Aneurisma}$$

A sentença acima afirma que doença hipertensiva é normalmente um fator de risco, o que é pouco convincente. Por outro lado, a doença hipertensiva certamente é um fator de risco para aneurisma cerebral. Assim, o que queremos dizer é que existe uma correlação forte entre os dois, e esta afirmação é de importância fundamental (mas existem, logicamente, outros fatores de risco também).

Estes exemplos demonstram os tipos de passos que teriam de ser tomados para que uma estrutura de LD fosse expandida, de tal forma que abrangesse certos tipos de conhecimento prévio, beneficiando-se, assim, da vantagem do apoio do raciocínio LD, sem incorrer no risco de modelos não intencionais.

Entretanto, a dificuldade de se representar todas as suposições ocultas implícitas no conhecimento prévio (e os problemas de desempenho que resultam da utilização da lógica complexa necessária) pode sugerir que utilizemos uma representação tripla muito mais simples, como mencionado na seção introdutória, e desenvolvamos dispositivos especiais de raciocínio para ela. Por outro lado, poderíamos lançar mão de uma variedade maior de artefatos de representação de conhecimento, tais como a lógica padrão (Reiter 1980), *frames* (Minsky 1974), F-logic (Kifer et al. 1989), e diversos outros tipos de extensões LD computacionalmente caras (Baader 2007, ch. 6). Os artefatos de representação de conhecimento resultantes, entretanto, não são ontologias formais, no sentido com o qual o termo é utilizado. Ainda assim, podemos reutilizar as classes formalmente definidas em uma ontologia como símbolos nesses formalismos, de acordo com as linhas gerais descritas nos exemplos acima.

## Representação de indivíduos

Se, por um lado, os três primeiros tipos de representação descritos acima fazem generalizações a respeito

de todas as entidades de determinado tipo, grande parte da medicina envolve descrições de entidades individuais, tais como um tumor, exame laboratorial ou tratamento específicos, ou a ocorrência de uma doença específica em determinado grupo de pacientes. As disciplinas de epidemiologia e saúde pública lidam com entidades políticas e geográficas, como o *Brasil*, *Nova Orleans*, *as ilhas do Pacífico Sul*, ou a *região superior do Rio Negro*.

Sentenças a respeito de fatos individuais podem ser expressas de maneira direta nos termos de LD como instâncias de classes correspondentes, ou, em outras palavras, como as chamadas afirmações de caixa-A (sendo que a letra A significa afirmativas a respeito de indivíduos), em contraste com o componente caixa-T de LDs que capturam o que é chamado de “conhecimento terminológico” (ou, talvez, melhor definido por “conhecimento pertencente aos tipos”). Considere, por exemplo:

$$\text{Hepatite\_162726} \in \text{Hepatite}$$

que afirma que uma determinada doença é uma ocorrência de hepatite.

Uma sentença de interação molecular como “Lmo-2 interage com Elf-2”, conforme publicado em um artigo científico é, primeiramente, uma afirmativa a respeito de determinados indivíduos, especificamente duas ocorrências de porções de Lmo-2 e Elf-2 (ou coleções moleculares), que comprovadamente mostraram alguma interação em um determinado ensaio (Schulz et al. 2008).

Assim, afirmamos certo evento de interação onde as duas porções de substâncias sob análise participam:

$$\text{Lmo-2.7760102} \in \text{Porção\_de\_Lmo-2}$$

$$\text{Elf-2.776010} \in \text{Porção\_de\_Elf-2}$$

$$\text{Interação.725322} \in \text{Interação}$$

$$\text{tem\_participante (Interação.725322, Lmo-2.7760102)}$$

$$\text{tem\_participante (Interação.725322, Elf-2.776010)}$$

Há áreas, como a geografia, em que indivíduos - e não classes - constituem os alvos principais de conhecimento. Qualquer descrição detalhada de divisões geográficas ou políticas que pudesse ser do interesse, por exemplo, da epidemiologia ou saúde pública, é abundante em referências a entidades particulares que exemplificam apenas um pequeno número de classes (SMITH et al. 2005). Por exemplo, pode-se criar uma completa divisão política dos EUA com base em quatro níveis agrupados (com uma ocorrência de países, com 50 ocorrências de estados, com 3.077 ocorrências de condados, e com mais de 50.000 ocorrências de municípios) (ver também entidades geográficas em GAZ CONSÓRCIO PADRÃO DE GENÔMICA (GENOMICS STANDARD CONSORTIUM 2008)). Observe a diferença em representação comparando-se às divisões anatômicas na Tabela 4.

**Tabela 4 – Exemplos de partonomia em geografia e anatomia**

<i>Orlando</i> $\in$ <i>Município</i>	<i>Polegar</i> $\sqsubseteq$ <i>Dígito</i>
<i>Condado de Orange</i> $\in$ <i>Condado</i>	<i>Mão</i> $\sqsubseteq$ <i>Parte_do_Corpo</i>
<i>Flórida</i> $\in$ <i>Estado</i>	<i>Extremidade_superior</i> $\sqsubseteq$ <i>Membro</i>
<i>EUA</i> $\in$ <i>País</i>	<i>Corpo</i> $\sqsubseteq$ <i>Estrutura_Anatômica</i>
$\langle \textit{Orlando}, \textit{Condado de Orange} \rangle \in \textit{parte\_de}$	<i>Polegar</i> $\sqsubseteq \exists \textit{ parte\_da.Mão}$
$\langle \textit{Condado de Orange}, \textit{Flórida} \rangle \in \textit{parte\_de}$	<i>Mão</i> $\sqsubseteq \exists \textit{ parte\_da.Extremidade\_superior}$
$\langle \textit{Flórida}, \textit{EUA} \rangle \in \textit{parte\_de}$	<i>Extremidade_Superior</i> $\sqsubseteq \exists \textit{ parte\_do.Corpo}$

Este exemplo demonstra que afirmações a respeito de classes diferem formalmente de afirmações a respeito de indivíduos. As relações empregadas, no entanto, são as mesmas, porque as LDs não permitem a existência de relações especiais entre classes. A ligação lógica das classes sempre exige a utilização de quantificadores, que não são necessários em afirmações que relacionam indivíduos. Isto explica por que, antes de qualquer representação baseada em lógica, deve-se esclarecer se as entidades sob análise são classes ou indivíduos. Isto não é comum, porém, no campo específico da biologia molecular. Assim, nosso exemplo de afirmação “*Lmo-2 interage com Elf-2*” pode ser perfeitamente bem entendido como uma sentença universal a respeito da classe das *moléculas de Lmo-2*, e, assim, como a expressão de um conhecimento de aptidão, no sentido de que:

Todas as moléculas de *Lmo-2* têm aptidão para interação com moléculas de *Elf-2*.

Há bons argumentos a favor das duas interpretações. Assim, a ambigüidade não pode ser desfeita sem que, primeiramente, seja analisado o contexto no qual a afirmação se dá.

Na prática, a fronteira indivíduo/classe é frequentemente definida de forma idiossincrática. Por exemplo, os registros do UniProt são feitos de forma a denotar “ocorrências” da classe proteína. Um profissional de informática poderia afirmar que esta escolha de terminologia é motivada, principalmente, pela visão que um modelador tem de determinada área: “Decidir se um dado conceito é uma classe em uma ontologia ou uma ocorrência individual depende de quais são as aplicações potenciais da ontologia”. (Noy & McGuinness 2001). Acreditamos, no entanto, que nenhuma arbitrariedade deva existir na distinção entre essa célula específica nesse tubo de ensaio específico aqui e agora (ocorrência), e uma *Célula* (classe). Além disto, incentivar a suposição de que exista tal arbitrariedade pode levar a uma bifurcação de representações que dificultarão a própria interoperabilidade que as ontologias de recursos de dados deveriam apoiar.

Na verdade, defendemos que uma análise ontológica formal apenas pode ser coerente com base num conceito da distinção entre indivíduos e classes como sendo a obtenção de distinção inalterável por parte das entidades em si. Indivíduos, por um lado, existem no espaço e no tempo; não têm relação de classificação entre si; podem ser chamados por nomes próprios e (em muitos casos) fotografados. As classes, por outro lado, não existem no espaço e no tempo; têm relação de classificação entre si; e podem ser chamadas por substantivos comuns. O fato de uma entidade ser particular, ou uma classe, ou um tipo, não se trata de escolha por parte do modelador. De acordo com nossa experiência, casos controversos que parecem sugerir a existência desta opção sempre revelam ambigüidades quando melhor examinados. Alguns defensores da opinião de que o gene humano *MPDU-1* é uma ocorrência da classe *Gene* referem-se aos genes como ocorrências de entidades de conteúdo de informação, como no OBI (2008). A mesma entidade de informação genética pode ser codificada em diferentes moléculas ácidas nucléicas, da mesma forma que um texto pode ser disseminado através de muitas cópias. Outros, entretanto, defendem que o gene humano *MPDU-1* não é uma ocorrência, e sim uma subclasse da classe *gene*; estão, assim, referindo-se a genes como tipos de seqüências macromoleculares, cujas ocorrências são seqüências de nucleotídeos replicadas nas células do nosso corpo.

Como vimos anteriormente na seção a respeito de conhecimento prévio, referência implícita a indivíduos é a base de sentenças probabilísticas típicas. A seguinte sentença exemplifica o que acabamos de dizer: “Em 2000, a prevalência mundial de diabetes mellitus foi de 2,8%”. Temos aqui duas classes, que são: *Humano* e (*caso de*) *Diabetes\_Humano*. Ambas as classes têm cardinalidade (valor inteiro), e a prevalência é dada pelo quociente entre as duas. A prevalência não é, assim, característica da doença, e sim da população de indivíduos que têm um caso da doença. Aqui, ampliamos a notação da LD ao simbolizar a cardinalidade da extensão de uma classe (por exemplo, o número de ocorrências) ao colocar o nome da classe entre “||”.



$Humano \sqsubseteq Objeto$

$Humano\_Diabético \equiv Humano \sqcap \exists portador\_de.Diabetes\_Mellitus$

$|Humano\_Diabético|/|Humano| = 0.028$

Isto demonstra que o conhecimento prévio probabilístico poderia ser expresso por caixas LD – A, ampliadas por operadores aritméticos (referindo-se a indivíduos). Isto não está, assim, dentro do escopo das ontologias formais, apenas nas abordagens alternativas, como as extensões caixa-T probabilísticas (Koller 1997, Klinov 2008). Além do mais, tampouco pode ser expresso pelas LDs atualmente disponíveis.

## Comentários e Conclusão

A disciplina da representação do conhecimento evoluiu no contexto da pesquisa da inteligência artificial, com o propósito de possibilitar que computadores tirem novas conclusões a partir de dados e informações existentes. Quando o termo “ontologias” se tornou popular na informática nos anos 90, foi muitas vezes considerado um novo nome para algo que já existia – os artefatos de representação do conhecimento. Entretanto, duas linhas de pesquisa se desenvolveram desde então, demonstrando a necessidade de uma metodologia mais baseada em princípios.

Primeiramente, a Lógica Descritiva (LDs) foi desenvolvida para ser fragmentos computáveis da Lógica de Primeira Ordem (LPO), que fossem suficientemente expressivos para permitir a formulação de afirmações a respeito de classes de indivíduos, bem como suas relações, de tal forma que novos teoremas pudessem ser automaticamente deduzidos. Isto necessitou uma semântica bem definida, exigindo uma divisão em classes e indivíduos; também exigiu uma descrição formal de classificação e da quantificação de papéis.

Enquanto nas representações mais primitivas, do tipo rede semântica, como o Metatesouro UMLS, sentenças como “aspirina é um salicilato”, “aspirina contém um anel aromático” e “aspirina impede infarto do miocárdio” parecem ser muito semelhantes, tentativas de representação mais formal revelam diferenças fundamentais. Na LD, a primeira sentença é direta, e não exige qualquer relação além daquela de subclasse; a segunda exige uma expressão quantificada de função; e a terceira não pode nem mesmo ser adequadamente representada.

Em segundo lugar, paralelamente à evolução das linguagens representacionais como OWL, os filósofos e cientistas computacionais confrontaram a disciplina experiente da ontologia filosófica com as exigências da sociedade de informação moderna, e criaram a disciplina da ontologia aplicada (Guarino 1998). A biomedicina tornou-se um laboratório para a convergência de LDs e ontologia aplicada. A iniciativa OBO Foundry e, cada vez mais, as atividades de reestruturação da SNOMED CT, são testemunhas disto.

Podemos, agora, resumir os resultados deste estudo através da delimitação aproximada de quatro tipos de sentença que apresentamos acima, que são: (i) representação léxico-semântica, (ii) representação de tipos de entidades, (iii) representação de conhecimento prévio, e (iv) representação de indivíduos.

(I) Estes são os tipos de sentenças que encontramos em grande parte do UMLS, assim como no WordNet e artefatos semelhantes, que se esforçam para representar o componente terminológico de uma área. Isto é feito através de relações como sinonímia, polissemia, mais abrangente, mais restrito; e são retirados dos reinos dos tesouros e léxicos semânticos. Alegamos que essa abordagem é útil para a recuperação de informação, mas não para inferência ou integração do conhecimento.

(ii) No extremo oposto estão os tipos de sentenças que encontramos em ontologias formais formuladas em termos de LD, onde o rigor formal e o poder de inferência são alcançados à custa de limitações na expressividade em diversas dimensões. Tais restrições podem não conseguir alcançar as exigências mínimas daqueles usuários que sempre esperam da ontologia de uma área mais que um simples repositório de verdades básicas. Por outro lado, mesmo os truismos podem ter um papel valioso como base para a formulação mais adequada de outros tipos de sentenças, especialmente no contexto dos sistemas de raciocínio.

(iii) Este grupo de sentenças constitui o que chamamos “conhecimento prévio”, uma questão de associação livre entre as classes, que não pode ser expressa pelo esquema “para todo... algum”, típico das LDs. Essas sentenças podem, até certo grau, ser “ontologizadas” pela introdução de classes de aptidão. Entretanto, sua introdução ocorre à custa de um aumento na complexidade. Existem outras abordagens da representação de conhecimento prévio, incluindo a lógica padrão (Reiter 1980), *frames* (Kifer et al. 1989), e diversos tipos de extensões LD computacionalmente caras (Baader 2007, cap. 6). Não se pode fazer uma recomendação geral a respeito de qual dessas - ou de outras - alternativas seria adequada: Isto depende grandemente da área de aplicação específica, e do caso de utilização específico para o qual os serviços de raciocínio são necessários.

(iv) O último conjunto de sentenças refere-se à representação de indivíduos. Isto poderia ser encarado como um pequeno problema, por exemplo, na biologia de leveduras, mas é de grande importância em áreas como a medicina, que se relaciona com o registro de informações a respeito de seres humanos. Mostramos, por exemplo, que sentenças probabilísticas a respeito de prevalência de doenças não são afirmações a respeito de classes, e sim a respeito de indivíduos.

A Tabela 5 recapitula os exemplos dados na Tabela 2 no início do artigo, e atribui cada um deles a uma das diferentes categorias de conhecimento que apresentamos acima.

**Tabela 5 – Afirmações em estilo Metatesauro UMLS (tabela mrrel) e categorias de representação de área correspondentes**

Conceito/Termo 1 (Objeto, Sujeito)	Relação (Atributo, Predicado)	Conceito/Termo 2 (Valor/Objeto)	Categoria de representação de Área
Aspirina	previne	Infarto_do_Miocárdio	BK
Aspirina	é_um	salicilato	ONT
Aspirina	tem_parte	Anel_aromático	ONT
Plasma_sanguíneo	Mais_restrito_que	Sangue	LS
Câncer	causa	Perda_de_peso	BK
Célula	tem_parte	Membrana_celular	ONT
Medida_contraceptiva	previne	Gravidez	BK
Diabetes_Mellitus	é_uma	Doença_frequente	BK
Diabetes_Mellitus	Tem_prevalência	2.8%	BK
Diclofenaco	Tem_efeito_colateral	Sangramento_gastrointestinal	BK
Difteria	é_uma	Doença_rara	BK
ELM-2	Interage_com	LMO-2	BK, INS
ELM-2	é_uma	Proteína	ONT
Febre	Sintoma_de	Malaria_Tropica	BK
Mão	tem_parte	polegar	ONT
Hepatite	Tem_localização	Fígado	ONT
Hepatite	Tem_tradução	Hepatitis	LS
Hipertensão	é_um	Fator_de_risco_cardiovascular	BK
Hipertermia	Tem_sinônimo	Febre	LS
Fígado	é_um	Órgão_do_corpo	ONT
Solução_NaCl	tem_parte	Ion_Cloreto	ONT
Faringite	Tem_sintoma	Hipertermia	BK
Fumar	causa	Câncer	BK
THC	é_um	Medicamento_Controlado_Schedule_III	BK
polegar	tem_parte	Unha	ONT
OMS	Localizada_em	Genebra	INS

BK = conhecimento prévio, INS = ocorrências, LS = representação léxico semântica, ONT = nível ontológico

Nossas distinções coincidem, até certo grau, com aquelas propostas pela OBRST (2006) no Espectro Ontológico (Ontology Spectrum). Nossa primeira categoria corresponde à sua “taxonomia e tesouros ineficazes”, e, a segunda, a teorias lógicas (“ontologias eficazes”). A categoria “ontologias ineficazes” do Espectro Ontológico integra aspectos de ambos, e é utilizada na modelagem de dados (UML), em vez de na representação de áreas. Enquanto Obrst menciona a classe versus distinção de ocorrências em sua descrição de ontologias eficazes, ele não se aprofunda na elaboração dessa distinção.

Isto está de acordo com o principal argumento que tentamos expor neste documento: demonstrar que a representação do conhecimento – que poderia ser mais adequadamente denominada de modelagem abrangente de crenças disseminadas entre cientistas – não é uma tarefa das ontologias formais. Tampouco as ontologias formais descrevem entidades pertencentes à área da linguagem humana. Elas têm representações distintas,

servem a diferentes propósitos, e utilizam diferentes formalismos. Supomos que uma compreensão mais clara dessas diferenças irá facilitar a definição de interfaces mais robustas e úteis entre elas, e assim reduzir a ocorrência de modelos não-intencionais, auxiliando na criação de uma base mais racional para sistemas semanticamente interoperáveis na biologia e na medicina.

## Agradecimentos

Este trabalho teve o apoio dos projetos @neurIST e DEBUGIT, da União Européia, e dos Institutos Nacionais de Saúde através do Roteiro NIH para Pesquisa Médica, Bolsa 1 U 54 HG004028.

## Nota

1. Existe uma “forma” de se representar relações n-árias em OWL através da reificação – ver <http://www.w3.org/TR/swbp-n-aryRelations>

## Referências bibliográficas

- Baader F, Lutz C, Suntisrivaraporn B. CEL – A Polynomial-time Reasoner for Life Science Ontologies. Proceedings of the International Joint Conference on Automated Reasoning, 8, 2006, Heidelberg: Springer; 2006. p. 287-291.
- Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF. The Description Logic Handbook Theory, Implementation, and Applications (2nd Edition). Cambridge: Cambridge University Press; 2007.
- Baader F, Peñaloza R, Suntisrivaraporn B. Pinpointing in the Description Logic EL. Description Logics 2007. <http://ceur-ws.org/Vol-250>.
- Beisswanger E, Stenzhorn H, Schulz S, Hahn U. BIO-TOP: An Upper Domain Ontology for the Life Sciences. A Description of its Current Structure, Contents, and Interfaces to OBO Ontologies. Accepted for publication in Applied Ontology; 2008.
- BMIR (Stanford Center for Biomedical Informatics Research). The Protégé Ontology Editor and Knowledge Acquisition System; 2008. Available from: <http://protege.stanford.edu>. Last accessed: 30 Jan. 2009
- Bodenreider O, Smith B, Kumar A, Burgun A. Investigating subsumption in DL-based terminologies: a case study in SNOMED-CT. First International Workshop on Formal Biomedical Knowledge Representation (KR-MED 2004); 2004. p. 12-20.
- Boeker M, Stenzhorn H, Kumpf K, Bijlenga P, Schulz S, Hanser S. The @neurIST ontology of intracranial aneurysms: providing terminological services for an integrated IT infrastructure. Proceedings of the 2007 Annual Symposium of the American Medical Informatics Association, Washington: AMIA, 2007; p. 39-50.
- Boyd R. Scientific Realism, Stanford Encyclopedia of Philosophy, 2002. Available from: <http://plato.stanford.edu/entries/scientific-realism>. Last accessed: 30 Jan. 2009.
- Ceusters W, Smith B, Flanagan J. Ontology and medical terminology: why description logics are not enough. Towards an Electronic Patient Record Proceedings of TEPR 2003, Boston: Medical Records Institute, 2003.
- Ceusters W, Elkin P, Smith B. Negative findings in electronic health records and biomedical ontologies: a realist approach. Intern J Med Inform. 2007; 76:326-33.
- Fellbaum C. WordNet: an electronic lexical database. Cambridge: MIT Press; 1998.
- Genomics Standard Consortium. The GAZ ontology. [http://gensc.org/gc\\_wiki/index.php/GAZ\\_Project](http://gensc.org/gc_wiki/index.php/GAZ_Project). Last accessed: 30 Jan. 2009.
- Grenon P. BFO in a nutshell: a bi-categorical axiomatization of BFO and comparison with DOLCE. IFOMIS Technical Report, 6; 2003.
- Gruber TR. A translation approach to portable ontology specifications. Knowledge acquisition. Special issue: Current issues in knowledge modeling. 1993; 5(2): 199-200.
- Guarino N. Formal ontology in information systems. Amsterdam: IOS Press; 1998.
- Guarino N. *Avoiding IS-A overloading: the role of identity conditions in ontology design*. international conference on spatial information theory: cognitive and computational foundations of geographic information science, Proceedings. 1999:221-34.
- Hoehndorf R, Loebe F, Kelso J, Herre H. Representing default knowledge in biomedical ontologies: application to the integration of anatomy and phenotype ontologies. BMC Bioinformatics. 2007; 8:377.
- Hofweber T. Logic and Ontology, Stanford Encyclopedia of Philosophy; 2004. Available from: <http://plato.stanford.edu/entries/logic-ontology>. Last accessed: 30 Jan. 2009.
- Horridge M, Drummond N, Goodwin J, Rector A, Stevens R, Wang H. The Manchester OWL Syntax. Proc. of the OWLED Workshop: Experiences and Directions 2006, 11, 2006. Available from: <http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-216>. Last accessed: 30 Jan. 2009.
- IHTSDO (International Health Terminology Standards Development Organisation). Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT), 2008. Available from: <http://www.ihtsdo.org/snomed-ct>. Last accessed: 30 Jan. 2009.
- Jansen L. "On ascribing dispositions". In: Max Kistler, Bruno Gnessounou, editors. Dispositions and causal powers, Aldershot: Ashgate; 2007:161-77.
- Kifer M, Lausen G. F-logic: a higher-order language for reasoning about objects, inheritance, and scheme. ACM SIGMOD Record. 2; 1989: 134-46.
- Klinov P. Pronto: A Non-Monotonic Probabilistic Description Logic Reasoner. Proceedings of the European Semantic Web Conference, 6, 2008. Heidelberg: Springer; 2008: 822-6.
- Klyne G, Carroll J. Resource Description Framework (RDF): concepts and abstract syntax; 2004. <http://www.w3.org/TR/rdf-concepts>. Last accessed: 30 Jan. 2009
- Koller D, Levy A, Pfeffer A. P-classic: A tractable probabilistic description logic. Proceedings of AAAI; 1997:390-7.
- Kusnierczyk W. Nontological engineering. Proceedings of the International Conference on Formal Ontology in Information Systems, 11, 2006. Amsterdam: IOS Press; 2006:39-50.
- Mccray At, Nelson SJ. The representation of meaning in the UMLS. Meth Inform Med. 1995; 34(1-2):193-201.
- Minsky M. A Framework for Representing Knowledge. MIT-AI Laboratory Memo 306, June; 1974. <http://web.media.mit.edu/~minsky/papers/Frames/frames.html>

- MITA (Medical Imaging and Technology Alliance). Digital imaging and communication in medicine (DICOM), 2008. Available from: <http://medical.nema.org>. Last accessed: 30 Jan. 2009.
- NCI (National Cancer Institute). NCI Enterprise Vocabulary Services (EVS), 2008. Available from: <http://www.cancer.gov/cancertopics/terminologyresources>. Last accessed: 30 Jan. 2009.
- Neuhaus F, Smith B. Modelling principles and methodologies. Relations in anatomical ontologies. In: Burger A, Davidson D, Baldock R, editors. *Anatomy ontologies for bioinformatics: principles and practice*; 2007.
- NHS (World Health Organization). Dictionary of medicines and devices (dm+d); 2008. Available from: <http://www.dmd.nhs.uk>. Last accessed: 30 Jan. 2009.
- NLM (United States National Library of Medicine). Medical Subject Headings (MeSH); 2008. Available from: <http://www.nlm.nih.gov/mesh>. Last accessed: 30 Jan. 2009.
- NLMa (United States National Library of Medicine). RxNorm; 2008. Available from: <http://www.nlm.nih.gov/research/umls/rxnorm>. Last accessed: 30 Jan. 2009.
- NLMb (United States National Library of Medicine). Unified Medical Language System (UMLS), 2008. Available from: <http://www.nlm.nih.gov/research/umls>. Last accessed: 30 Jan. 2009.
- Northrop Grumman. Medical Dictionary for Regulatory Activities (MedDRA); 2008. Available from: <http://www.meddrasso.com>. Last accessed: 30 Jan. 2009.
- Noy NF, McGuinness DL. Ontology development 101: a guide to creating your first ontology; 2001, Technical Report, <http://ce.sharif.edu/~daneshpajouh/ontology/ontology-tutorial-noy-mcguinness.pdf>
- OBI (Ontology of Biomedical Investigation Consortium). The ontology of biomedical investigations. <http://purl.obofoundry.org/obo/obi>. Last accessed: 30 Jan. 2009.
- Patel-Schneider PF, Hayes P, Horrocks I. OWL - Web Ontology Language Semantics and Abstract Syntax. W3C Recommendation; 2004. Available at <http://www.w3.org/TR/owl-semantics>. Last accessed: 30 Jan. 2009.
- Quine O. On what there is. In: Gibson R, editor. *Quintessence - Basic readings from the philosophy of W. V. Quine*. Cambridge: Belknap Press, Harvard University; 2004.
- Rector AL, Bechhofer S, Goble CG, Horrocks I, Nowlan WA, and Solomon WD. The GRAIL concept modelling language for medical terminology. *Artificial Intelligence in Medicine*. 1997; 9(2):139-71.
- Rector AL. Defaults, context, and knowledge: Alternatives for OWL-Indexed Knowledge Bases. *Pacific Symposium on Biocomputing*; 2004: 226-37.
- Rector AL. Barriers, approaches and research priorities for integrating biomedical ontologies; 2008. Available from: [www.semantichealth.org/DELIVERABLES/SemanticHEALTH\\_D6\\_1.pdf](http://www.semantichealth.org/DELIVERABLES/SemanticHEALTH_D6_1.pdf). Last accessed: 30 Jan. 2009.
- Regenstrief Institute. Logical Observation Identifiers Names and Codes (LOINC); 2008. Available from: <http://loinc.org>. Last accessed: 30 Jan. 2009.
- Reiter R. A logic for default reasoning. *Artificial Intelligence*. 1980; 13:81-132.
- Schulz S, Hahn U. Medical knowledge reengineering - converting major portions of the UMLS into a terminological knowledge base. *Intern J Med Inform*. 2001; 64(2-3): 207-21.
- Schulz S, Jansen L. Molecular interactions: on the ambiguity of ordinary statements in biomedical literature; 2008. Forthcoming in *Applied Ontology*.
- Shortliffe EH, Davis R, Axline SG, Buchanan BG, Green CC, Cohen SN. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. *Comp Bio Res*. 1975; 8(8):303-20.
- Sirin E, Parsia B, Cuenca Grau B, Kalyanpur A, Katz Y. Pellet: a practical OWL DL reasoner. *J Web Semantics*. 2007; 5(2):51-3.
- Smith B. Beyond concepts: ontology as reality representation. *Proceedings of the International Conference on Formal Ontology in Information Systems*, 11; 2004. p. 39-50.
- Smith B, Köhler J, Kumar A. On the application of formal principles to life science data: a case study in the gene ontology. *Proceedings of Data Integration in the Life Sciences (DILS 2004)*, Berlin: Springer; 2004. p. 79-94.
- Smith B, Mejino JLV, Schulz S, Rosse C. Anatomical information science. In: *COSIT 2005: spatial information theory. Foundations of Geographic Information Science, Lecture Notes in Computer Science*, Springer; 2005. p. 149-64.
- Smith B, Mejino Jr JLV, Schulz S, Kumar A, Rosse C. Anatomical Information Science. In: Cohn AG, Mark DM, editors. *Spatial information theory. Proceedings of COSIT 2005, Heidelberg*: Springer; 2005. p. 149-64.
- Smith B, Kusnierczyk W, Schober D, Ceusters W. Towards a Reference Terminology For Ontology Research And Development In The Biomedical Domain. *Proceedings of KR-MED - Biomedical Ontology in Action*; 2006. p. 57-66.
- Smith M, Welty C, McGuinness DL. OWL Web ontology language guide, W3C Recommendation; 2004. Available from: <http://www.w3.org/TR/owl-guide>. Last accessed: 30 Jan. 2009.
- Soualmia LF, Golbreich C, Darmoni SJ. Representing the MeSH in OWL: towards a semi-automatic migration. *Workshop on Formal Biomedical Knowledge Representation (KR-MED)*, 7; 2004. p. 81-7.



Tsarkov D, Horrocks I. FaCT++ Description logic reasoner: system description. Proceedings of the Third International Joint Conference on Automated Reasoning, 8, 2006. Heidelberg: Springer; 2006. p. 292-7.


UMC (Uppsala Centre for International Drug Monitoring). WHO Drug Dictionary Enhanced; 2008. Available from: <http://www.unc-products.com>. Last accessed: 30 Jan. 2009.

UNIPROT (Universal Protein Resource Consortium). UniProt Protein Knowledgebase; 2008. Available from: <http://www.uniprot.org>. Last accessed: 30 Jan. 2009.

W3C (World Wide Web Consortium). Semantic Web Activity, 2008. Available from: <http://www.w3.org/2001/sw>. Last accessed: 30 Jan. 2009.

Welly C, Guarino N. Supporting ontological analysis of taxonomic relationships”, Data & Knowledge Engineering 39. Elsevier; 2001

WHO (World Health Organization). International Classification of Diseases (ICD); 2008. Available from: <http://www.who.int/classifications/icd>. Last accessed: 30 Jan. 2009.

WHOC (WHO Collaborating Centre for Drug Statistics Methodology). Anatomical Therapeutic Chemical Classification System (ATC), 2008. Available from: <http://www.whoc.no/atcddd>. Last accessed: 30 Jan. 2009. 

## Sobre os autores

### *Stefan Schulz*

É formado em medicina pela Heidelberg University, Alemanha, e é pesquisador sênior e professor do Instituto de Biometria Médica e Informática da Medicina do Centro Médico Universitário Freiburg, onde chefia o Grupo de Pesquisas em Informática na Medicina. Seu trabalho se concentra em terminologias e ontologias biomédicas, representação do conhecimento biomédico, recuperação de documentos médicos multilíngües, mineração de texto e dados em repositórios de documentos clínicos, aprendizado eletrônico na Medicina, e informática da saúde em países em desenvolvimento.

Após executar trabalhos clínicos em cirurgia e medicina interna, obteve seu diploma de doutorado na área da higiene tropical, onde efetuou um estudo de campo parasitológico em São Luís, Brasil. Após obter qualificação técnica em computação médica, mudou-se para a Universidade de Freiburg, onde participou de projetos de desenvolvimento de software clínico e educacional, e de diversos projetos de pesquisa na área da extração de informações, terminologias biomédicas, engenharia da linguagem médica, e tecnologias semânticas. Tem desempenhado papéis de liderança em diversos projetos financiados pela União Européia. Stefan Schulz é autor de mais de cem publicações revisadas por especialistas, e recebeu vários prêmios. Tem oferecido repetidas contribuições a projetos de pesquisa na área da informática de saúde brasileira desde 2001, como pesquisador convidado da Pontifícia Universidade Católica do Paraná (PUC-PR).

### *Holger Stenzhorn*

É lingüista computacional (Universidade Saarland, Alemanha) e pesquisador adjunto do Instituto de Biometria Médica e Informática da Medicina do Centro Médico Universitário Freiburg, Alemanha. Seu trabalho enfoca a representação e gerenciamento de informação e dados, ontologias e tecnologias da Semantic Web, informática biomédica, processamento de linguagem natural, interfaces de usuário multimodais, e projeto e desenvolvimento de software. Já participou do desenvolvimento de recuperação de documentos multilíngües, extração de informação, e sistemas de geração de linguagem natural, tanto na indústria quanto no meio acadêmico. Atualmente está envolvido em diversas tarefas de engenharia ontológica: uma ontologia para a pesquisa de aneurismas cerebrais (projeto @neurIST, financiado pela União Européia); uma para os testes clínicos de nefroblastoma e câncer de mama (projeto ACGT, financiado pela União Européia); e a ontologia BioTop, de todas as áreas. Holger é membro do Grupo W3C de Participação em Saúde e Ciências Biológicas.



**RECIIS**

Revista Eletrônica de Comunicação  
Informação & Inovação em Saúde

[www.reciis.cict.fiocruz.br]

ISSN 1981-6278

**Artigos originais**

# Uma análise ontológica do eletrocardiograma

DOI: 10.3395/reciis.v3i1.242pt



**Bernardo  
Gonçalves**

Departamento de Ciência  
da Computação, Univer-  
sidade Federal do Espírito  
Santo, Vitória, Brasil  
bgoncalves@inf.ufes.br



**Veruska  
Zamborlini**

Departamento de Ciência  
da Computação, Univer-  
sidade Federal do Espírito  
Santo, Vitória, Brasil  
veruska@inf.ufes.br

## **Giancarlo Guizzardi**

Departamento de Ciência da Computação, Universidade Federal do Espírito Santo, Vitória, Brasil  
gguizzardi@inf.ufes.br

### **Resumo**

A bioinformática tem sido um campo fértil para aplicação da disciplina de ontologia formal. A representação com princípios de entidades biomédicas tem alimentado progressivamente a pesquisa biológica, com benefícios diretos que vão da reformulação das terminologias médicas à introdução de novas perspectivas para modelos aprimorados de Registros Eletrônicos de Saúde (RES). Este artigo apresenta uma análise ontológica do eletrocardiograma (ECG) independente de aplicação e fundamentada na Ontologia Fundamental Unificada. Com o objetivo de investigar o fenômeno subjacente a esse exame cardiológico, trabalhamos com os subdomínios da eletrofisiologia e anatomia do coração humano. Então delineamos uma Ontologia do ECG construída sobre a Ontologia Relacional OBO (Open Biomedical Ontology). Além disso, a ontologia de domínio esboçada aqui se inspira tanto no Modelo Fundamental de Anatomia como na Ontologia de Funções proposta sob o amparo do programa de pesquisa em Ontologia Geral Formal (OGF).

### **Palavras-chave**

ontologia biomédica; eletrocardiograma; eletrofisiologia cardíaca

### **Introdução**

O campo da Bioinformática está testemunhando a aplicação da *ontologia formal* (a disciplina) à representação de entidades biológicas (ex. Schulz & Hahn 2007) e a (re-organização das terminologias médicas também em virtude dos registros eletrônicos de saúde (RES) (ex.

Schulz et al. 2008). A motivação é (basicamente) estabelecer bases para (i) biólogos e médicos armazenarem e comunicarem informação biomédica e dados relativos a pacientes com eficiência; (ii) gradualmente integrarem esses recursos no desenvolvimento da próxima geração de aplicativos computacionais biomédicos baseados em

dados. Esses aplicativos destinam-se a oferecer suporte em ciência básica e pesquisa clínica, bem como na prestação de serviços de saúde mais eficientes. Conforme colocado por Rosse e Mejino Jr. (2003), “*tal foco ampliado em bioinformática é inevitável na era pós-genoma e o processo de fato já começou*”.

Uma iniciativa notável para reunir ontologias biomédicas seguindo princípios é a *Open Biomedical Ontologies (OBO) foundry* (Smith et al. 2007). Até agora ela consiste em 60 ontologias que, sem variar muito em termos de granulosidade, canonicidade e estágio de desenvolvimento, em que cada uma visa representar um objeto de estudo bem delimitado. Entre as ontologias mais consultadas do OBO, está o Modelo Fundamental de Anatomia (FMA) (Rosse & Mejino Jr. 2003). O FMA lida com a estrutura anatômica do corpo dos mamíferos (especialmente o humano). Entretanto, apesar do fato de que o domínio da eletrofisiologia do coração humano é de interessante significativo em Biomedicina, ainda falta uma ontologia da eletrofisiologia do coração na OBO, bem como na literatura de ontologia biomédica<sup>1</sup>. Ademais, embora o eletrocardiograma (ECG) defina uma das formas notáveis de dados biomédicos, até onde sabemos, ele ainda não foi tratado na literatura da ontologia biomédica.

O ECG é o teste mais freqüentemente utilizado na Cardiologia para medir a atividade do coração (Geselowitz 1989). Nos últimos anos, tanto o armazenamento e a transmissão dos registros de ECG foram objeto de iniciativas de padronização. Entre os padrões de referência, pode-se recorrer a SCP-ECG<sup>2</sup>, FDA XML<sup>3</sup> ou HL7 ECG Mensagem de Notação v3<sup>4</sup>. Entretanto, o foco desses padrões é geralmente como dados e informações devem ser representado em sistemas de computador e transmissão de mensagens (Smith et al. 2007, p. 1252, Yu 2006, p. 254). Por outro lado, existe a necessidade de se concentrar na representação adequada da realidade biomédica sendo examinada (Smith 2006). Isto é, em *o que é o ECG*, tanto pelo lado do paciente como do médico. Isso é claramente relevante, já que o ECG, como um sinal vital, é uma peça importante na composição do RES de hoje e provavelmente no RES do futuro.

Nos últimos anos, temos lidado com o ECG como objeto de investigação ontológica. Um esforço inicial de representação de dados de ECG aplicando-se técnicas de ontologia formal resultaram em uma ontologia de domínio ECG preliminar relatada em (Gonçalves et al. 2007, Zamborlini et al. 2008). Desde então, temos revisado a base subjacente a esse empenho pioneiro. Isso nos levou a reformular nossa representação ontológica do ECG em prol de uma maior especialização, nível de detalhes, densidade e conectividade para citar os termos transmitidos por Rector et al. (2006, p. 335). Este artigo apresenta uma análise ontológica do eletrocardiograma (ECG) independente de aplicação. Nossa análise é baseada na Ontologia Fundamental Unificada (UFO) (Guizzardi & Wagner 2009). A UFO começou como uma unificação da OGF (Ontologia Formal Geral) (Heller & Herre 2004) e a ontologia de nível superior

de universais com base no OntoClean (<http://www.ontoclean.org>). Entretanto, conforme demonstrado em (Guizzardi & Wagner 2009), existem inúmeras questões problemáticas relacionadas ao objetivo específico de desenvolver fundamentos ontológicos gerais para modelagem conceitual que não são cobertas de forma satisfatória pelas ontologias fundamentais existentes como a OGF, DOLCE ou OntoClean. Por esse motivo, a UFO foi desenvolvida em uma ontologia de referência madura baseada em diversas teorias da Ontologia Formal, Lógica Filosófica, Filosofia da Linguagem, Linguística e Psicologia Cognitiva. Essa ontologia é apresentada com profundidade e formalmente caracterizada em (Guizzardi 2005). Neste artigo, apresentamos nossa caracterização formal dessa análise ontológica usando a Lógica de Primeira Ordem (FOL) padrão.

Empregando os resultados dessa análise ontológica, delineamos uma ontologia de domínio para o ECG que incorpora os subdomínios da eletrofisiologia e da anatomia do coração. A Ontologia de ECG também se utiliza de diversas teorias fundamentais existentes, a saber: (i) a Ontologia Relacional OBO, que fornece relações básicas a serem usadas nas ontologias biomédicas (Smith et al. 2005); (ii) o Modelo Fundamental de Anatomia (FMA), ao lidar com a anatomia humana para o ECG; (iii) a Ontologia de Funções (OF) desenvolvida sob o guarda-chuva do programa de pesquisa em Ontologia Geral Formal (OGF) (Burek et al. 2007) para lidar com as funções eletrofisiológicas do coração. Essa ontologia de ECG delineada é atualmente implementada com uma combinação da linguagem de representação OWL DL e sua extensão SWRL (Horrocks et al. 2005).

## Materiais e métodos

Os princípios metodológicos para a engenharia ontológica vêm atraindo atenção crescente na literatura ontológica biomédica, cf. (Yu 2006). Na nossa análise ontológica do ECG, empregamos diversos princípios da engenharia ontológica e da ontologia fundamental, que também favorecem o raciocínio automático eficiente e a integração ontológica.

## Engenharia ontológica

Empregamos uma abordagem de engenharia ontológica baseada na premissa de que promover uma ontologia de domínio (no contexto da IA) exige dois instrumentos ontológicos diferentes (Guizzardi & Halpin 2008), a saber, uma teoria do domínio de objeto bem fundamentada ontologicamente cujo propósito é ser fortemente axiomática para limitar o máximo possível o significado pretendido pela teoria e outra cujo propósito é ser uma ferramenta computacional para o raciocínio automatizado e recuperação de informação. Em Bittner e Donnelly (2007), os autores evidenciam uma linha análoga de argumento e propõem o uso de FOL como um formalismo para o primeiro e alguma espécie de Lógica Descritiva (DL) para a última. Seguimos a mesma escolha de linguagens de representação aqui, particularmente através do uso de OWL DL (espelhada para uma

DL) com a extensão SWRL para regras com a última (Horrocks et al. 2005). Além disso, como é tradicional em Engenharia Ontológica (Yu 2006, p.255), especificamos um conjunto de perguntas de competência para delimitar o escopo e o propósito do domínio que temos à mão. Essa técnica metodológica também é benéfica no fim do ciclo de desenvolvimento como forma de avaliar o instrumento resultante.

## Engenharia ontológica

Chamamos atenção para o fato de que criar uma ontologia de domínio (Biomédica) com base em alguma fundamentação ontológica é útil a até necessário. Uma estrutura ontológica de nível superior não só fornece um suporte para a tomada de decisões ontológicas (Guarino & Welty 2002), mas também permite que tomemos decisões da forma mais transparente possível na ontologia de domínio resultante. Nosso estudo ontológico sobre ECG está fundamentado na estrutura de nível superior da UFO (Guizzardi & Wagner 2009). A UFO contém categorias ontológicas superiores (ex. resistente, persistente, tipo, papel, coletivo, relator e assim por diante) que são *demonstrados pelos* universais do domínio de ECG (ex. coração, registro de ECG).

## Garantindo um raciocínio automatizado eficiente

Um dos principais objetivos práticos desta pesquisa é usar os resultados do nosso estudo ontológico de ECG para dar suporte ao raciocínio automatizado sobre os universais e particulares do ECG e eletrofisiologia do coração. Temos procurado (praticamente) o ponto ideal para expressar o máximo possível da teoria ontológica de ECG que desenvolvemos aqui em uma combinação do OWL DL e sua extensão SWRL, mantendo a decidibilidade e maneabilidade computacional. Como as lógicas de ordem superior comprometem o objetivo do raciocínio automatizado prático, as categorias UFO são expressas na implementação da ontologia de ECG resultante meramente como notações OWL. Apesar disso, sustentamos que a estrutura de princípios da ontologia (ex. a solidez ontológica de taxonomias parte-de e subordinação) ainda é preservada na implementação.

## Integração ontológica

Buscamos a integração ontológica, especialmente com respeito ao OBO foundry e a iniciativa da web

semântica. Esta última nos influenciou para selecionar a combinação OWL DL/SWRL como linguagem de código ontológica. Com relação ao primeiro, a princípio nossa teoria ontológica de ECG se baseia no FMA no que cobre os conceitos anatômicos humanos que são relevantes para uma teoria ECG. Além disso, aplicamos a Ontologia de Funções (OF) proposta em (Burek et al. 2007) como uma estrutura ontológica de nível superior para modelar funções eletrofisiológicas do coração. Em segundo lugar, tomamos emprestadas relações do OBO Ontologia Relacional (domínio cruzado) (Smith et al. 2005), que são especialmente valiosas para definir relações espaciais ao longo do tempo. Nós os usamos em combinação com relações específicas de domínio cunhadas aqui e relações complementares formalmente descritas em UFO.

## A Ontologia Relacional OBO

Uma distinção essencial em muitas ontologias fundamentais (ex. DOLCE, OGE, UFO) e, em particular, na Ontologia Relacional (OR) OBO, é a distinção entre *permanentes* e *processos*. Dito literalmente (Smith et al. 2005).

*Permanentes são aquelas entidades que persistem, continuam a existir através do tempo passando por diferentes tipos de mudanças, incluindo mudanças espaciais. Processos são entidades que se desdobram em fases temporais sucessivas.*

De modo geral, a noção de permanente é semelhante ao que se chama de *resistente* em UFO, enquanto que o processo pode ser entendido como semelhante ao *resistente*. A tabela 1 apresenta as relações OR que empregamos aqui. Uma discussão nessas relações pode ser encontrada em (Smith et al. 2005). Nessa seção, por uma questão de brevidade, mantivemos inicialmente a sintaxe semi-formal empregada nesse artigo para posteriormente partir para seus contrapartes da Lógica de Primeira Ordem (LPO) correspondentes. As seguintes variáveis e faixas são usadas em seguida.

- $C, C_1, \dots$  para faixas em classes permanentes
- $P, P_1, \dots$  para faixas em processos permanentes
- $c, c_1, \dots$  para faixas em instâncias permanentes
- $p, p_1, \dots$  para faixas em instâncias de processo
- $r, r_1, \dots$  para faixas de regiões espaciais em três dimensões
- $t, t_1, \dots$  para variações em instâncias de tempo

**Tabela 1 - As relações usadas na Ontologia Relacional OBO**

Relação	Definição
$c \text{ instance\_of } C \text{ at } t$	Uma relação primitiva entre uma instância permanente e uma classe que instancia e um tempo específico
$p \text{ instance\_of } P$	uma relação primitiva entre uma instância de processo e uma classe que instancia independente de tempo

Cont.



$c \text{ part\_of } c_1 \text{ at } t$	uma relação primitiva entre duas instâncias permanentes e um tempo em que um é parte do outro
$c \text{ located\_in } r \text{ at } t$	Uma relação primitiva entre uma instância permanente, a região espacial que ela ocupa em um tempo específico
$r_1 \text{ part\_of } r_2$	uma relação primitiva parte-de, independente de tempo (ex. constante) entre regiões espaciais (uma sendo sub-região da outra)
$r \text{ adjacent\_to } r_1$	uma relação primitiva de proximidade entre duas regiões espaciais disjuntas
$t_1 \text{ earlier } t_2$	uma relação primitiva entre dois tempos
$p \text{ has\_participant } c \text{ at } t$	Uma relação primitiva entre um processo, um permanente em um tempo específico
$p \text{ has\_agent } c \text{ at } t$	uma relação primitiva entre um processo, um permanente em um tempo $t$ específico em que o permanente está casualmente ativo no processo
$c \text{ exists\_at } t$	para algum $p$ , $p \text{ has\_participant } c \text{ at } t$
$p \text{ occurring\_at } t$	para algum $c$ , $p \text{ has\_participant } c \text{ at } t$
$t \text{ first\_instant } p$	$p \text{ occurring\_at } t$ e para todo $t$ , se $t$ , $t$ anterior então não $p \text{ occurring\_at } t$
$t \text{ last\_instant } p$	$p \text{ occurring\_at } t$ e para todo $t$ , se $t$ , $t$ anterior então não $p \text{ occurring\_at } t$
$c \text{ located\_in } c_1 \text{ at } t$	para algum $r$ , $r_1$ ( $c \text{ located\_in } r \text{ at } t$ em $t$ e $c_1$ , $\text{located\_in } r_1$ em $t$ e $r \text{ part-of } r_1$ )

## Resultados

### Anatomia para o ECG

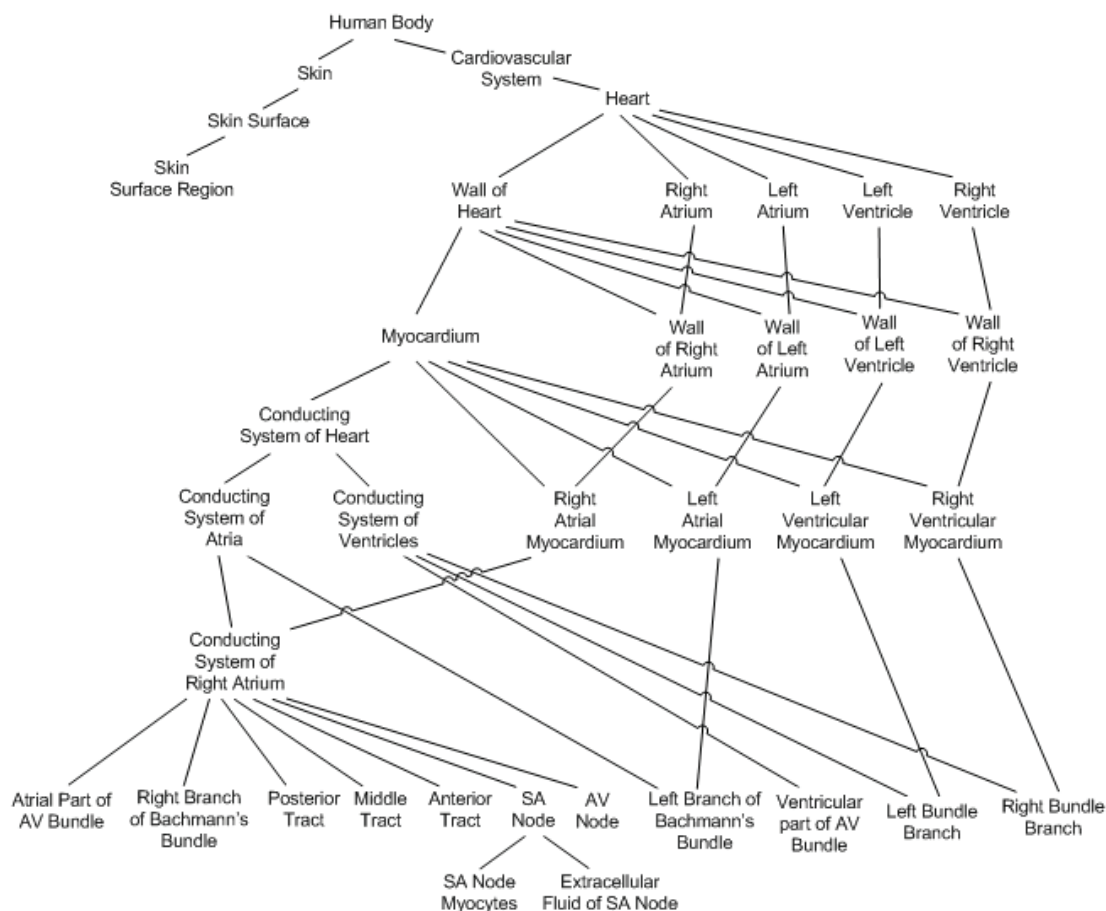
Essa sessão fornece um relato ontológico dos permanentes anatômicos do corpo humano diretamente envolvidos no ECG. Tomamos a FMA como referência e depois consideramos os universais permanentes com os mesmos termos empregados pela FMA ou com seus sintomas. Contudo, não seguimos rigidamente as escolhas de modelagem da FMA já que elas não são totalmente apoiadas pelos fundamentos ontológicos. Por exemplo, Donnelly et al. (2005) apontam alguns problemas com relação à concepção de relações parte-todo na FMA, enquanto Kumar et al. (2004, p. 505) e Rector et al. (2006, p. 345) discutem problemas na FMA com relação à granularidade.

Em consonância com a FMA (ver Figura 1 para a taxonomia anatômica parte-de), começamos com o *corpo humano* e elaboramos nas partes que compõem o *coração humano*. Incluímos no nosso modelo *pele*, *superfície da pele* e *região da superfície da pele* porque, para antecipar a seção que lida com o ECG, a última é parte é a parte do corpo humano que é objeto de mensuração por um aparelho de registro para aquisição de ECG. No nosso escopo, vale a pena dizer que o coração tem como partes o *átrio direito* e *esquerdo*, o *ventrículo direito* e o *esquerdo* e a *parede cardíaca*. Enquanto o *átrio* e os *ventrículos* são subtipos

das *cavidades de órgão*, a parede do coração é um subtipo de parede de órgão. A parede do coração tem como partes as camadas do endocárdio, epicárdio e *miocárdio*. A última é um subtipo de *camada muscular de um órgão*, que é ainda dividida (não completamente) em *miocárdio atrial direito* e *esquerdo* e *miocárdio ventricular direito* e *esquerdo*. São todos tipos de *regiões do miocárdio* e têm como partes os sistemas condutores do *átrio direito* e *esquerdo* e dos *ventrículos direito* e *esquerdo* respectivamente.

Consideramos aqui somente o *sistema condutor do átrio direito*, já que ele exemplifica uma divisão completa em variadas partes finais do coração no nosso escopo. Diferente dos curadores de FMA, não incluímos na partonomia anatômica principal universais em diferentes níveis de granularidade (cf. Rector et al. 2006), ex. *miócito* do nó SA. Supomos aqui que tal universal é uma granulação do conjunto de *miócitos* do nó SA. Esse conjunto de *miócitos*, por sua vez, é um componente funcional do (um tipo específico de parte) nó SA, que emerge do conjunto de células associadas a um *fluido extracelular*. No nosso entendimento, o grânulo de *miócito* do nó SA não é parte do nó SA. A noção de conjunto é contemplada na ontologia UFO (Guizzardi 2005, Capítulo 5) e também é discutida em profundidade em (Rector et al. 2006).

Na partonomia anatômica da Figura 1, usamos a relação de parte-de adotando o que é conhecido em Ontologia Formal como *mereologia mínima* (Guizzardi 2005, Capítulo



**Figura 1** - Partonomia de uma anatomia para o ECG. As linhas representam relações parte-de (de baixo para cima) entre as entidades anatômicas.

5). As ligações parte-de mostradas na Figura 1 representam uma relação de nível universal (entre dois universais, ex. o átrio direito é parte do coração) definido a partir de um nível-instância parte\_de (entre dois indivíduos, ex. meu átrio direito é parte do meu coração em particular). A relação de parte de nível universal é definida considerando a versão nível-instância. Esta é uma relação primitiva caracterizada pelas meta-propriedades de irreflexibilidade, assimetria e transitividade. Formalmente, isso significa que:

**Irreflexibilidade:**  $\forall c, t \neg \text{part\_of}(c, c, t)$

**Assimetria:**  $\forall c_1, c_2, t \text{part\_of}(c_1, c_2, t) \rightarrow \neg \text{part\_of}(c_2, c_1, t)$

**Transitividade:**  $\forall c_1, c_2, c_3, t \text{part\_of}(c_1, c_2, t) \wedge \text{part\_of}(c_2, c_3, t) \rightarrow \text{part\_of}(c_1, c_3, t)$

A relação de parte de nível universal (as ligações na Figura 1) pode então ser obtida como a seguir:

$$\text{part\_of}(C_1, C_2) =_{\text{def}} \forall c_1 \exists t_1 \text{instance\_of}(c_1, C_1, t_1) \rightarrow \forall t (\text{instance\_of}(c_1, C_1, t) \rightarrow \exists c_2 (\text{instance\_of}(c_2, C_2, t) \wedge \text{part\_of}(c_1, c_2, t)))$$

Perceba também na Figura 1 que algumas entidades na partonomia só têm uma parte. Embora isso não seja

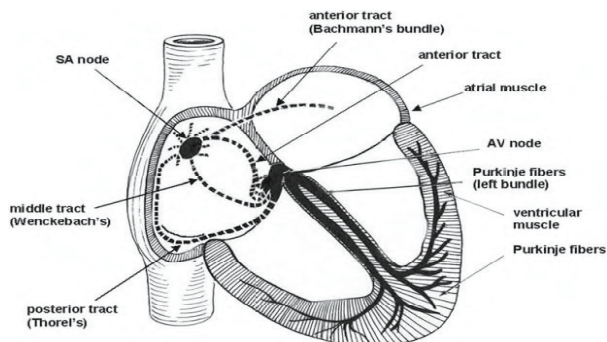
problema ao se adotar a mereologia mínima, a razão verdadeira aqui é outra. Essas entidades tem outros universais como partes, mas não são relevantes para a representação de ECG. Finalmente, é importante destacar que usamos a relação parte-de aqui para representar uma relação de parte adequada. Se necessário, uma relação de parte inadequada pode ser definida como de costume:

$$\text{improper\_part\_of}(c_1, c_2, t) =_{\text{def}} \text{part\_of}(c_1, c_2, t) \vee (c_1 = c_2 \text{ at } t)$$

## Eletrofisiologia do coração humano

Fontes bioelétricas surgem espontaneamente no coração no nível celular. Os miócitos (células musculares) do coração são imersas em fluido extracelular separados do seu interior por membranas que realizam o controle do transporte de íons. No estado de repouso, o interior desses miócitos tem um potencial negativo com relação ao exterior, ex. as células são eletricamente polarizadas. Entretanto, especialmente nos nós sinoatrial (SA) e atrioventricular (AV), partes do miocárdio (ver Figura 1), os miócitos abruptamente despolarizam e depois voltam para seu valor de repouso. Esse fenômeno é um resultado dos íons passando em qualquer direção através da membrana da célula (Geselowitz 1989).

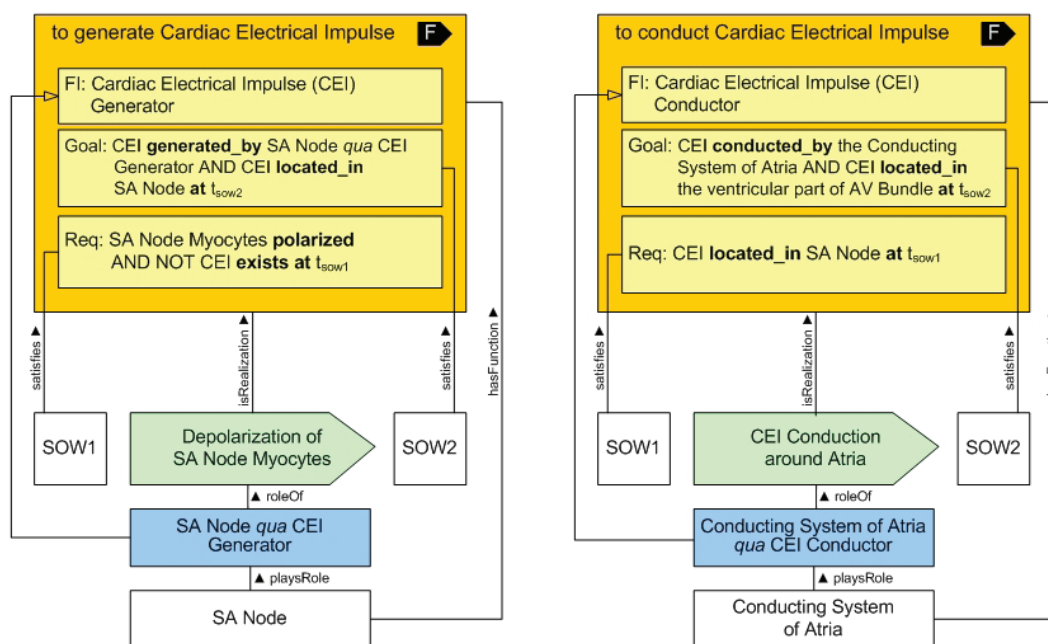
Portanto, especialmente os miócitos do nó SA e AV originam impulsos elétricos que são propagados para os miócitos vizinhos e normalmente atingem todo o coração. Por isso os nós SA e AV são chamados de marca-passos do coração. Entretanto, como esse tipo de impulso elétrico surge no nó SA em um ritmo mais rápido e com maior intensidade, diz-se que o impulso elétrico do nó AV é forçado pelo impulso no nó SA (Geselowitz 1989). Para conduzir o impulso elétrico cardíaco (surgido no nó SA) por todo o coração, existem outros miócitos além dos miócitos do SA e AV (as fibras de Purkinje) que constituem o *sistema condutor do coração* (ver Figura 2). O principal percurso de condução é o chamado sistema His-Purkinje. Ele é composto do feixe atrioventricular (feixe AV ou feixe de His), então bifurcado nos *ramos de feixe direito e esquerdo* (Laske & Iaizzo 2005, Guyton & Hall 2006). Como resposta ao impulso elétrico cardíaco conduzido pelo sistema, o miocárdio tem contrações em suas partes atrial e ventricular para lançar o sangue respectivamente nos ventrículos e na circulação sistêmica ou pulmonar.



**Figura 2** - O sistema de condução do coração (fonte: Laske & Iaizzo 2005).

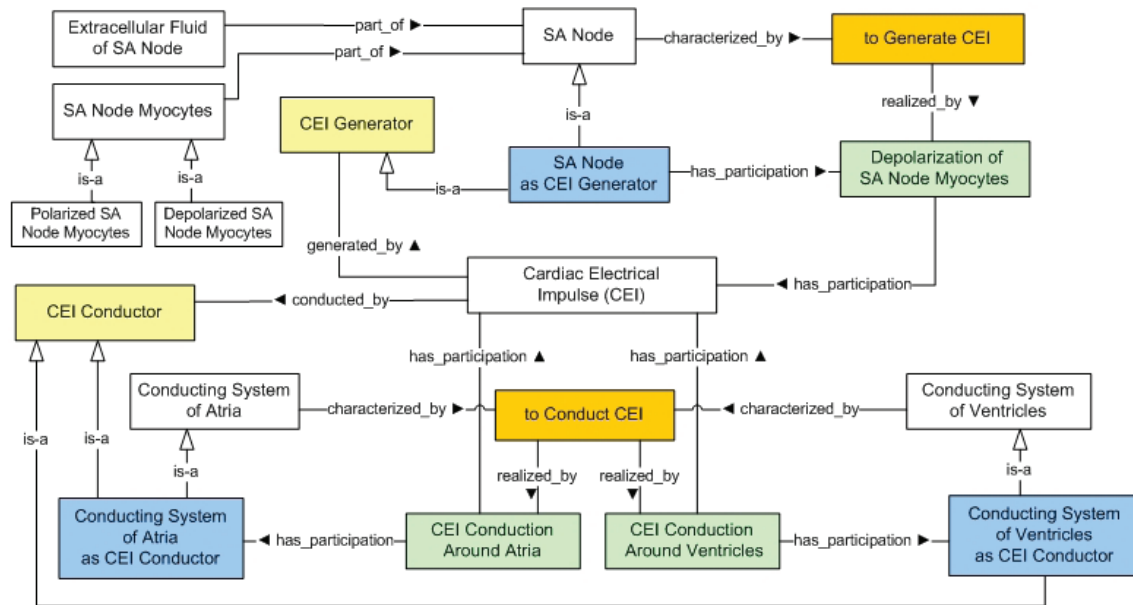
Dada essa visão geral, agora nos concentramos na representação ontológica da eletrofisiologia do coração humano significativa para a representação do ECG. Para tal, elaboramos com base na Ontologia de Funções (OF) proposta por Burek et al. (2007). Basicamente, buscamos fornecer uma estrutura clara das funções eletrofisiológicas do coração (*o que são*), e *como e por quem* podem ser realizadas. Pretendemos, dessa forma, conseguir reconstruir essas entidades fisiológicas de um ECG específico.

A estrutura básica de uma função, conforme apresentada em (Burek et al. 2007) é um conjunto de rótulos, um item funcional, um conjunto de exigências a serem cumpridas caso a função seja realizável e um objetivo a ser satisfeito no caso da função ser de fato realizada. A função está ligada ao permanente que *tem* a função e que pode *realizá-la* exercendo um *papel* específico (o item funcional). Esse papel é exercitado pelo que é chamado em Filosofia de *qua indivíduo* (Guizzardi 2005, Capítulo 7). Por exemplo, se John casa com Mary, os direitos e deveres (legalmente falando) devem ser desse momento em diante satisfeitos por John-qua-marido-de-Mary. Finalmente, a função é realizada por meio de um processo. Esse processo garante a transição de um estado do mundo (SOW) no qual as exigências da função são satisfeitas, para o SOW no qual o objetivo da função é satisfeita. Esse processo é chamado a *realização* da função. A realização pode ser considerada *real* ou *circunstancial*. Isto é, o processo pode ser a capacidade de realização da função, mesmo se essa capacidade nunca é efetivada, ex. no caso de algum mau funcionamento.



**Figura 3** - Funções eletrofisiológicas do coração representadas na estrutura de OF.

A Figura 3 ilustra dois exemplos de funções eletrofisiológicas do coração representadas usando a estrutura de OF, isto é, *para gerar impulso elétrico cardíaco* (IEC) e *para conduzir IEC*. Enquanto que o primeiro é realizado por meio do processo de *despolarização dos miócitos do nó SA*, o último, em sua manifestação atrial, é realizado por meio de um processo de *condução do IEC em torno do átrio*. A função *para conduzir o IEC* é manifestada pelo processo de *condução de IEC em torno dos ventrículos* de forma semelhante. A Figura 4 fornece uma representação adaptada dessas funções (incluindo a manifestação ventricular da condução de IEC). Sua aplicabilidade é esclarecida com mais detalhes na sessão sobre ECG.



**Figura 4** - Parte do modelo eletrofisiológico do coração. A função para *conduzir IEC* caracteriza ambos os sistemas condutores, tanto do átrio como do ventrículo. Essa função pode ser realizada pelo processo de condução do IEC pelo átrio ou pelo ventrículo. A função para *gerar IEC* em seguida caracteriza o nó SA e pode ser realizada pelo processo de despolarização dos miócitos do nó SA.

Usamos relações específicas definidas a seguir. Primeiramente, antes de definirmos o que significa dizer que o permanente foi *gerado por* outro, precisamos definir a noção de produção. A relação instância-nível produzida\_por existe entre um permanentes e um processo. Como formalmente descrito abaixo, o permanente  $c$  é produzido por um processo  $p$  se e somente se existe um e somente um instante de tempo  $t_i$ , tal que  $t_i$  é o último instante de  $p$ ,  $p$  tem  $c$  como participante em  $t_i$  e em todos os instantes de tempo anteriores a  $t_i$  então  $c$  não existe em  $t$ . A relação universal-nível produzida\_por é definida a seguir.

$$\begin{aligned} \text{produced\_by}(c, p) &=_{\text{def}} \\ \exists! t_i ( &\text{last\_instant}(p, t_i) \wedge \text{has\_participant}(p, c, t_i) \wedge \\ &\forall t ( \text{earlier}(t, t_i) \rightarrow \neg \text{exists}(c, t) ) ) \\ \text{produced\_by}(C, P) &=_{\text{def}} \\ \forall c \exists t \text{instance\_of}(c, C, t) &\rightarrow \exists p \text{instance\_of}(p, P) \wedge \\ &\text{produced\_by}(c, p) \end{aligned}$$

Perceba que dizer que um permanente participa em algum processo acarreta que ele existe durante o proces-

so, ver Tabela 1. Agora podemos prosseguir dando uma definição para a noção de geração. Um permanente  $c$  é gerado por outro permanente  $c_i$  sse existe um processo  $p$  tal que, para todos os instante de tempo  $t$  no qual  $p$  está ocorrendo então  $p$  tem  $c_i$  participando como agente e  $c$  é produzido por  $p$ . Ver também a versão nível universal.

$$\begin{aligned} \text{generated\_by}(c, c_i) &=_{\text{def}} \\ \exists p ( &\forall t ( \text{occurring\_at}(p, t) \rightarrow \text{has\_agent}(p, c_i, t) ) \wedge \\ &\text{produced\_by}(c, p) ) \\ \text{generated\_by}(C, C_i) &=_{\text{def}} \\ \forall c \exists t ( &\text{instance\_of}(c, C, t) \rightarrow \exists c_i, t_i ( \text{instance\_of}(c_i, \\ &C_i, t_i) \wedge \text{generated\_by}(c, c_i) ) ) \end{aligned}$$

A noção de condução, por sua vez, é um pouco mais complexa. Primeiro, segundo a UFO levamos a categoria *modo* em consideração. A razão é que uma entidade que é objeto de condução, como um impulso elétrico cardíaco (CEI), precisa ser *inerente* a um condutor para existir (Guizzardi 2005, Capítulo 6). Portanto, ela é



existencialmente dependente de algum condutor. O IEC é modelado aqui como um modo, somente como sintoma, que somente existe por ser inerente a um paciente. Antes de fornecer uma definição para condução, apresentamos abaixo a relação primitiva instância-nível de inerência, juntamente com a relação correlata universal-nível de *caracterização*. A inerência é um tipo irreflexível, assimétrico e intransitivo de relação dependente existencial; a caracterização só pode ser aplicada se F (ver fórmula abaixo) é uma instância da categoria *universal momento* (da qual modo é uma especialização). Nesse caso, adicionamos a restrição de que a variável *F* varia sobre funções (um tipo específico de modo).

**Irreflexivity:**  $\forall c, t \neg \text{inheres}(c, c, t)$

**Asymmetry:**  $\forall c, c_1, t \text{ inheres}(c, c_1, t) \rightarrow \neg \text{inheres}(c_1, c, t)$

**Intransitivity:**  $\forall c_1, c_2, c_3, t \text{ inheres}(c_1, c_2, t) \wedge \text{inheres}(c_2, c_3, t) \rightarrow \neg \text{inheres}(c_1, c_3, t)$

**Existential Dependency:**

$\forall c_1, c_2, \exists t_1 \text{ inheres}(c_1, c_2, t_1) \rightarrow \forall t (\text{exists}(c_1, t) \rightarrow \text{exists}(c_2, t) \wedge \text{inheres}(c_1, c_2, t))$

$\text{characterized\_by}(C, F) =_{\text{def}}$

$\forall c \exists t_1 \text{ instance\_of}(c, C, t_1) \rightarrow \forall t (\text{instance\_of}(c, C, t) \rightarrow \exists f (\text{instance\_of}(f, F, t) \wedge \text{inheres}(f, c, t)))$

Podemos então proceder para a descrição formal da relação conduzida por entre dois permanentes *c* e *c<sub>r</sub>*. Essa relação é caracterizada aqui usando-se as três fórmulas abaixo. A primeira delas estabelece que se *c* é conduzido por *c<sub>r</sub>*, então existe um proceddo (de condução) *p* que ocorre no devido tempo e que, em todos os instantes em que esse processo ocorre, tanto *c* como *c<sub>r</sub>* participam desse processo. Ademais, a fórmula estabelece que é inerente a *c<sub>r</sub>* durante o esse processo inteiro e somente durante esse processo. Criando essa fórmula com a condição de dependência existencial para a relação de inerência definida acima, temos que participar do processo de condução é uma condição essencial para *c*.

$\text{conducted\_by}(c, c_r) \rightarrow \exists p, t_1 \text{ occurring\_at}(p, t_1) \wedge \forall t (\text{occurring\_at}(p, t) \rightarrow \text{has\_participant}(p, c, t) \wedge \text{has\_participant}(p, c_r, t)) \wedge (\forall t_2 \text{ inheres}(c, c_r, t_2) \leftrightarrow \text{occurring\_at}(p, t_2))$

A próxima fórmula estabelece que em todos os instantes em que *c* é inerente a *c<sub>r</sub>* (ex. em todos os instantes em que *c* existe), *c* ocupa uma região espacial *r<sub>1</sub>* que é uma parte adequada da região espacial *r* ocupada pelo seu portador (o condutor). Ademais a fórmula estabelece que dado um instante de tempo *t*, só existe uma região ocupada por *c* nesse instante (analogamente para o condutor *c<sub>r</sub>*). Finalmente, a fórmula (indiretamente) estabelece que durante o processo de condução *p* (ex. durante a vida de *c*), *c* ocupa todas as partes próprias de *r* mas também que nenhuma parte própria de *r* é ocupada por *c* mais do que uma vez durante o processo *p*.

$\text{conducted\_by}(c, c_r) \rightarrow \forall t (\text{inheres}(c, c_r, t) \rightarrow \exists r, r_1 (\text{located\_in}(c_r, r, t) \wedge \text{located\_in}(c, r_1, t) \wedge \text{part\_of}(r_1, r) \wedge \forall r_2, r_3 (\text{located\_in}(c_r, r_2, t) \wedge \text{located\_in}(c, r_3, t) \rightarrow (r_2 = r) \wedge (r_3 = r_1)) \wedge \forall r_4 (\text{part\_of}(r_4, r) \rightarrow \exists! t_1 \text{ inheres}(c, c_r, t_1) \wedge \text{located\_in}(c, r_4, t_1))))$

Finalmente, a fórmula a seguir estabelece que dados dois instantes *t<sub>1</sub>* e *t<sub>2</sub>* tal que *c* é inerente a *c<sub>r</sub>* tanto em *t<sub>1</sub>* como em *t<sub>2</sub>* e que *t<sub>1</sub>* é o instante imediatamente anterior a *t<sub>2</sub>* então, em cada desses instantes, *c* ocupa regiões adjacentes.

$\text{conducted\_by}(c, c_r) \rightarrow \forall t_1, t_2, r_1, r_2 (\text{inheres}(c, c_r, t_1) \wedge \text{inheres}(c, c_r, t_2) \wedge \text{located\_in}(c, r_1, t_1) \wedge \text{located\_in}(c, r_2, t_2) \wedge \text{immediately\_earlier}(t_1, t_2) \rightarrow \text{adjacent\_to}(r_1, r_2))$

Consulte abaixo a relação de *immediately\_earlier* (imediatamente anterior) existente entre dois instantes de tempo.

$\text{immediately\_earlier}(t_1, t_2) =_{\text{def}} \text{earlier}(t_1, t_2) \wedge \neg \exists t (\text{earlier}(t, t_2) \wedge \text{earlier}(t_1, t))$

A versão universal-nível da relação *conducted\_by* (conduzida por) é a seguinte.

$\text{conducted\_by}(C, C_r) =_{\text{def}}$

$\forall c \exists t \text{ instance\_of}(c, C, t) \rightarrow \exists c_r \text{ instance\_of}(c_r, C_r, t) \wedge \text{conducted\_by}(c, c_r)$

## O eletrocardiograma

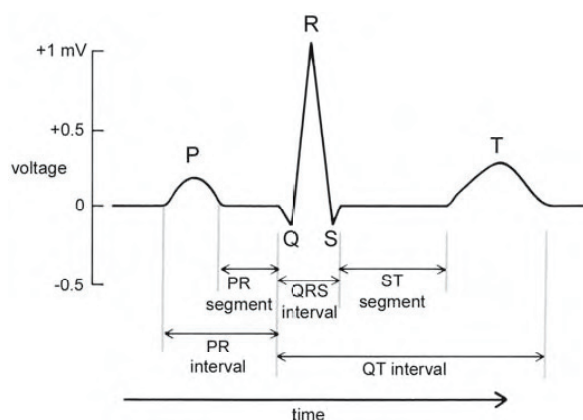
Uma vez estabelecidas as bases de anatomia e fisiologia, podemos finalmente nos concentrar na nossa análise ontológica no ECG em si. O ECG (em alemão, electrokardiogram, EKG) foi provavelmente o primeiro sinal diagnóstico a ser estudado com o propósito de interpretação automática por programas de computadores (Geselowitz 1989). A razão para tal interesse em registros computacionais de ECG é que a análise do formato de onda do ECG pode ajudar a identificar uma ampla gama de doença cardíacas, que são distintas por modificações nas *formas elementares* de ECG.

Do lado do *paciente*, o ECG é obtido no contexto de uma *sessão de registro/gravação* em que um *aparelho de gravação/registro* é usado para realizar *observações* uniformemente espaçadas no tempo para medir diferenças elétricas potenciais (d.p.) em torno na *superfície da pele* do paciente e com *amostras* produzidas no resultado. Conforme discutido na seção anterior, essas d.p.'s são resultado da atividade elétrica do coração. As observações são feitas ao mesmo tempo a partir de *disposição de eletrodos* em locais diferentes para fornecer pontos de vista múltiplos da atividade do coração (chamados *condutores*). Essas observações correlatas formam *séries de observações* correlatas. Cada série de observação então produz uma *seqüência amostral*.

Agora mudando para a perspectiva do médico, é válido mencionar que os batimentos cardíacos são espe-

lhados em ciclos cardíacos que compõem o formato de onda do ECG. Um ciclo canônico, conforme apresentado por W. Einthoven, tem *ondas* (subtipos de formas elementares) chamadas PQRST. Elas são delineadas como *ondas* P, a soma mereológica de ondas Q, R e S (o chamado complexo QRS), e *onda* T (ver Figura 5). A

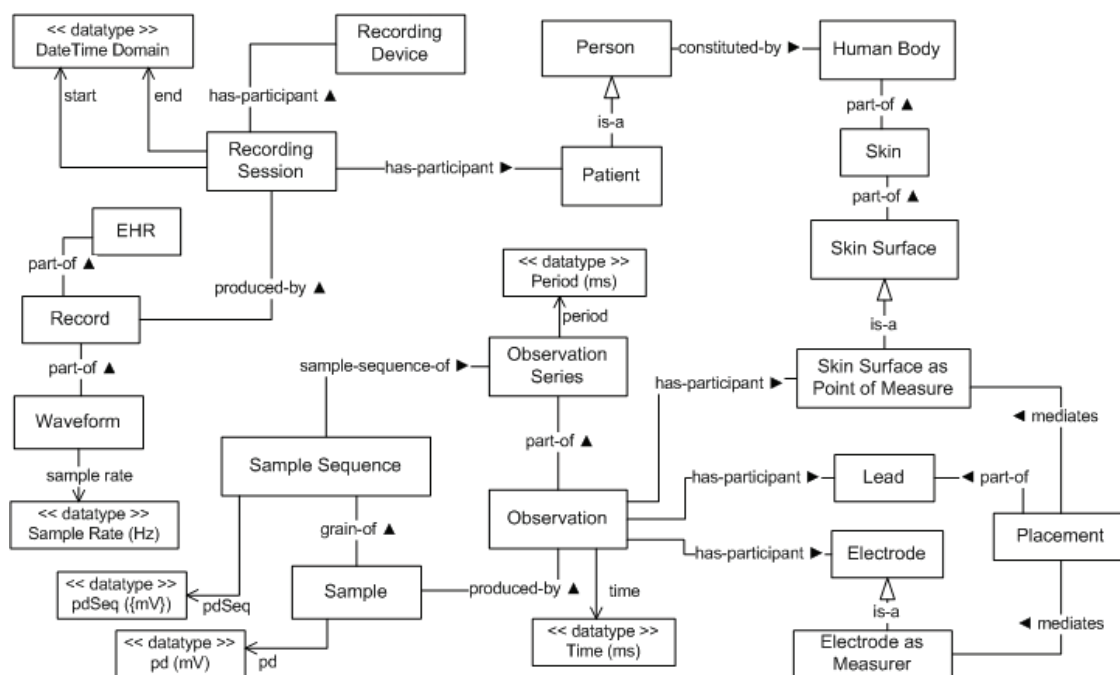
onda P e o complexo QRS mapeiam a despolarização dos átrios e ventrículos, respectivamente. As contrações do miocárdio no átrio e ventrículo começam normalmente no pico dessas ondas. A onda T, por sua vez, mapeia a repolarização dos ventrículos<sup>5</sup> (Geselowitz 1989, Guyton & Hall 2006).



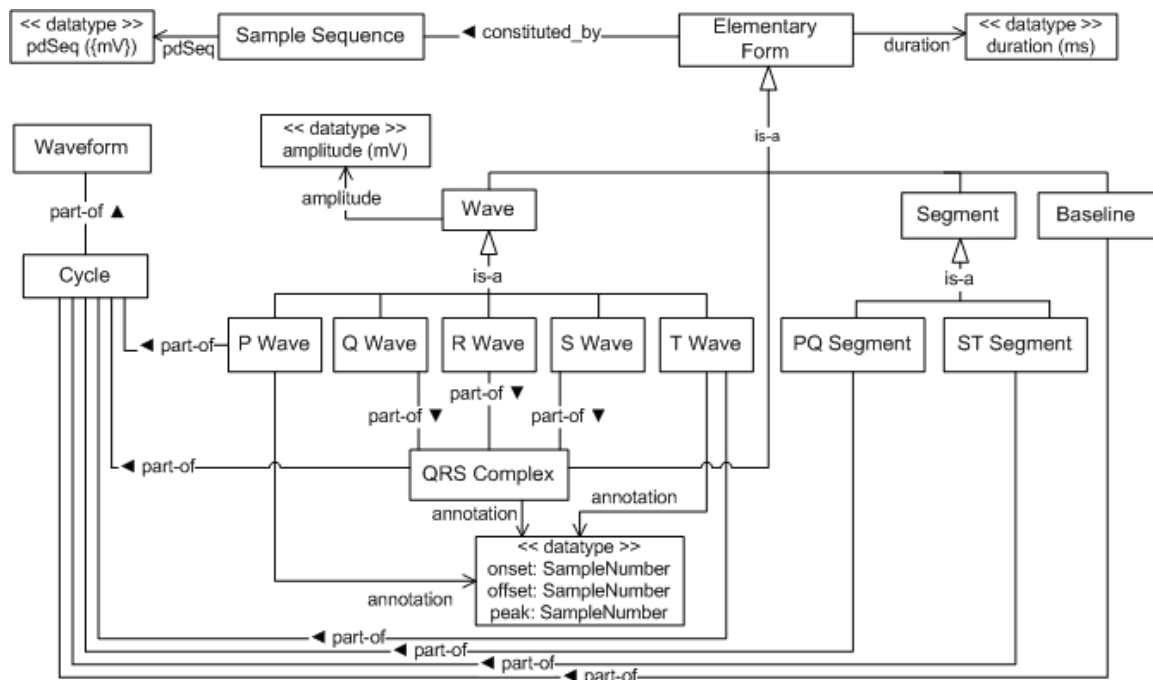
**Figura 5** - Um ciclo típico (refletindo um batimento cardíaco) no formato de onda do ECG (fonte: Laske; Iaizzo 2005). Dois ciclos são conectados pela linha de base, que reflete o estado de repouso do coração.

As Figuras 6 e 7 fornecem representações gráficas do ECG dos lados do paciente e do médico respectivamente. Na nossa ontologia, esses dois modelos são dotados de uma axiomatização LPO correspondente, as quais, por brevidade, são omitidas aqui. Aqui, nessas figuras, os modelos destinam-se exclusivamente à representação

visual do domínio do ECG sem nenhuma intenção de ser completo. Esses modelos são baseados em evidências presentes em livros médicos mas também sintetizam questões presentes em padrões de ECG atuais (deixando de fora aspectos tecnológicos).



**Figura 6** - Modelo de ECG do lado do paciente. Ele ou ela participam de uma sessão de registro destinada a produzir um registro de ECG. Nessa sessão, diversas observações são feitas por eletrodos posicionados na superfície da pele do paciente. Cada observação produz uma amostra, que é um grânulo da sequência amostral (um conjunto ordenado de amostras). Por uma questão de brevidade, omitimos aqui a representação de diversas configurações de colocação de eletrodos em regiões específicas da superfície da pele que compõem os condutores de ECG (ver I, II, III, aVL, aVR, aVF, V1, ..., V6).



**Figura 7** - Modelo de ECG do lado do médico. Ele ou ela pode analisar o formato de onda de ECG, ciclo por ciclo. Cada um representa um batimento cardíaco. O ciclo tem como partes muitas formas elementares diferentes. Uma forma elementar é constituída de uma seqüência amostral, que é um conjunto ordenado de amostras. Observe que esse modelo não cobre nenhuma anormalidade na ECG.

As noções de constituição e mediação usadas nas relações *constituídas por* e *mediadas* são não triviais (Guizzardi 2005). Por uma questão de brevidade, não daremos suas definições neste texto. Uma discussão profunda destas relações pode ser encontrada em (Masolo et al. 2003) e (Guizzardi 2005, Capítulo 6), respectivamente.

## Do ECG à eletrofisiologia do coração

Agora temos material para unir os domínios do ECG e da eletrofisiologia do coração. A interpretação de um ECG envolve diversos detalhes sutis que normalmente existem tacitamente na mente do cardiologista. Nosso empenho aqui está em fornecer um método capaz de desvelar explicitamente, a primeira vista, o que um ECG mapeia com relação à eletrofisiologia cardíaca canônica. Portanto, apresentamos a relação chamada *mapas* cujo propósito é associar cada uma das formas elementares de ECG que aparecem no ECG para sua realidade eletrofisiológica subjacente. Pode ser definida nos níveis instância e universal como a seguir.

Primeiro, podemos caracterizar formalmente a relação *observation\_series\_of* entre um processo de série de observação  $o$  e um processo (condução)  $p$ . A fórmula abaixo estabelece que se  $o$  é uma série de observação do processo  $p$  então cada observação (atômico) que é parte de  $o$  é uma observação de uma parte de  $p$  (e só pode ser uma observação de um processo que é parte de  $p$ ).

$$\text{observation\_series\_of}(o,p) \rightarrow \forall o_1 (\text{part\_of}(o_1,o) \rightarrow \exists p_1 (\text{part\_of}(p_1,p) \wedge \text{observation\_of}(o_1,p_1)^6) \wedge \forall p_2 (\text{observation\_of}(o_1,p_2) \rightarrow \text{part\_of}(p_2,p)))$$

Na seqüência, estabelecemos que se temos duas observações  $o_1$  e  $o_2$  que são parte de  $o$  e que são observações de partes  $p_1$  e  $p_2$  (partes de  $p$ ), respectivamente, tal que  $o_2$  segue  $o_1$  na série  $o$ , então suas partes de processo observadas respectivas também se seguem da mesma forma (ex.,  $p_2$  follows  $p_1$ ).

$$\text{observation\_series\_of}(o,p) \rightarrow \forall o_1, o_2, p_1, p_2 (\text{part\_of}(o_1,o) \wedge \text{part\_of}(p_1,p) \wedge \text{observation\_of}(o_1,p_1) \wedge \text{part\_of}(o_2,o) \wedge \text{part\_of}(p_2,p) \wedge \text{observation\_of}(o_2,p_2) \wedge \text{follows}(o_2,o_1) \rightarrow \text{follows}(p_2,p_1))$$

A relação *follows* (segue) existente entre dois processos  $p_2$  e  $p_1$  implica que

$$\text{follows}(p_2, p_1) \rightarrow \exists t_1, t_2 (\text{last\_instant}(t_1, p_1) \wedge \text{first\_instant}(t_2, p_2) \wedge \text{earlier}(t_1, t_2))$$

Agora, podemos caracterizar a correspondência entre uma série de observação e uma seqüência de amostras representando essa série. As duas primeiras fórmulas são análogas à fórmula apresentada para a série de observações com duas diferenças importantes. Se  $s$  é uma seqüência amostral da série de observação  $o$ , então: (i) cada amostra em  $s$  é produzida por exatamente uma observação em  $o$ ; (ii) existe uma correspondência direta entre observações em  $o$  e amostras em  $s$ .

$$\text{sample\_sequence\_of}(s,o) \rightarrow \forall s_1 (\text{grain\_of}(s_1,s) \rightarrow \exists o_1 (\text{part\_of}(o_1,o) \wedge \text{produced\_by}(s_1,o_1)) \wedge (\forall o_2 \text{ produced\_by}(s_1,o_2) \rightarrow (o_1 = o_2)))$$

$\text{sample\_sequence\_of}(s,o) \rightarrow \forall s_1, s_2, o_1, o_2 (\text{grain\_of}(s_1, s) \wedge \text{produced\_by}(s_1, o_1) \wedge \text{grain\_of}(s_2, s) \wedge \text{produced\_by}(s_2, o_2) \wedge \text{successor\_of}(s_2, s_1) \rightarrow \text{directly\_follows}(o_2, o_1))$

A relação *successor\_of* (sucessor de) é definida como de costume entre um elemento em uma sequência e um sucessor (direto) desse elemento naquela sequência (seguindo o critério ordenador intrínseco dessa sequência). A relação de *directly\_follows* (segue diretamente) é definida como:

$\text{directly\_follows}(p_2, p_1) =_{\text{def}} \text{follows}(p_2, p_1) \wedge \neg \exists p_3 (\text{follows}(p_3, p_1) \wedge \text{follows}(p_2, p_3))$

Finalmente, podemos definir a relação de *maps* (mapas) entre uma forma elementar *c* e um processo (condutor) *p*:

$\text{maps}(c,p) =_{\text{def}} \exists s,o \text{ constituted\_by}(c,s) \wedge \text{sample\_sequence\_of}(s,o) \wedge \text{observation\_series\_of}(o,p)$

e a relação correspondente em um nível-universal.

$\text{maps}(C, P) =_{\text{def}} \forall c \exists t \text{ instance\_of}(c, C, t) \rightarrow \exists p, t_1 \text{ instance\_of}(p, P, t_1) \wedge \text{maps}(c, p)$

Empregando as noções recém discutidas, damos significado às formas elementares do ECG. Também especificamos um conjunto de regras para reconstruir a partir do formato de onda do ECG os processos eletrofisiológicos correlatos ocorridos sobre permanentes anatômicos. Essas regras fazem uso das nossas representações de função. Como exemplo, considere as regras R1 a R<sup>6</sup> dadas abaixo. Elas dão significado a onda-P baseada na função da direita da Figura 3. Portanto, o que podemos inferir uma vez que tivermos anotado fielmente (portanto, reconhecido) a onda-P?

Primeiramente, cada onda-P *mapeia* um e somente um processo eletrofisiológico da condução do impulso elétrico cardíaco (IEC) em torno do átrio.

(R1)  $\forall c \text{ PWave}(c) \rightarrow \exists p (\text{CEIConductionAroundAtria}(p) \wedge \text{maps}(c, p) \wedge \forall p_1 (\text{maps}(c, p_1) \rightarrow (p_1 = p)))$

Além disso, todo processo como esse é associado a um e somente um IEC e a um e somente um sistema condutor do átrio desempenhando o papel de condutor de IEC. Efetivamente, eles precisam participar em todo o processo. Formalmente (ver R2),

(R2)  $\forall p (\text{CEIConductionAroundAtria}(p) \rightarrow \exists t_1 (\text{occurring}(p, t_1) \wedge \exists! c_1, c_2 (\text{CEI}(c_1) \wedge \text{ConductingSystemOfAtriaAsCEIConductor}(c_2) \wedge \forall t (\text{occurring}(p, t) \rightarrow \text{has\_participant}(p, c_1, t) \wedge \text{has\_participant}(p, c_2, t))))))$

Além disso, para cada processo, existe uma e somente uma função *to conduct CEI* (para conduzir IEC) tal que a última é potencialmente realizada pelo processo. Quer dizer (ver R3), eles são associados um ao outro pela disposição o processo tem que ser a realização da função, mesmo que essa disposição não venha a se tornar real.

(R3)  $\forall p (\text{CEIConductionAroundAtria}(p) \rightarrow \exists! f (\text{toConductCEI}(f) \wedge \text{disp\_realized\_by}(f, p)))$

No entanto, se temos o processo, somos capazes de inferir que (ver R4) havia um estado do mundo SOW1 no qual suas exigências foram satisfeitas (ver Figura 3).

(R4)  $\forall p (\text{CEIConductionAroundAtria}(p) \rightarrow \exists! c_1, c_2, t_{\text{SOW1}} (\text{CEI}(c_1) \wedge \text{first\_instant}(p, t_{\text{SOW1}}) \wedge \text{SANode}(c_2) \wedge \text{exists}(c_1, t_{\text{SOW1}}) \wedge \text{located\_in}(c_1, c_2, t_{\text{SOW1}})))$

O reconhecimento da realização de *to conduct CEI* (para conduzir IEC) depende da notação, se a onda-P à mão é normal ou não. Isso pode ser formalmente descrito por R5 como se segue.

(R5)  $\forall p, c, f ((\text{CEIConductionAroundAtria}(p) \wedge \text{NormalPWave}(c) \wedge \text{toConductCEI}(f) \wedge \text{maps}(c, p) \wedge \text{disp\_realized\_by}(f, p)) \rightarrow \text{actual\_realized\_by}(f, p))$

Nesse caso, podemos então inferir que os objetivos do *to conduct CEI* foram satisfeitos pelo processo de condução de IEC pelo átrio.

(R6)  $\forall p, f ((\text{CEIConductionAroundAtria}(p) \wedge \text{toConductCEI}(f) \wedge \text{actual\_realized\_by}(f, p)) \rightarrow \exists c_1, c_2, c_3, t_{\text{SOW2}} (\text{CEI}(c_1) \wedge \text{ConductingSystemOfAtria}(c_2) \wedge \text{VentricularPartOfAVBundle}(c_3) \wedge \text{last\_instant}(p, t_{\text{SOW2}}) \wedge \text{conducted\_by}(c_1, c_2) \wedge \text{located\_in}(c_1, c_3, t_{\text{SOW2}})))$

## Para onde: uma ontologia de ECG

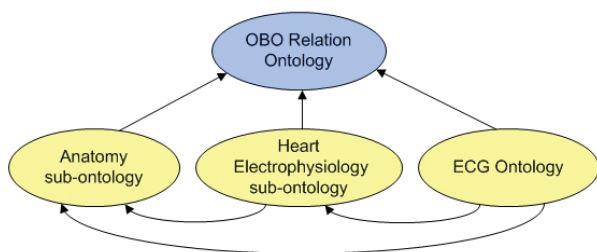
Os resultados do nosso estudo ontológico de Eletrocardiograma foram a fonte do conhecimento de domínio na construção de uma ontologia de ECG. Constitui uma teoria independente de solução de eletrocardiograma, que deve ser reutilizada em diversos aplicativos. Em sua essência, a Ontologia de ECG lida com o que o ECG é tanto do lado do paciente como do médico. Como vimos, isso depende de várias noções relacionadas à eletrofisiologia do coração, que se realiza em entidades anatômicas. O scopo da Ontologia de ECG pode ser definida através das seguintes questões de competência (QC).

QC1. O que compõe essencialmente um registro de ECG?  
QC2. Como é obtido o registro de ECG?



- QC3. O que no formato de onda de ECG é objeto da análise do médico para interpretar um comportamento cardíaco correlato?
- QC4. Para todas as formas elementares, que funções eletrofisiológicas do coração elas mapeiam?
- QC5. Para todas as funções eletrofisiológicas cardíacas, que entidades anatômicas são capazes de realizá-las?
- QC6. Para todas as funções eletrofisiológicas cardíacas, que exigências devem ser satisfeitas para possibilitar sua realização?
- QC7. Para todas as funções eletrofisiológicas cardíacas, que objetivos devem ser satisfeitos para concluir sua realização?

A Ontologia de ECG é então composta por duas subontologias extras, a saber, a anatomia para a subontologia de ECG e eletrofisiologia do coração. Também importa a Ontologia Relacional (OR) OBO, ver Figura 8.



**Figura 8** - Importar relacionamentos da Ontologia de ECG. As setas apontam em direção à ontologia sendo importada. A Ontologia Relacional OBO é importada aqui para nos dar relações básicas usadas nas outras.

A Ontologia de ECG foi implementada na linguagem de codificação ontológica OWL DL e sua extensão SWRL. A versão atual da Ontologia de ECG implementada está disponível para download no website do projeto<sup>7</sup>.

## Discussão

### Competência

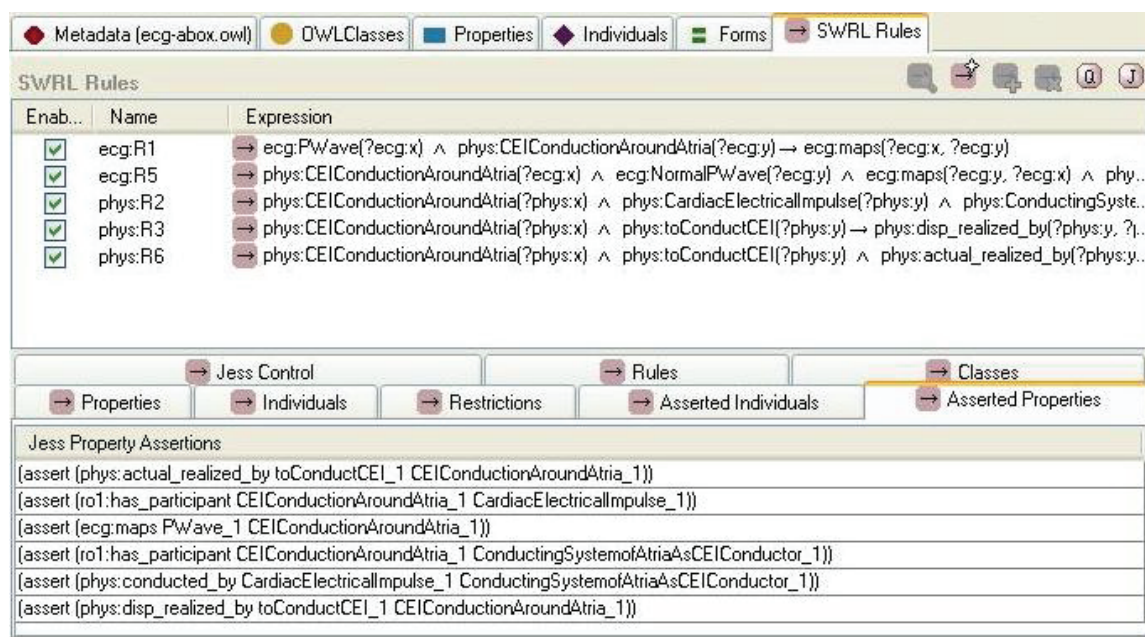
As QC's da Ontologia de ECG foi axiomatizada e também implementada em OWL DL/SWRL. Como tal, elas compreendem uma forma de avaliação aproveitando os serviço de raciocínio. Damos dois exemplos abaixo com relação a axiomatização da QC4 e QC7 (mais uma vez tomando a onda-P como exemplo). Elas são respondidas por raciocínio automatizado como demonstrado na Figura 9.

QC4. Para todas as formas elementares, que funções eletrofisiológicas do coração elas mapeiam?

$$\forall c \ ( \text{Pwave}(c) \rightarrow \exists p \ ( \text{ImpulseConductionAroundAtria}(p) \wedge \text{maps}(c, p) ) )$$

QC7. Para todas as funções eletrofisiológicas cardíacas, que objetivos devem ser satisfeitos para concluir sua realização?

$$\forall f, c, c_1, p, t_{\text{SOW2}} \ ( \ ( \text{toConductCEI}(f) \wedge \text{ConductingSystemOfAtria}(c) \wedge \text{characterized\_by}(c, f) \wedge \text{CEIConductionAroundAtria}(p) \wedge \text{actual\_realized\_by}(f, p) \wedge \text{last\_instant}(p, t_{\text{SOW2}}) )$$

$$\wedge \text{CEI}(c_1) \wedge \text{has\_participant}(p, c_1, t_{\text{SOW2}}) ) \rightarrow ( \text{conducted\_by}(c_1, c) \wedge \exists c_2, c_3 \ ( \text{VentricularPartOfAVBundle}(c_2) \wedge \text{ConductingSystemOfHeart}(c_3) \wedge \text{part-of}(c, c_3) \wedge \text{part-of}(c_2, c_3) \wedge \text{located\_in}(c_1, c_2, t_{\text{SOW2}}) ) )$$


**Figura 9** - Imagem de um serviço de raciocínio que responde as QC's da Ontologia de ECG usando sua própria implementação OWL DL/SWRL.

## Aplicabilidade

Uma ontologia de domínio independente de aplicativo como a Ontologia de ECG pode ser aplicada para muitos fins diferentes. Os exemplos incluem o que será brevemente discutido abaixo: (i) gerenciamento da heterogeneidade da informação e (ii) raciocínio sobre os universais e particulares, semelhante ao que foi referido por Burgun (2006).

## Gerenciamento da heterogeneidade da informação

Uma vez que supomos que a Ontologia de ECG é muito importante na representação *do que é ECG* e somente isso (ex., independente de questões tecnológicas), ela pode ser usada para dar suporte ao design de versões interoperáveis de formatos de dados de ECG como SCP-ECG, FDA XML e HL7. Tomando a Ontologia de ECG como referência, as entidades presentes nesses formatos de dados podiam ser semanticamente espelhados nos universais ontológicos, em vez de ser objeto de mapeamentos em pares. Com isso, os formatos de dados de ECG devem estar de acordo com o desiderato CIMINO, a saber: (i) *não incerteza*, as entidades que formam os nodos do formato de dados devem corresponder a pelo menos um universal; e (ii) eles devem corresponder a não mais que um universal, isto é, *não ambigüidade*. Como a axiomatização da Ontologia de ECG permite pouca liberdade tanto para a incerteza quanto para a ambigüidade, essas soluções iriam pelo menos forçar os formatos de dados para fazer suas suposições explícitas. Além disso, essa proposta é custo-eficiente, já que  $n$  formatos de dados exigem  $n$  mapeamentos para uma ontologia de referência, enquanto mapeamentos em pares  $n(n-1)/2$  seriam exigidos (Burgun 2006).

## Raciocínio sobre universais e particulares

A Ontologia de ECG delineada aqui foi integralmente implementada em uma linguagem de codificação ontológica. No nosso projeto, nós usamos o OWL DL e sua extensão SWRL em virtude do instrumento de raciocínio produzido em série disponível, ex. Pellet (Sirin et al. 2007). O arquivo OWL DL/SWRL é então suscetível para ser eficientemente usado para o raciocínio automatizado, embora não mantendo a axiomatização da ontologia (ver a Seção “Métodos”).

A Ontologia ECG representa um modelo canônico da anatomia do coração e um modelo canônico da eletrofisiologia do coração. O modelo ECG, antagonicamente, pode ser preenchido por qualquer instância de registro de ECG real. Entretanto, um complexo QRS deformado (possivelmente indicando alguma patologia) não teria um impulso elétrico cardíaco não canônico para o qual mapear. Dada essa elucidação, vamos lançar luz sobre o que pode ser feito. Usando-se uma instância de um registro<sup>8</sup> de ECG normal (uma ferramenta de estudo), podemos reconstruir a eletrofisiologia (canônica) por trás. Portanto, de uma instância normal de um complexo QRS (fielmente anotado), somos capazes de reconstruir

o impulso elétrico cardíaco por trás e a anatomia no qual ocorreu.

Uma aplicação característica para isso é um sistema para dar suporte ao aprendizado em eletrofisiologia cardíaca e ECG. De fato, criamos esse sistema que usa uma versão anterior da Ontologia de ECG (ver Gonçalves et al. 2009). Nesse sistema, uma tabela de ECG é criada a partir de um arquivo OWL de ECG (com dados preenchidos para os indivíduos da ontologia). Além disso, uma ilustração do sistema de condução do coração é capaz de mostrar animações em resposta a cliques do usuário tanto neste último como em um ponto da tabela de ECG. Esses cliques pedem um procedimento de raciocínio que enfatiza uma forma elementar no formato de onda de ECG e seleciona o fenômeno de condução correlato a ser animado.

Tudo isso também poderia ser feito com um registro de ECG não canônico se tivéssemos um modelo não canônico de fisiologia para reconstruir. Até onde investigamos, isso parece ser possível estendendo a subontologia da eletrofisiologia do coração para lidar com a imprecisão (incerteza) da realização das funções eletrofisiológicas do coração.

## Limitações e trabalhos futuros

Conforme exposto na discussão acima, as limitações dos nossos resultados são em maioria devidas à complexidade em lidar com aspectos fisiológicos do coração humano. Isso é particularmente difícil quando questões fenotípicas devem ser cobertas. Portanto, um forte empenho de pesquisa é demandado para estender a teoria ontológica de ECG apresentada aqui com tal propósito.

Entre as direções enfrentadas para trabalhos futuros nós incluímos: (i) a liberação de uma versão atualizada do sistema baseado em raciocínio da web proposto em (Gonçalves et al. 2009) para colocar em uso online a Ontologia de ECG implementada; (ii) a investigação de como capturar do ECG a imprecisão inerente de se uma função eletrofisiológica do coração foi de fato realizada. Nós acreditamos que este é um ponto de partida importante para lidar com casos patológicos específicos.

## Conclusões

Neste artigo, fornecemos um relato ontológico do exame cardiológico ECG e sua correlação com a eletrofisiologia do coração humano. A Ontologia de ECG delineada aqui constitui uma teoria de domínio axiomatizada fundamentada em uma base ontológica de princípios. A aplicabilidade dessa ontologia também foi ilustrada por dois propósitos diferentes, a saber, o gerenciamento da heterogeneidade dos padrões de formato de dados de ECG e raciocínio automatizado. Com este último em mente, traduzimos os modelos e a fórmula LPO que apresentamos aqui em uma linguagem de codificação ontológica OWL DL com sua extensão SWRL.

Como parte de um empenho de pesquisa mundial em andamento para fomentar as representações ontológicas da realidade biomédica, nosso empreendimento é apropriado. Naturalmente, nossa investigação ontológica de ECG pode ser elaborada para aumentar, digamos, o *nível de detalhe* e até mesmo cobrir uma lacuna eventual. Enquanto isso, o desafio da integração da ontologia ainda é difícil mesmo nesse campo de pesquisa cada vez mais firme chamado ontologia biomédica. Entretanto, lutando para manter a conformidade com as iniciativas correlacionadas, nós nos empenhamos nesse sentido. De qualquer maneira, o caso é que “O valor de qualquer tipo de dados [ou ontologia] é muito intensificado quando ele existe de uma forma que permite que seja integrado com outros dados [ou ontologia]” (Smith et al. 2007). Nesse espírito, a Ontologia de ECG pode ser entendida como uma contribuição a ser agregada no empenho ontológico biomédico.

## Agradecimentos

Essa pesquisa foi parcialmente apoiada pelos projetos MODELA (financiado pela agência de fomento brasileiro - Facitec) bem como os projetos INFRA-MODELA e software livre e interoperabilidade em saúde (financiado pela agência de financiamento Fapes).


## Notas

1. Temos ciência de duas iniciativas de pesquisa em andamento que se encaixam em linhas gerais na eletrofisiologia do coração. Rubin et al. (2006) apresentam uma metodologia simbólica, voltada para a ontologia, para representar um modelo fisiológico da circulação como uma alternativa aos modelos matemáticos comumente empregados. Por sua vez, Cook et al. (2004) estão se empenhando em uma extensão do FMA para cobrir a fisiologia.
2. Protocolo de Comunicação Padrão para Eletrocardiografia auxiliada pelo computador <http://www.openecg.net/>.
3. Especificação de Design do Formato De Dados FDA XML <http://xml.coverpages.org/FDA-EGC-XMLData-Format-C.pdf>.
4. Mensagem de Notação V3 HL7ECG <http://www.hl7.org/V3AnnECG>.
5. A repolarização do átrio não pode ser vista no formato de onda de ECG já que seus potenciais resultantes são pequenos em amplitude e então excedido pelo complexo QRS. Uma onda U também é frequentemente mencionada, mas sua origem ainda não é completamente conhecida.
6. Aqui supomos que se *observation\_of* (o,p) então o processo ocorre sincronicamente ou após o processo p. Intuitivamente, não pode haver observação do futuro.
7. <<http://nemo.inf.ufes.br/biomedicine/ecg.html>>
8. A Physionet (Goldberger 2000), por exemplo, fornece padrões de dados ECG com anotações feitas por outros médicos ou programas de computador. Essas anotações são em geral para marcar e classificar as formas elementares de ECG (ex. a onda-P, o complex QRS e por aí vai).

## Referências bibliográficas

- Bittner T, Donnelly M. Logical properties of foundational relations in bio-ontologies. *Artificial Intelligence in Medicine*. 2007; 39(3):197-216. [doi: 10.1016/j.artmed.2006.12.005]
- Burek P. et al. A top-level ontology of functions and its application in the Open Biomedical Ontologies. *Bioinformatics*. 2006; 22(14):e66-e73. [doi: 10.1093/bioinformatics/btl266]
- Burgun A. A desiderata for domain reference ontologies in Biomedicine. *J Biomed Info*. 2006; 39(3):307-13. [doi: 10.1016/j.jbi.2005.09.002]
- Cimino J. Desiderata for controlled medical vocabularies in the twenty-first century. *Meth Info Med*. 1998; 37(4-5): 394-403.
- Cook D et al. Evolution of a Foundational Model of Physiology: Symbolic representation for functional bioinformatics. In: *World Congress on Medical Informatics*, 11<sup>th</sup>, San Francisco, USA, 2004, Proceedings. IOS Press. 2004:336-40.
- Donnelly M, Bittner T, Rosse C. A formal theory for spatial representation and reasoning in biomedical ontologies. *Artificial Intelligence in Medicine*. 2006; 36(1):1-27. [doi: 10.1016/j.artmed.2005.07.004]
- Geselowitz D. On the Theory of the Electrocardiogram. *Proceedings of the IEEE*. 1989; 77(6):857-76. [doi: 10.1109/5.29327]
- Goldberger A et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*. 2000;101(23): e215-e220.
- Gonçalves B, Guizzardi G, Pereira Filho JG. An electrocardiogram (ECG) domain ontology. In: *Workshop on Ontologies and Metamodels for Software and Data Engineering*, 2nd, João Pessoa, Brazil, 2007. Proceedings. 2007:68-81.
- Gonçalves B, Zamborlini V, Guizzardi G, Pereira Filho JG. An ontology-based application in heart electrophysiology: Representation, reasoning and visualization on the web. In: *ACM Symposium on Applied Computing (SAC 2009)*, 24th, Hawaii, USA, 2009. Proceedings. 2009.
- Guarino N, Welty C. Evaluating ontological decisions with OntoClean. *Commun ACM*. 2002; 45(2):61-5. [doi: 10.1145/503124.503150]
- Guizzardi G. Ontological foundations for structural conceptual models. *Telematica Institute Fundamental Research Series*, Vol.15, Universal Press, The Netherlands. 2005. Available at: <<http://purl.org/utwente/50826>>.
- Guizzardi G, Wagner G. Using the Unified Foundational Ontology (UFO) as a foundation for general conceptual modeling languages. In: Poli R, editors, *Theory and Application of Ontologies*, vol. 2, Springer-Verlag: Berlin; 2009.



- Guizzardi G, Halpin T. Ontological foundations for Conceptual Modeling. *J Appl Ontology*. 2008; 3(1-2):1-12. IOS Press. [doi: 10.3233/AO-2008-0049]
- Guyton A, Hall J. Textbook of medical physiology. 11th edition. Elsevier Saunders:Philadelphia; 2006.
- Heller B, Herre H. Ontological categories in GOL. *Axiomathes*. 2004; 14(1):57-76.
- Horrocks I et al. OWL rules: A proposal and prototype implementation. *J Web Semantics*. 2005; 3(1):23-40. [doi: 10.1016/j.websem.2005.05.003]
- Laske T, Iaizzo P. The cardiac conduction system. In: Iaizzo P, editors. *Handbook of cardiac anatomy, physiology, and devices*. Humana Press: New Jersey; 2005.
- Masolo C. et al. *Ontology Library: WonderWeb Deliverable D18*. 2003. Available at: <www.loa-cnr.it/Papers/D18.pdf>
- Kumar A, Smith B, Novotny D. Biomedical Informatics and granularity. *Comparative and Functional Genomics*. 2004; 5(6-7):501-508. [doi: 10.1002/cfg.429]
- Rector A, Rogers J, Bittner T. Granularity, scale and collectivity: when size does and does not matter. *J Biomed Info*. 2006; 39(3):333-49. [doi: 10.1016/j.jbi.2005.08.010]
- Rosse C, Mejino J. A reference ontology for bioinformatics: The Foundational Model of Anatomy. *J Biomed Info*. 2003; 36(6):478-500. [doi: 10.1016/j.jbi.2003.11.007]
- Rubin D et al. Ontology-based representation of simulation models of physiology. In: *AMIA Annual Symposium*, Washington DC, USA, 2006. Proceedings; 2006. p. 664-68.
- Schulz S, Hahn U. Towards the ontological foundations of symbolic biological theories. *Artificial Intelligence in Medicine*. 2007; 39(3):237-50. [doi: 10.1016/j.artmed.2006.12.001]
- Schulz S et al. SNOMED reaching its adolescence: Ontologists' and logicians' health check. *Int J Med Info*. 2008. [doi: 10.1016/j.ijmedinf.2008.06.004]
- Sirin E et al. Pellet: a practical OWL-DL reasoner. *J Web Semantics*. 2007; 5(2):51-3. [doi: 10.1016/j.websem.2007.03.004]
- Smith B. From concepts to clinical reality: An essay on the benchmarking of biomedical terminologies. *J Biomed Info*. 2006; 39(3):288-98. [doi: 10.1016/j.jbi.2005.09.005]
- Smith B et al. Relations in biomedical ontologies. *Gen Biol*. 2005; 6(5):R46. [doi: 10.1186/gb-2005-6-5-r46]
- Smith B et al. The OBO foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnol*. 2007; 25(11):1251-5. [doi: 10.1038/nbt1346]
- Yu A. Methods in biomedical ontology. *J Biomed Info*. 2006; 39(3):252-66. [doi: 10.1016/j.jbi.2005.11.006]
- Zamborlini V, Gonçalves B, Guizzardi G. Codification and application of a well-founded heart-ECG ontology. In: *Workshop on Ontologies and Metamodels for Software and Data Engineering*, 3rd, Campinas, Brazil, 2008. Proceedings. 2008. 

## Sobre os autores

### *Bernardo Gonçalves*

Bacharel em Ciência da Computação pela Universidade Federal do Espírito Santo (UFES) em 2006. Atualmente, é estudante de mestrado em Ciência da Computação na mesma universidade sob supervisão de Giancarlo Guizzardi. Ele trabalha com a aplicação da Ontologia Formal à Biomedicina. Seus interesses científicos incluem a Ontologia Formal, Lógica Aplicada IA Simbólica e Modelagem Conceitual.

### *Veruska Zamborlini*

Bacharel em Ciência da Computação na Universidade Federal do Espírito Santo (UFES). Atualmente, é estudante de mestrado em Ciência da Computação na mesma universidade sob supervisão de Giancarlo Guizzardi. Seus interesses científicos incluem questões de Lógica, Modelagem Conceitual e Ontologia Formal.



*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 contêm 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<sup>1</sup>, 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 consequência, 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 satisfa-

zer 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 frequência 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 ambígua, (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.

Ranganathan 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 consequência, 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 a 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<sup>2</sup>. 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, Hjørland 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 Tripanosomatídeos 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.

Hjørland (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 Laboratório, 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<sup>3</sup>.

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 GARS (Davila et al. 2005)<sup>4</sup>, 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 (Ranganathan 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 Bio-medicina, 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 sofre 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.

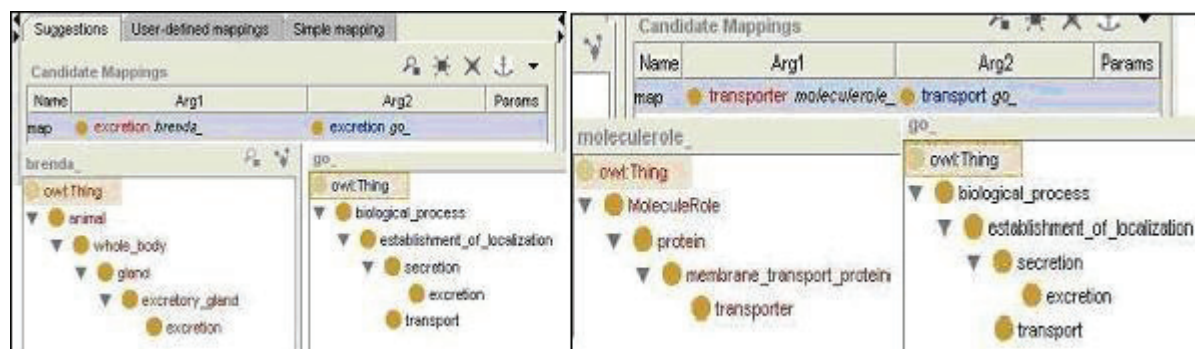


Figura 1 - Falsos positivos sugeridos para mapeamento pela ferramenta Prompt



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 <i>transporter</i> na ontologia <i>system biology</i> 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 definições, 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 definitivo 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, fenômeno.
(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 pesquisas 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... Marília, 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 establishment 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 PE, Grisard E C. GARS: 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 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 common 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 LE. 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 analysis 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 Tripanosomatídeos 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 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 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, 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.

## *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.



*Artigos originais*

## **Extração automática de termos compostos para construção de ontologias: um experimento na área da saúde**

DOI: 10.3395/reciis.v3i1.244pt



*Lucelene Lopes*

Universidade Católica do  
Rio Grande do Sul, Porto  
Alegre, Brasil  
lucelene.lopes@pucrs.br



*Renata Vieira*

Universidade Católica do  
Rio Grande do Sul, Porto  
Alegre, Brasil  
renata.vieira@gmail.com

*Maria José Finatto*

Universidade Federal do Rio Grande do Sul, Porto  
Alegre, Brasil  
mfinatto@pq.cnpq.br

*Adriano Zanette*

Universidade Federal do Rio Grande do Sul, Porto  
Alegre, Brasil

*Daniel Martins*

Universidade Católica do Rio Grande do Sul, Porto  
Alegre, Brasil

*Luiz Carlos Ribeiro Jr*

Universidade Católica do Rio Grande do Sul, Porto  
Alegre, Brasil

### **Resumo**

Neste artigo mostramos o uso da ferramenta OntoLP no processo de construção de ontologias em um experimento na área da Saúde. Especificamente, faz-se a extração de termos com base em um corpus da área de Pediatria. Comparamos o resultado obtido pela ferramenta com os resultados de referência de uma lista de termos obtida manualmente. Nessa comparação, são analisados bi-gramas e tri-gramas obtidos através de diferentes métodos. Concluímos o trabalho observando as vantagens do processamento com inclusão de informação lingüística complexa, como análise sintática e semântica.

### **Palavras-chave**

processamento de linguagem natural; ontologias; construção de ontologias para área da saúde; extração semi-automática de termos

## Introdução

O desenvolvimento da informatização e também a evolução constante dos meios para armazenamento de grandes massas de dados, nos mais diversos setores da sociedade, deram origem a uma significativa quantidade de bases digitais e fontes de dados. A maior parte dessas disponíveis na internet, e das mais variadas espécies, tais como textos, imagens, vídeos, serviços, hipertextos etc. Nesse sentido, faz-se necessário aprimorar modelos descritivos do conhecimento disponibilizado por esses recursos para que sejam recuperados quando necessários. A falta de padronização na representação de conhecimento pode dificultar a compreensão do conteúdo das diversas bases, inviabilizando, conseqüentemente o seu uso. Como uma alternativa para resolver esse problema, a área de sistemas de informação e ciência da computação tem adotado o uso de representação de conhecimento por ontologias (Gomez-Perez et al. 2004).

Ontologias têm sido empregadas como forma de conceituar, estruturar e representar, em um documento, o conhecimento de um domínio de forma que possa ser compartilhado. Esta prática tem sido adotada em vários domínios, e em especial na Biologia, Bioinformática, Biomedicina e na Medicina, que é o domínio de conhecimento explorado nesse artigo. Entretanto, é sabido que ontologias têm um processo de construção trabalhoso, que exige muito tempo e esforço, principalmente na sua utilização em larga escala (Brewster et al. 2003). Uma solução para isso é investir em pesquisas para que se consiga automatizar a tarefa de construção de ontologia de domínios específicos (Buitelaar et al. 2003). Essas pesquisas consideram, muitas vezes, as bases de texto como fontes de conhecimento. Essas fontes, por sua vez, estão expressas em diferentes idiomas, fazendo com que métodos baseados no uso de informações lingüísticas sejam desenvolvidos para as diferentes línguas.

Nesse contexto, apresentamos um estudo voltado especificamente para a língua portuguesa considerando o domínio da Medicina, em particular a área de Pediatria. Neste estudo destacamos a comparação de alternativas de abordagem para identificação de termos compostos (conceitos expressos em mais de uma palavra), considerando que esta é apenas a etapa inicial no processo de construção de ontologias.

Na área de Medicina, mais especificamente no tema da prevenção e promoção de Saúde, por exemplo, o tratamento automatizado da informação textual tende a ajudar pesquisadores e gestores de políticas de informação a reconhecer os melhores modos de apresentar os dados mais relevantes em função dos objetivos e da situação comunicativa que se tenha. Desenha-se aqui a área de pesquisa já reconhecida, fora do Brasil, como *e-health*. A idéia de compactação e da representação de extratos da informação em ciências da saúde (e seus desafios) tem tido espaço para discussão não só no âmbito do reconhecimento de linguagens e de terminologias científicas que precisam ser “facilitadas” para o leigo, mas também nas próprias publicações médicas. Para tanto, é preciso que sejam desenvolvidos sistemas de processamento da in-

formação e esse desenvolvimento precisa estar a cargo de equipes multidisciplinares compostas por profissionais da saúde, estudiosos da língua e da comunicação em ciências e engenheiros que constroem sistemas informatizados com interfaces práticas e de uso simples.

O Journal of Medical Internet Research – JMIR <<http://www.jmir.org/>>, por exemplo, contempla temas de *e-health*. Os resumos, extratos de texto ou representações de conteúdo são gerados com apoio de softwares ou de programas e têm saída em forma de:

1) texto sintético recortado a partir do todo de um texto fonte ou de grupos de textos;

2) esquemas do tipo mapa conceitual a partir de um texto ou grupos de textos fonte;

3) esquemas de relações hierárquicas de nódulos conceituais em ontologias a partir de um ou de mais textos.

Em meio ao cenário brasileiro, podemos pinçar um exemplo ilustrativo da utilidade de sistemas de *e-health*. No caso do Ministério da Saúde, dada a propaganda institucional veiculada a respeito no final de 2008 e início de 2009, sabemos que, por exemplo, a hanseníase ainda é uma doença de considerável impacto no Brasil. Entretanto, apesar da mobilização do Ministério e da sociedade civil organizada, a população ainda parece refratária às campanhas informativas que divulgam medidas preventivas. Entre vários recursos, as matérias sobre prevenção da hanseníase veiculadas na televisão e mídias impressas parecem não surtir um efeito desejado.

Nesse caso particular, um mapeamento da informação disponível, sobretudo em veículos de popularização de temas de saúde, feito a partir de em grandes massas de dados textuais *on-line*, de textos científicos a textos científicos dirigidos a leigos, pode vir a mostrar, por exemplo, que a palavra “lepra”, por mais carregada negativamente ou estigmatizante que seja, raramente é mencionada nos textos de divulgação que tratam sobre hanseníase. Talvez essa lacuna de inter-relação possa justificar o não entendimento mais imediato das mensagens dirigidas pelo público leigo. Um mapeamento semelhante poderia também mostrar situações de usos de termos conexos e de seus equivalentes mais ou menos populares em textos dirigidos a especialistas ou a semi-especialistas e se haveria, inclusive, alguma ponderação mais ou menos presente a respeito do emprego dessas terminologias, em seus diferentes matizes, por parte da própria comunidade científica.

Assim, seria possível que o gestor percebesse, a partir de um levantamento estatístico da configuração da linguagem em textos que visam informar os leigos, que a linguagem empregada precisaria favorecer *links* nocionais ou lógicos entre um nódulo conceitual X e um nódulo conceitual Y (com a devida informação sobre o caráter das diferentes denominações). Diagnosticada a falta e suprida a informação, um cidadão leigo, como no caso acima, poderia relacionar noções e, acionando sua memória cultural sobre uma dada doença ou risco, poderia colocar-se de modo pró-ativo perante a informação que recebe.

Um outro exemplo de utilidade desses sistemas, mas em outra dimensão, pode revelar como está sendo empregado um termo como “prevalência” pela comunidade de

médicos-pesquisadores em publicações científicas em um intervalo, por exemplo, de 2 anos de publicações ao longo de um corpus de mais de um milhão de palavras. Em um acervo de textos de Pediatria, publicados pelo Jornal de Pediatria <<http://www.jped.com.br>>, vê-se como o emprego dessa expressão pode gerar mal-entendidos quando tende a extrapolar a seara do termo técnico e funde-se à palavra da língua comum, cujo sentido é “o que predomina”, enquanto que, na terminologia médica, trata-se de um termo que corresponde a uma medida estatística em Epidemiologia. Nesses textos, a presença de uma construção como “prevalência de uso de chupeta” pode sinalizar uma informação importante para o editores da revista. Além de casos específicos sobre o emprego de termos técnicos, há outros casos indicados de modo esparso pela própria comunidade médica. São situações que extrapolam as terminologias e alcançam o estatuto de construções recorrentes nos textos, tal como vemos no trabalho *Expressões médicas: falhas e acertos* (Bacelar et al. 2003).

Para além da observação de usos de linguagem, a prospecção da informação científica disponível em grandes acervos de periódicos científicos é importante também porque “o aumento da produção de conhecimento - e, portanto, do número de periódicos - na segunda metade do século passado levou a comunidade de profissionais e pesquisadores a encarar o desafio de desenvolver critérios de qualidade que pudessem orientar os leitores na seleção da melhor evidência científica” (Blank et al. 2006, p.97). Esses tipos de dados, ao serem obtidos e estudados extensivamente em grandes bases de textos disponíveis na internet, oferecem informações valiosas para os gestores e editores de periódicos especializados à medida que possibilitam um retrato da recorrência de temas e de noções de uma dada comunidade científica. Essa comunidade teria, então, a partir das técnicas de exploração e de representação informatizada de bases de conhecimento, acesso a um amplo quadro de informações sobre a sua própria prática de comunicação, o que pode ser útil para ponderar sobre a feição e a circulação do conhecimento produzido.

Os exemplos citados, relativos a simples incidências de palavras em textos, visam ilustrar as possibilidades de uso de sistemas avançados de tratamento de informação textual. Esses sistemas, além da localização da informação propriamente dita, mostram como a informação está configurada lingüística e nocionalmente em diferentes situações. Revelam também que outras unidades de vocabulário acompanham-nas de modo mais freqüente ou raro. São sistemas inovadores que integram buscadores e ordenadores de dados em acervos textuais, que integram estatística lexical, sumarização (sintetização de conteúdos) e ontologias. São ferramentas de seleção de informação que precisam de metodologias qualificadas para o tratamento dos fenômenos da linguagem científica em uso. Nesses sistemas, a cooperação entre profissionais de saúde, de lingüistas e cientistas de computação é uma necessidade.

Neste artigo é apresentada a ferramenta OntoLP (Ribeiro 2008), que visa auxiliar de forma semi-automática os engenheiros de ontologias de língua portuguesa, sejam

eles especialistas do domínio em questão (profissionais da saúde) ou lingüistas. A ferramenta mostra sugestões de termos, conceitos e de organização de hierarquias da ontologia, com base no conhecimento registrado em uma base textual ou corpus de domínio.

Mais especificamente, o artigo apresenta um estudo sobre a análise e identificação de termos compostos, ou seja, termos que contêm duas (bi-gramas) ou mais palavras (n-gramas). No contexto deste trabalho, apenas bi-gramas e tri-gramas são extraídos e isso corresponde à etapa inicial no complexo processo de construção de ontologias. Métodos alternativos de processamento textual são comparados. O estudo é desenvolvido através de uma aplicação na área da saúde, considerando um corpus de Pediatria e uma lista de termos de referência para avaliação dos métodos.

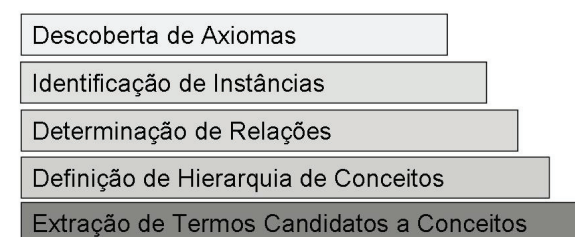
## Ferramenta OntoLP

O processamento de linguagem natural (PLN) se faz presente na construção de ontologias através de textos. O PLN utiliza técnicas lingüísticas baseadas em análises sintáticas e morfológicas dos textos representando as informações em vários níveis.

Como as ontologias criadas ficam restritas a um idioma específico, a dificuldade aumenta no domínio do Português, língua sobre a qual pouco foi feito se comparada com o Inglês. Em resposta a essa dificuldade, novas metodologias para construção semi-automática de ontologias vêm sendo criadas juntamente com as ferramentas de auxílio para essa construção.

OntoLP é uma ferramenta, na verdade um *plug-in* para o editor de ontologias Protégé (Gennari et al. 2002), um editor bastante utilizado na comunidade científica e que dá suporte à construção de ontologias, seguindo as tecnologias da Web Semântica, como por exemplo, a construção de ontologias OWL *Web Ontology Language*, conforme o padrão definido pelo *World Wide Web Consortium* (W3C) (McGuinness et al. 2004).

O processo de construção automática de ontologias é dividido em cinco etapas básicas (Buitelaar 2005): extração de termos candidatos a conceitos de um domínio; identificação da relação hierárquica entre os termos; identificação de relações não-hierárquicas; identificação de instâncias e extração de regras (axiomas). Esse processo pode ser representado em camadas de acordo com a representação feita na Figura 1.



**Figura 1 - Etapas básicas de construção de ontologias.**



Do ponto de vista deste trabalho, que tem por objetivo em longo prazo a construção automática de ontologias, a extração de termos é a tarefa inicial e fundamental, pois os termos extraídos representam os conceitos de uma área específica e são a base para a execução das demais fases.

Para extração de termos, existem três abordagens principais:

- Estatística – os documentos contidos no corpus são vistos como um conjunto de termos e é medida sua frequência de ocorrência;
- Lingüística – os textos são anotados com informações lingüísticas (morfológicas, sintáticas e semânticas) e estas informações são levadas em consideração no processo de extração;
- Híbrida – considera a união das duas abordagens (estatística e lingüística).

A ferramenta OntoLP (Ribeiro 2008) é composta de uma série de métodos híbridos. Estes métodos de extração de termos estão agrupados em duas etapas:

- CorpusXCES: nesta etapa é realizada a anotação do corpus com informações lingüísticas pelo *parser* PALAVRAS (Bick 2000). O corpus anotado contém informações morfológicas, sintáticas e semânticas, representadas no formato XCES/PLN-BR (Ide et al. 2000). O processamento textual para extração de termos é baseado neste corpus anotado. Através da análise morfossintática, informações são adicionadas ao texto original que permitem explorar o emprego de métodos menos ou mais informados lingüisticamente. Nesse trabalho, as informa-

ções lingüísticas empregadas pelos métodos de extração são as categorias gramaticais das palavras (por exemplo, *substantivo, verbos, adjetivos*), categorias semânticas prototípicas (por exemplo, *humanos, animais, doenças*) e a identificação de grupos gramaticais nominais (sintagmas nominais como *aleitamento materno exclusivo*);

- Extração de termos: para essa etapa, são aplicados diferentes métodos que combinam medidas estatísticas de frequência com as informações lingüísticas mencionadas acima, com a finalidade de extrair os termos simples (uni-gramas) e de termos complexos (n-gramas, onde  $n > 1$ ).

No *plug-in* OntoLP, os métodos de extração possuem um conjunto de funcionalidades para auxiliar o engenheiro de ontologias nas etapas que podem ter interação humana. A interface de extração divide-se em três partes: (1) seleção de grupos semânticos; (2) extração de termos simples; e (3) extração de termos compostos.

A Figura 2 mostra de maneira geral como essas funcionalidades aparecem na ferramenta OntoLP, onde os números 1, 2 e 3 indicam as interfaces de extração citadas acima. Na figura, o grupo semântico selecionado é (H - Humano). A janela mais à direita mostra as palavras do grupo em forma de *tag clouds*, onde os termos mais frequentes são destacados, no caso <paciente> e <criança> aparecem como as mais frequentes. O engenheiro de ontologias pode optar por excluir ou não um grupo semântico. Nesse exemplo, o grupo H (humano) não seria excluído devido à sua relevância no corpus analisado.

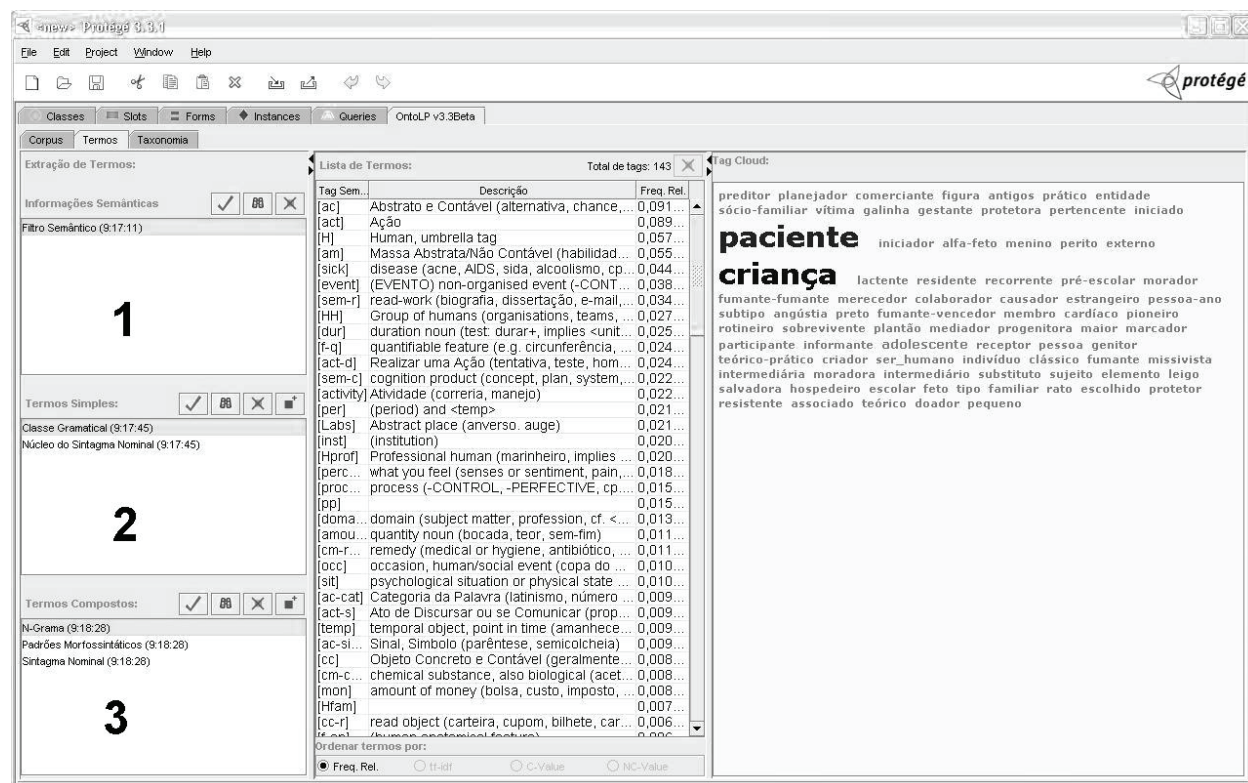
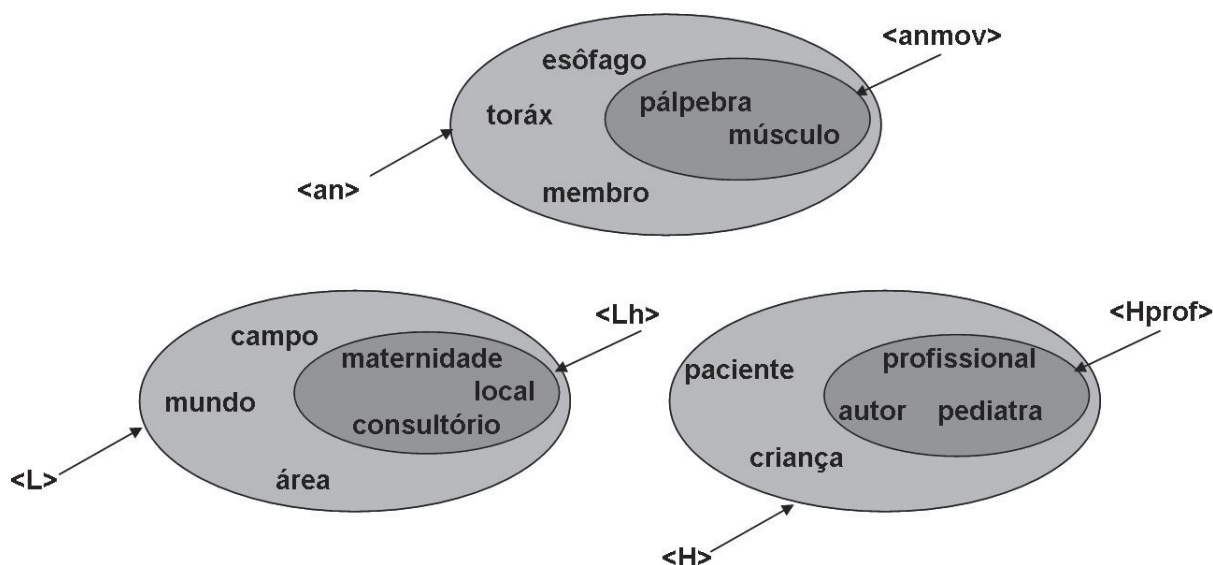


Figura 2 - Interface de extração de termos e suas etapas propostas para tarefa.



A etapa de Seleção de Grupos Semânticos é opcional. A ferramenta mostra ao usuário as informações semânticas que o *parser* (etiquetador) PALAVRAS associa às palavras do corpus. São informações prototípicas que classificam nomes comuns em classes gerais, por exemplo,

a tag “<an>” atribuída ao substantivo “músculo”, indica que a palavra pertence à classe “Anatomia”. A Figura 3 mostra alguns exemplos destes grupos e subgrupos semânticos existentes no corpus de Pediatria analisado neste artigo.



**Figura 3 - Exemplos de grupos e subgrupos semânticos.**

Dessa forma, os substantivos etiquetados com uma mesma *tag* (etiqueta) são agrupados em conjuntos semânticos, por exemplo:

- Grupo <an> (Anatomia): {esôfago, tórax, membro, pálpebra, músculo}
- Grupo <L> (Lugar): {campo, mundo, área, maternidade, local, consultório}
- Grupo <H> (Humano): {paciente, criança, profissional, autor, pediatra}

Os grupos semânticos podem ainda apresentar subdivisões, como pode ser observado na Figura 3: a) anatomia (<an>) e anatomia de movimento (<anmov>); b) lugar (<L>) e lugares funcionais (<Lh>); c) humano (<H>) e humano profissional (<Hprof>).

A ferramenta OntoLP disponibiliza ao usuário o método Filtro por Grupo Semântico, que emprega os seguintes passos:

- As *tags* semânticas presentes no corpus de entrada são extraídas;
- O cálculo de frequência relativa (FR) é aplicado a listas de *tags* semânticas que são apresentadas ao engenheiro ordenadas conforme essa medida;
- O engenheiro exclui os grupos semânticos que considera não ter relação com o domínio representado pelo corpus de entrada.

Esse método pode ser considerado como a construção de uma lista de *stopwords* específicas para um domínio. Essas *stopwords* são itens a não considerar. A

seleção correta dos grupos depende do conhecimento do engenheiro de ontologia sobre a área de conhecimento implicada. A ferramenta auxilia o engenheiro ao mostrar as ocorrências dos termos de cada grupo e sua relevância pelo método de *tags clouds*, ou seja, um método que atribui fontes maiores e destaque textual para termos mais frequentes no corpus.

Após a seleção de grupos semânticos, são executadas as segunda e terceira etapas da extração (extração de termos simples e extração de termos compostos), implementadas por métodos híbridos (estatístico e lingüístico). Na segunda etapa, é realizada a extração de termos simples. O método utilizado é o método de classes gramaticais, detalhado em (Ribeiro 2008).

A terceira etapa, extração de termos compostos, foco deste trabalho, consiste em identificar bi-gramas e tri-gramas. Nessa fase, utilizamos três diferentes métodos, com diferentes complexidades lingüísticas. Em primeiro lugar, é realizada a extração de n-gramas por frequência de co-ocorrência, simplesmente, com aplicação de filtros simples como eliminação de termos com preposições iniciais ou finais. O segundo método leva em consideração a classe gramatical dos termos e padrões de extração, tais como:

- substantivo adjetivo - aleitamento materno
- substantivo preposição substantivo - saturação de oxigênio

O terceiro método realiza a extração dos sintagmas nominais, tais como reconhecidos pelo analisador sintático. Este é um nível de informação lingüístico estrutural

mais complexo e a sua produção requer ferramentas especializadas.

No decorrer das etapas de extração de termos simples e compostos, os métodos recebem a lista de grupos semânticos gerada na primeira etapa, e percorrem o corpus selecionando os termos que fazem parte de pelo menos um grupo presente na lista de entrada. A ferramenta oferece como opção quatro medidas de relevância: FR, *tf-idf* (Manning et al. 1999), NC-Value e C-Value (Frantzi et al. 1998). Os termos extraídos são organizados de forma decrescente com base nos resultados da aplicação dessas medidas, podendo o engenheiro de ontologias analisar e editar a lista final de termos. Cabe ressaltar que as análises feitas neste artigo levaram em conta a Frequência Relativa (FR), a qual considera o número de vezes que um termo aparece no documento dividido pelo total de palavras do documento.

## Experimentos

O corpus utilizado nos experimentos com a ferramenta é composto por 283 textos em português extraídos

do Jornal de Pediatria, num total de 785.448 palavras. Sobre esse corpus foram realizados os experimentos que podem ser descritos conforme a Figura 4. Inicialmente o corpus foi anotado pelo *parser* PALAVRAS, que gerou uma representação XML, sendo convertida para arquivos tipo XCES. Esse corpus anotado em formato XCES foi lido pelo plug-in OntoLP. A partir disso, a extração de termos compostos seguiu as etapas:

- extração de grupos semânticos - através de uma análise manual, o especialista exclui os grupos semânticos considerados não relevantes, os grupos semânticos são criados através da etiquetagem semântica do PALAVRAS e são verificados pelo usuário por meio das *tag clouds*;
- extração dos termos simples; e
- extração dos termos compostos em que foram analisados os métodos: n-gramas, padrões morfossintáticos e o sintagma nominal.

Neste artigo, comparamos os resultados obtidos com e sem a extração de grupos semânticos. Na extração sem exclusão de grupos semânticos, todos os termos são considerados na computação das etapas posteriores.

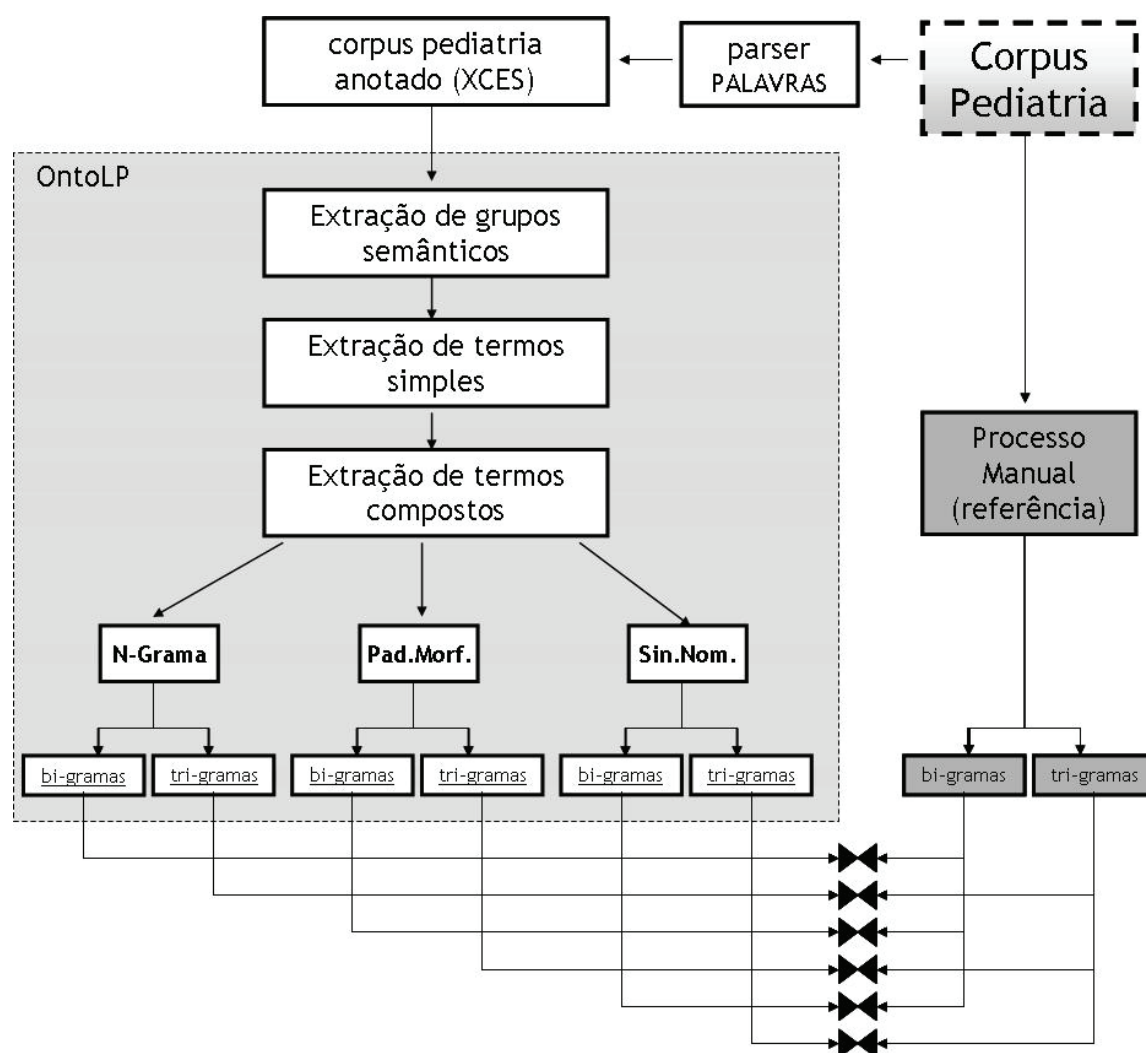


Figura 4 - Metodologia usada nos experimentos.

As etapas de extração de termos simples e compostos são reproduzidas para os dois grupos de termos identificados (com e sem exclusão de grupos semânticos) de maneira idêntica, tendo como saída do processo seis listas de bi-gramas e seis listas de tri-gramas. Cada uma delas foi comparada com as listas de referência de bi-gramas e tri-gramas. As listas de referências foram construídas por um processo fortemente apoiado em tarefas manuais, executado pelo Grupo TEXTQUIM/TERMISUL da Universidade Federal do Rio Grande do Sul (TEXTQUIM/UFRGS, <<http://www.ufrgs.br/textquim>>). O trabalho de extração de termos presentes no corpus de Pediatria visou à elaboração de um glossário para apoio aos estudantes de tradução. Esse material gerado também abastece os itens de um Catálogo de Expressões Recorrentes em Pediatria. O glossário e catálogo, desenhados como recursos *on-line* para educação à distância, visam auxiliar a formação de tradutores e de revisores de textos de Pediatria.

Na geração das listas de referência, foram inicialmente considerados n-gramas com mais de 5 ocorrências no corpus, extraídos automaticamente. A partir desta lista de 36.741 n-gramas, partiu-se para um processo de filtragem automático baseado em heurísticas. Por exemplo, termos que começavam ou terminavam por preposições foram transformados pela exclusão destas preposições; n-gramas contidos em n-gramas maiores foram excluídos. Dessa forma, um bi-grama que aparecia em um tri-grama foi descartado, pois, para fins de aprendizagem de tradução, termos com um maior número de palavras são preferíveis a termos menores. Por exemplo, o termo “aleitamento materno exclusivo” consta como um tri-grama, portanto “aleitamento materno” não consta na lista de bigramas. O processo resultou em uma lista com 3.645 n-gramas. Esta lista foi conferida manualmente por estudantes de tradução, resultando em uma lista final com 2.407 termos, sendo 1.293 bi-gramas, 775 tri-gramas e 339 termos de composição maior que 3 palavras.

A comparação das listas obtidas pela ferramenta OntoLP com as listas de referência foi feita através das seguintes métricas: precisão (P), abrangência (A) e *f-measure* (F). A precisão (P) indica a capacidade do método de identificar os termos corretos, considerando a lista de referência e é calculada pela fórmula (1).

$$P = (\text{Termos Referência} \cap \text{Termos Extraídos}) / \text{Termos Extraídos} \quad (1)$$

A abrangência (A) avalia a quantidade de termos corretos extraídos pelo método e é calculada através da fórmula (2).

$$A = (\text{Termos Referência} \cap \text{Termos Extraídos}) / \text{Termos Referência} \quad (2)$$

A *f-measure* (F) é a medida harmônica entre a precisão e abrangência, e é dada pela fórmula (3).

$$F = (2 * P * A) / (P + A) \quad (3)$$

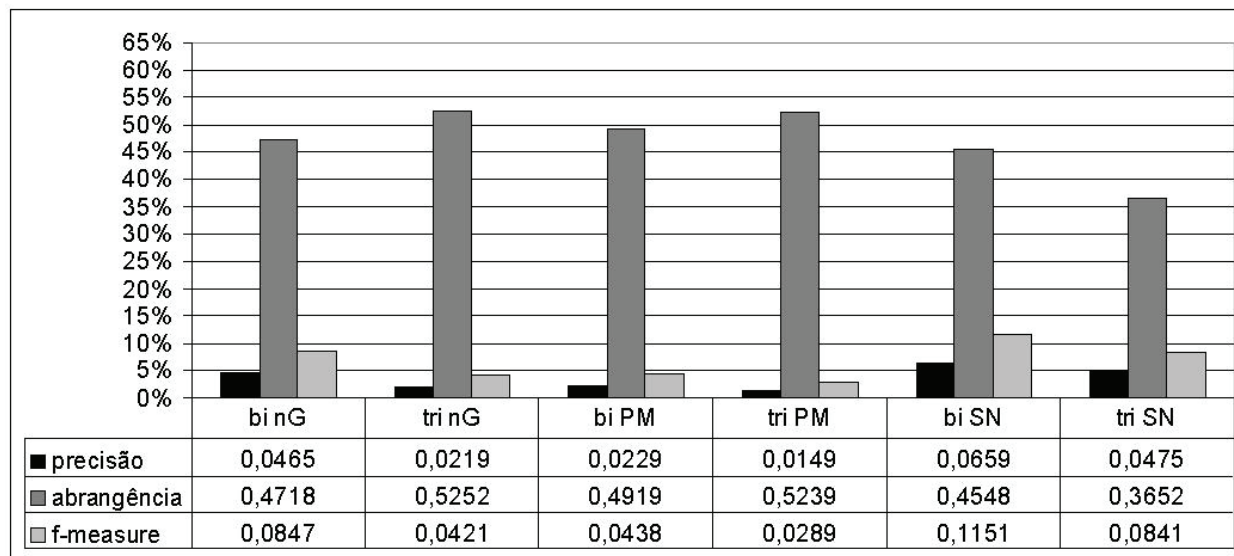
A Tabela 1 apresenta o número total de termos encontrados nos experimentos para cada uma das análises feitas. Adicionalmente, são mostrados quantos desses termos estão presentes na lista de referência que possui um total de 1.293 bi-gramas e 775 tri-gramas. O número de termos recuperado é bem superior ao número de termos da referência, uma vez que todos os termos extraídos do corpus são considerados, sem a utilização de um ponto de corte por frequência. Obviamente, este número diminui com a exclusão de grupos semânticos. Nesse caso, como a proporção de termos não relevantes excluídos é maior do que a de termos relevantes, observa-se um aumento na precisão.

**Tabela 1 - Termos extraídos e listas de referência**

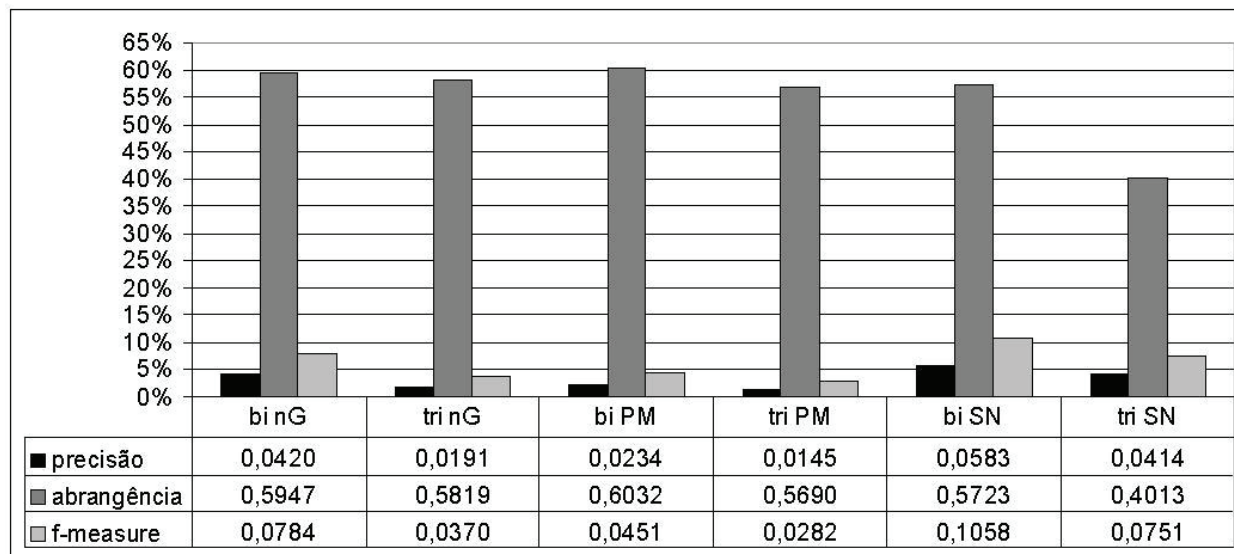
Com exclusão de grupos semânticos	n-gramas		Padrões morfosintáticos		Sintagma nominal	
	Bi-gramas <i>bi nG</i>	Tri-gramas <i>tri nG</i>	Bi-gramas <i>bi PM</i>	Tri-gramas <i>tri PM</i>	Bi-gramas <i>bi SN</i>	Tri-gramas <i>tri SN</i>
Total de Termos	13.115	18.554	27.763	27.322	8.926	5.959
Termos presentes na referência	610	407	636	406	588	283
Sem exclusão de grupos semânticos	n-gramas		Padrões morfosintáticos		Sintagma nominal	
	Bi-gramas <i>bi nG</i>	Tri-gramas <i>tri nG</i>	Bi-gramas <i>bi PM</i>	Tri-gramas <i>tri PM</i>	Bi-gramas <i>bi SN</i>	Tri-gramas <i>tri SN</i>
Total de Termos	18.325	23.588	33.276	30.497	12.691	7.509
Termos presentes na referência	769	451	780	441	740	311

As métricas finais da avaliação são apresentadas nas Figuras 5 e 6. O método que apresenta um melhor balanço entre precisão e abrangência (ambos para bi-gramas e tri-gramas) é a extração por sintagmas nominais, com exclusão de termos por grupos semânticos (f-measure de 11,51% e 8,41% para bi-gramas e tri-gramas, respectivamente). Estes são os métodos que empregam um maior nível de processamento linguístico, tanto em relação ao

processamento sintático como processamento semântico. Esta observação indica que pré-processamento linguístico do corpus tende a contribuir positivamente para a extração de termos compostos. Porém, ressaltamos que a informação semântica é empregada de forma semi-automática, os grupos são apresentados ao especialista do domínio que indica os grupos semânticos a serem desconsiderados no processo.



**Figura 5** - Gráfico de métricas com exclusão de grupos semânticos.



**Figura 6** - Gráfico de métricas sem exclusão de grupos semânticos.

A abrangência atingiu apenas 61% dos termos da referência, ou seja, um número significativo dos termos da referência não foi encontrado por nenhum dos métodos empregados. Este aspecto ainda deve ser investigado.

Nos apêndices (1-4) estão listados os primeiros termos das listas extraídas, ordenados por frequência. Foram destacados em negrito os termos constantes da lista de referência. Pode-se observar que alguns termos

aparentemente relevantes, encontrados pelos métodos, não constam na referência, por exemplo, “idade gestacional”, “fator de risco” e “aleitamento materno”. Em alguns casos, bigramas relevantes não foram constatados na referência, pois esta não inclui sub-termos. O bi-grama “aleitamento materno”, por exemplo, está ausente da lista de referência, pois este é um sub-termo do o tri-grama “aleitamento materno exclusivo”. Fatos como



esse sugerem a necessidade de uma revisão do método de avaliação, por refinamento da referência, uma vez que os propósitos do presente trabalho diferem dos de auxílio à tradução. Por outro lado, destacamos a importância deste recurso, pois a área de aprendizagem de ontologias, por ser muito recente, ainda enfrenta sérios problemas relacionados à carência de recursos para avaliação. Avaliações, apesar de difíceis, são cruciais para o desenvolvimento das técnicas. Avaliações nesta área, além de lidarem com a escassez de recursos, enfrentam problemas relativos à subjetividade. É comum que diferentes especialistas tenham julgamentos diferenciados em relação à relevância dos termos.

## Trabalhos relacionados

Em Cimiano (2006) aborda-se de forma bastante completa o problema de extração de ontologias de texto, um problema com muitas questões de pesquisa ainda em aberto.

Suchanek (2006) discute de forma geral o emprego de análise lingüística na extração de informação em bases textuais. Em particular, a extração de termos compostos é um tema bastante investigado (Ramisch et al. 2008) é um exemplo de trabalho recente nesta linha. O propósito de investigação de termos compostos, no entanto, é variável, nem sempre são considerados como parte do levantamento de conceitos em um domínio.

Trabalhos anteriores relativos ao português são poucos; (Baségio 2006), por exemplo, apresenta uma primeira abordagem para o problema de extração de ontologias de texto baseada em corpus de domínio. A abordagem de Baségio foi avaliada para o domínio do Turismo através do julgamento de especialistas, sem uma lista de referência. Nessa forma de avaliação, entretanto, não é possível calcular métricas como abrangência e *f-measure*.

Ribeiro e Vieira (2008) avaliam os métodos do Plug-in OntoLP num corpus de ecologia em relação aos primeiros 1000 termos extraídos por cada método. O impacto da extração de grupos semânticos é avaliado num conjunto de 150 termos. Neste trabalho apresentamos uma avaliação dos métodos a partir de uma lista de referência mais extensa, em outro domínio, a Pediatria, e as avaliações consideram o conjunto total de termos extraídos.

O fato de a área de aprendizagem de ontologias ser muito recente, torna-se difícil apresentar uma análise comparativa com outros trabalhos, uma vez que ainda não há testes padrões disponíveis.

## Conclusão

Apresentamos aqui uma avaliação inicial do uso de técnicas de Processamento de Linguagem Natural aplicadas ao problema de construção de ontologias. Os experimentos realizados estão relacionados com a primeira etapa do complexo processo de construção de ontologias, qual seja, a etapa de identificação dos termos candidatos a conceitos. Os experimentos realizados consideram um corpus e uma lista de referência de construções relevantes para a área de ensino de tradução de textos de Medicina/Pediatria. Esta é

uma área, entre outras, que pode se beneficiar do desenvolvimento de técnicas de processamento textual e estruturação de informação, uma vez que há muito conhecimento específico adquirido, que está registrado em texto.

Apesar de preliminares, os resultados podem ser usados para observação do comportamento dos diferentes métodos de extração empregados. Os métodos informados linguisticamente mostraram vantagens em relação aos métodos menos informados.

Como trabalho futuro, pretendemos refinar tanto a lista de termos utilizada como referência como também o processo de extração de termos. Como o engenheiro de ontologias recebe uma lista de termos ordenada por frequência, uma avaliação importante a ser realizada é a análise da precisão dos termos mais frequentes; será importante avaliar também a evolução do balanço entre precisão e abrangência conforme o ponto de corte. Este trabalho já está em andamento. É importante identificar um balanço entre precisão/abrangência que seja útil, isto é, que possa contribuir positivamente ao engenheiro de ontologias.

Além disso, planejamos avançar nas outras etapas de construção de ontologias. Para isso, em um primeiro momento, trabalharemos com o agrupamento semântico das expressões, identificando hierarquias e similaridades entre os termos.

As técnicas avaliadas neste artigo são incorporadas ao editor de ontologias Protégé, por meio de um plug-in. O plug-in, bem como outros recursos para desenvolvimento de pesquisa em ontologias, está disponível em <<http://www.inf.pucrs.br/~ontolp>>.

## Agradecimentos

Agradecemos aos financiamentos da Capes, CNPq e SEAD/UFRGS concedidos aos autores deste trabalho.

## Referências bibliográficas

- Bacelar S, Galvão CC, Alves E, Tubino P. Expressões médicas: falhas e acertos. Rev Bras Cirurgia Cardiovascular, São Paulo. Jul./Set., 2003; 18(3).
- Baségio T. Uma Abordagem Semi-Automática para Identificação de Estruturas Ontológicas a partir de Textos na Língua Portuguesa do Brasil. 2006. Dissertação (Mestrado em Ciência da Computação), Pontifícia Universidade Católica do Rio Grande do Sul - PUCRS, Porto Alegre.
- Bick E. The parsing System "Palavras": Automatic grammatical analysis of portuguese in a constraint grammar framework. PhD thesis, Arhus University. 2000.
- Blank D, Rosa LO, Gurgel RQ, Goldani MZ. Produção brasileira de conhecimento no campo da saúde da criança e do adolescente. J Pediatria, Rio de Janeiro. 2006; 82(2): 97-102.
- Brewster C, Ciravegna F, Wilks Y. Background and foreground knowledge in dynamic ontology construction. In: SIGI, Proceedings of the Semantic Web Workshop, 2003, Toronto. Proceedings. Toronto: August, 2003.

Buitelaar P, Cimiano P, Magnini B. Ontology learning from text: An overview. In: Buitelaar P, Cimiano P, Magnini B. (editors). Ontology learning from text: methods, evaluation and applications, v. 123 of Frontiers in Artificial Intelligence and Applications. IOS Press, 2005.

Cimiano P. Ontology learning and population from text: Algorithms, evaluation and applications. Heidelberg: Springer-Verlag, 2006.

Frantizi KT, Ananiadou S, Ichi Tsujii J. The c-value/nc-value method of automatic recognition for multi-word terms. In: ECDL'98: Proceedings of the second european conference on research and advanced technology for digital libraries, 1998, London. Proceedings. Heidelberg: Springer-Verlag, 1998, p. 585-604.

Gennari J et al. The evolution of protégé: an environment for knowledge-based systems development. Technical Report SMI-2002-0943. 2002.

Gomez-Perez A, Corcho O, Fernandez-Lopez M. Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. Heidelberg: Springer-Verlag. 2004.

Ide N, Bonhomme P, Romary L. Xces: An xml-based encoding standart for linguistic corpora. In: Proceedings of the second international language resources and evaluation conference, 2000. Proceedings. Paris: European Language Resources Association, 2000.

Manning CD, Schutze H. Foundations of statistical natural language processing. Cambridge, Massachusetts: The MIT Press, 1999.


Mcguinness DL, Van Harmelen F. OWL web ontology language overview. World Wide Web Consortium (W3C) recommendation. <<http://www.w3.org/TR/owl-features>>. Acesso em: 1 Feb. 2004.

PROTÉGÉ <<http://protege.stanford.edu>> Acesso em: 25 ago. 2008.

Ramisch C, Schreiner P, Idiart M, Villavicencio A. An evaluation of methods for the extraction of multiword expressions. In: LREC 2008 MWE workshop: towards a shared task on multiword expressions, Marrakesh, 2008. Proceedings. Paris: European Language Resources Association, 2008.

Suchanek FM, Ifrim G, Andweikum G. Leila: Learning to extract information by linguistic analysis. In: Proceedings of the 2nd workshop on ontology learning and population: bridging the gap between text and knowledge, Sydney, Australia, 2006. Proceedings. Association for Computational Linguistics, 2006.

Ribeiro LC. OntoLP: Construção semi-automática de ontologias a partir de textos da língua portuguesa. Dissertação (Mestrado em Computação Aplicada), Universidade do Vale do Rio dos Sinos - UNISINOS, São Leopoldo. 2008.

Ribeiro LC, Vieira R. ontolp: engenharia de ontologias em língua portuguesa. In: Anais do xxviii congresso da SBC - SEMISH - Seminário integrado de software e hardware, Belém do Pará, 2008. Anais. Porto Alegre: Sociedade Brasileira de Computação, 2008. 

## Sobre os autores

### *Lucelene Lopes*

É doutoranda do Curso de Ciências da Computação da PUCRS desde 2008. Possui Mestrado em Tecnologia em Saúde pela PUCPR (2007). Graduada em Ciências com Habilitação Plena em Matemática pela UNIVALE (2000). Atua principalmente em Aprendizagem de Máquina (Inteligência Artificial) desde o mestrado e mais recentemente, com o início do doutorado, tem atuado em extração de termos dentro da área de Processamento de Linguagem Natural.

### *Renata Vieira*

Possui título de PhD em Informática pela University of Edinburgh (1998). É professora da PUCRS onde atua em pesquisa e ensino na graduação e pós-graduação na área de inteligência computacional, com ênfase em processamento de linguagem natural, representação do conhecimento, ontologias, agentes e web semântica. Possui experiência em coordenação de projetos, é membro de comitês científicos das principais conferências internacionais da área de lingüística computacional e agentes inteligentes. Participa ativamente no desenvolvimento da área de processamento de linguagem natural no país.

## Apêndice 1

Este Apêndice apresenta os 70 primeiros termos identificados em cada método (bi-gramas) **com** exclusão de termos. Os termos extraídos que constam na referência estão destacados.

### Bigramas NG

aleitamento materno, idade gestacional, ventilação mecânica, faixa etária, terapia intensiva, hipertensão arterial, baixo peso, grupo controle, **diferença estatística**, perímetro cefálico, grande número, período neonatal, alto risco, massa óssea, **exame físico**, vitamina d, **nível sérico**, grande risco, amamentação exclusiva, carga viral, **infecção urinária**, **diferença significativa**, grande frequência, significância estatística, **baixa estatura**, cicatriz renal, insuficiência adrenal, **otite média**, choque séptico, saúde pública, **diagnóstico diferencial**, insuficiência respiratória, regressão logística, nível plasmático, **prática clínica**, evolução neurológico, solução salina, **quadro clínico**, ventilação pulmonar, via oral, fibrose cística, **idade inferior**, relaxamento muscular, primeiro mês, grande incidência, grande prevalência, **tubo endotraqueal**, **frequência respiratória**, anemia falciforme, dieta isento, escolaridade materna, **avaliação clínica**, obesidade infantil, desconforto respiratório, dor abdominal, **escore z**, disfunção miccional, perda auditiva, hipertensão pulmonar, grau I, **escore clínico**, **evolução clínica**, **deposição pulmonar**, pressão intracraniano, alta hospitalar, perímetro braquial, fase aguda, tempo médio, sexto mês, longo prazo

### Bigramas PM

aleitamento materno, faixa etária, idade gestacional, ventilação mecânica, criança pequena, terapia intensiva, hipertensão arterial, **diferença significativa**, perímetro cefálico, cicatriz renal, **otite média**, **efeito colateral**, período neonatal, criança grande, paciente pediátrico, **nível sérico**, **exame físico**, **infecção urinária**, **manifestação clínica**, massa óssea, amamentação exclusiva, **efeito adverso**, carga viral, significância estatística, **quadro clínico**, choque séptico, solução salina, ensaio clínico, saúde pública, insuficiência respiratória, atividade física, **doença crônica**, regressão logística, nível plasmático, **prática clínica**, evolução neurológica, evento adverso, **avaliação clínica**, via oral, **frequência respiratória**, **escore clínico**, ventilação pulmonar, **idade inferior**, grau I, dor abdominal, relaxamento muscular, fibrose cística, **tubo endotraqueal**, **critério diagnóstico**, **cardiopatia congênito**, **infecção respiratória**, exame complementar, anemia falciforme, pressão intracraniano, **evidência científica**, escolaridade materna, perda auditiva, desconforto respiratório, hipertensão pulmonar, exame laboratorial, obesidade infantil, **evolução clínica**, tempo médio, infecção congênito, disfunção miccional, **fator prognóstico**, **volume corrente**, paciente portador, **alergia alimentar**, doença pulmonar

## Bigramas SN

aleitamento materno, idade gestacional, ventilação mecânica, período neonatal, hipertensão arterial, perímetro cefálico, cicatriz renal, terapia intensiva, alto risco, criança pequena, baixo peso, **exame físico**, massa óssea, **diferença significativa**, significância estatística, amamentação exclusiva, paciente pediátrico, saúde pública, **infecção urinária**, choque séptico, criança grande, **baixa estatura**, último ano, solução salina, via oral, **tubo endotraqueal**, fibrose cística, anemia falciforme, grande frequência, **cardiopatia congênito**, obesidade infantil, **prática clínica**, **efeito colateral**, **doença crônica**, pressão intracraniana, escolaridade materna, infecção congênito, alta hospitalar, **alergia alimentar**, **manifestação clínica**, exame complementar, **otite média**, **efeito adverso**, **frequência cardíaca**, evolução neurológica, doença pulmonar, disfunção miccional, **diagnóstico diferencial**, dois paciente, **frequência respiratória**, criança estudada, dor abdominal, **aleitamento exclusivo**, regressão logística, **quadro clínico**, hipertensão pulmonar, faixa etária, modo geral, bom resultado, carga viral, orelha média, idade escolar, longo prazo, alto frequência, evento adverso, ventilação pulmonar, curta duração, tamanho amostral, presente trabalho, quatro paciente

## Apêndice 2

Este Apêndice apresenta os 70 primeiros termos identificados em cada método (tri-gramas) **com** exclusão de termos. Os termos extraídos que constam na referência estão destacados.

### Trigrama NG

ano de idade, fator de risco, mês de vida, ano de vida, mês de idade, peso de nascimento, dia de vida, **aleitamento materno exclusivo**, unidade de terapia, **intervalo de confiança**, nível de significância, qualidade de vida, país em desenvolvimento, serviço de saúde, semana de vida, diferença estatística significativa, hora de vida, critério de inclusão, velocidade de crescimento, **curva de crescimento**, coleta de dado, diferença estatística significante, problema de saúde, **tipo de parto**, média de idade, **taxa de mortalidade**, termo de consentimento, tempo de internação, **grupo de risco**, consumo de medicamento, produção de leite, confiança de 95, **dieta de exclusão**, trabalho de parto, número de paciente, déficit de atenção, **saturação de oxigênio**, estilo de vida, prevalência de asma, necessidade de ventilação, **terapia intensivo neonatal**, uso de antibiótico, densidade mineral óssea, criança com idade, **risco de infecção**, paciente com doença, **faixa etária pediátrico**, livre de doença, volume de leite, plasmático de vitamina, **modelo de regressão**, transtorno de ansiedade, uso em criança, índice de massa, uso de droga, **ingestão de cálcio**, **radiografia de tórax**, período de tempo, ventilação não invasivo, **síndrome de Down**, equipe de saúde, **ponto de corte**, **solução salina hipertônica**, uso de

oxigênio, uso de medicamento, análise de variância, uso de medicação, **ventilação pulmonar mecânica**, nível de escolaridade, selo de água

## Trigrama PM

ano de idade, fator de risco, mês de vida, ano de vida, mês de idade, peso de nascimento, profissional de saúde, dia de vida, grupo de paciente, unidade de terapia, **intervalo de confiança**, qualidade de vida, nível de significância, país em desenvolvimento, grupo de criança, serviço de saúde, semana de vida, hora de vida, critério de inclusão, maioria do paciente, velocidade de crescimento, problema de saúde, coleta de dado, **curva de crescimento**, tamanho da amostra, **tipo de parto**, número de paciente, média de idade, **taxa de mortalidade**, termo de consentimento, maioria do caso, tempo de internação, **grupo de risco**, **ponto de corte**, paciente do grupo, consumo de medicamento, produção de leite, início do sintoma, **dieta de exclusão**, trabalho de parto, paciente com doença, vida da criança, déficit de atenção, uso de antibiótico, **saturação de oxigênio**, estilo de vida, criança com idade, aumento da pressão, necessidade de ventilação, duração do aleitamento, aplicação da vacina, volume de leite, prevalência de asma, saúde da criança, **risco de infecção**, maioria da criança, uso em criança, **modelo de regressão**, **transtorno de ansiedade**, idade da criança, criança do sexo, **ingestão de cálcio**, índice de massa, **radiografia de tórax**, período de tempo, uso de droga, **tempo de queixa**, uso de oxigênio, uso de medicamento, início na infância

## Trigrama SN

fator de risco, **aleitamento materno exclusivo**, profissional de saúde, peso de nascimento, critério de inclusão, coleta de dado, serviço de saúde, **tipo de parto**, tempo de internação, país em desenvolvimento, **faixa etária pediátrico**, trabalho de parto, **terapia intensivo neonatal**, **ventilação pulmonar mecânica**, uso de antibiótico, produção de leite, **intervalo de confiança**, **saturação de oxigênio**, velocidade de crescimento, **otite média agudo**, terapia intensivo pediátrico, qualidade de vida, uso de oxigênio, sala de parto, **dieta de exclusão**, nível de significância, ventilação não invasiva, estilo de vida, equipe de saúde, consumo de medicamento, análise de variância, **radiografia de tórax**, acidente de transporte, **grupo de risco**, **farelo de trigo**, selo de água, termo de consentimento, **relaxamento muscular inadequado**, uso de medicamento, uso de medicação, **doença de base**, densidade mineral óssea, **suplemento de cálcio**, diferença estatística significativa, **taxa de mortalidade**, critério de exclusão, **vacina contra influenza**, **centro de referência**, **ansiedade de separação**, nível de linfócito, deficiência de vitamina, criança mais velha, **tempo de queixa**, **curva de crescimento**, esquizofrenia com início, escape de ar, situação de estresse, **síndrome de abstinência**, infecção respiratória aguda, **amostra de sangue**, **tubo de ventilação**, **hemorragia digestiva alta**, hipótese de nulidade, centro de saúde, **controle sem hepatopatia**, **risco de infecção**, **uso de chupeta**,

**vacinação contra influenza**, **aspiração de mecônio**, local de trabalho

## Apêndice 3

Este Apêndice apresenta os 70 primeiros termos identificados em cada método (bi-gramas) **sem** exclusão de termos. Os termos extraídos que constam na referência estão destacados.

## Bigramas NG

aleitamento materno, idade gestacional, leite materno, faixa etária, ventilação mecânica, presente estudo, terapia intensiva, hipertensão arterial, leite humano, baixo peso, primeiro ano, estado nutricional, grupo controle, **diferença estatística**, perímetro cefálico, um ano, massa óssea, período neonatal, grande número, alto risco, 1 ano, **escore z**, **exame físico**, vitamina d, **nível sérico**, grande risco, análise estatística, carga viral, amamentação exclusiva, **infecção urinária**, **diferença significativa**, grande parte, dieta isenta, grande frequência, significância estatística, cicatriz renal, insuficiência adrenal, **baixa estatura**, **otite média**, história familiar, vez grande, choque séptico, **quadro clínico**, lesão pulmonar, saúde pública, regressão logística, insuficiência respiratória, **diagnóstico diferencial**, nível plasmático, pressão intracraniano, **prática clínica**, evolução neurológica, teste t, população estudado, solução salina, **diagnóstico precoce**, ventilação pulmonar, via oral, medicamento não, primeiro mês, **idade inferior**, fibrose cística, corticosteroide antenatal, **avaliação clínica**, relaxamento muscular, grande incidência, grande prevalência, anemia falciforme, resposta inflamatória, **tubo endotraqueal**

## Bigramas PM

aleitamento materno, faixa etária, idade gestacional, leite materno, ventilação mecânica, criança pequena, terapia intensiva, hipertensão arterial, leite humano, estado nutricional, **diferença significativa**, perímetro cefálico, cicatriz renal, massa óssea, **efeito colateral**, alimento complementar, **otite média**, período neonatal, análise estatística, **nível sérico**, paciente pediátrico, criança grande, **exame físico**, **manifestação clínica**, **infecção urinária**, **efeito adverso**, amamentação exclusiva, carga viral, dieta isento, **quadro clínico**, significância estatístico, vez grande, ensaio clínico, história familiar, solução salina, choque séptico, lesão pulmonar, população estudado, pressão intracraniano, atividade física, saúde pública, insuficiência respiratória, **prática clínica**, **doença crônica**, regressão logística, nível plasmático, maioria do paciente, lesão cerebral, evolução neurológico, evento adverso, **avaliação clínica**, **frequência respiratória**, ventilação pulmonar, **diagnóstico precoce**, **escore clínico**, via oral, **idade inferior**, relaxamento muscular, fibrose cística, dor abdominal, grau I, **tubo endotraqueal**, **critério diagnóstico**, cardiopatia congênito, resposta inflamatório, anemia falciforme, **infecção respiratório**, exame complementar



## Bigramas SN

aleitamento materno, presente estudo, idade gestacional, leite materno, ventilação mecânica, dois grupo, 6 mês, leite humano, 2 ano, período neonatal, hipertensão arterial, cicatriz renal, perímetro cefálico, terapia intensivo, alimento complementar, alto risco, tabela 2, massa óssea, 5 ano, estado nutricional, um ano, criança pequena, baixo peso, **exame físico**, 1 ano, população estudado, **diferença significativa**, dois ano, cinco ano, análise estatística, significância estatística, tabela 3, amamentação exclusiva, paciente pediátrico, último década, tabela 1, 24 hora, saúde pública, criança grande, pressão intracraniano, choque séptico, **infecção urinária**, **baixo estatura**, último ano, seis mês, **tubo endotraqueal**, solução salina, via oral, fibrose cística, anemia falciforme, grande frequência, corticosteróide antenatal, 12 mês, amostra estudada, faixa etário, **cardiopatia congênito**, 10 ano, **efeito colateral**, obesidade infantil, quatro mês, país desenvolvido, mãe adolescente, 6 ano, **prática clínica**, **doença crônica**, 30 minuto, escolaridade materno, infecção congênito, **efeito adverso**, alta hospitalar

## Apêndice 4

Este Apêndice apresenta os 70 primeiros termos identificados em cada método (tri-gramas) **sem** exclusão de termos. Os termos extraídos que constam na referência estão destacados.

## Trigrama NG

ano de idade, fator de risco, mês de vida, ano de vida, mês de idade, **leite de vaca**, peso de nascimento, profissional de saúde, dia de vida, **aleitamento materno exclusivo**, unidade de terapia, grupo de paciente, **intervalo de confiança**, nível de significância, qualidade de vida, país em desenvolvimento, semana de vida, serviço de saúde, diferença estatística significativa, hora de vida, **ponto de corte**, critério de inclusão, velocidade de crescimento, polissacarídeo de soja, coleta de dado, diferença estatística significante, **curva de crescimento**, problema de saúde, isento de leite, t de Student, **tipo de parto**, grupo de criança, **taxa de mortalidade**, média de idade, termo de consentimento, tempo de internação, **grupo de risco**, confiança de 95, consumo de medicamento, produção de leite, **dieta de exclusão**, trabalho de parto, número de paciente, casca de banana, déficit de atenção, **saturação de oxigênio**, estilo de vida, necessidade de ventilação, densidade mineral óssea, criança com idade, uso de antibiótico, volume de leite, **terapia intensivo neonatal**, **risco de infecção**, prevalência de asma, livre de doença, paciente com doença, **faixa etária pediátrico**, uso em criança, plasmático de vitamina, **transtorno de ansiedade**, **modelo de regressão**, fórmula de soja, índice de massa, **radiografia de tórax**, uso de droga, **ingestão de cálcio**, período de tempo, **síndrome de Down**, tempo de queixa

## Trigrama PM

ano de idade, fator de risco, mês de vida, ano de vida, mês de idade, **leite de vaca**, peso de nascimento, profissional de saúde, dia de vida, grupo de paciente, unidade de terapia, **intervalo de confiança**, nível de significância, qualidade de vida, grupo de criança, país em desenvolvimento, serviço de saúde, semana de vida, **ponto de corte**, hora de vida, critério de inclusão, maioria do paciente, velocidade de crescimento, problema de saúde, **curva de crescimento**, coleta de dado, **tipo de parto**, tamanho da amostra, número de paciente, maioria da vez, **taxa de mortalidade**, média de idade, maioria do caso, termo de consentimento, **grupo de risco**, tempo de internação, filho de mãe, consumo de medicamento, paciente do grupo, produção de leite, paciente com doença, trabalho de parto, início do sintoma, **dieta de exclusão**, vida da criança, casca de banana, déficit de atenção, **saturação de oxigênio**, uso de antibiótico, estilo de vida, maioria do estudo, criança com idade, prevalência de asma, duração do aleitamento, volume de leite, aplicação da vacina, necessidade de ventilação, aumento da pressão, saúde da criança, **risco de infecção**, maioria da criança, uso em criança, **transtorno de ansiedade**, **modelo de regressão**, idade da criança, período de tempo, índice de massa, **ingestão de cálcio**, criança do sexo, **radiografia de tórax**

## Trigrama SN

fator de risco, **aleitamento materno exclusivo**, profissional de saúde, **leite de vaca**, peso de nascimento, critério de inclusão, coleta de dado, serviço de saúde, **tipo de parto**, país em desenvolvimento, tempo de internação, polissacarídeo de soja, **faixa etária pediátrico**, trabalho de parto, **ventilação pulmonar mecânica**, **terapia intensivo neonatal**, casca de banana, uso de antibiótico, produção de leite, **saturação de oxigênio**, **intervalo de confiança**, velocidade de crescimento, qualidade de vida, terapia intensiva pediátrica, **ponto de corte**, **otite médio agudo**, ventilação não invasivo, nível de significância, consumo de medicamento, **dieta de exclusão**, sala de parto, uso de oxigênio, **lesão pulmonar agudo**, estilo de vida, equipe de saúde, acidente de transporte, **radiografia de tórax**, **grupo de risco**, farelo de trigo, análise de variância, termo de consentimento, **fórmula de soja**, critério de exclusão, densidade mineral óssea, **suplemento de cálcio**, **taxa de mortalidade**, uso de medicação, diferença estatística significativa, **doença de base**, selo de água, **relaxamento muscular inadequado**, uso de medicamento, **centro de referência**, **vacina contra influenza**, década de 70, **ansiedade de separação**, deficiência de vitamina, criança mais velha, estudo de corte, escape de ar, nível de linfócito, esquizofrenia com início, **curva de crescimento**, **tempo de queixa**, **risco de infecção**, **aspiração de mecônio**, **uso de chupeta**, **amostra de sangue**, primeiro 24 hora, **resposta inflamatório sistêmico**



**RECIIS**

Revista Eletrônica de Comunicação  
Informação & Inovação em Saúde

[www.reciis.cict.fiocruz.br]

ISSN 1981-6278

*Resenhas*

## **Semantic Web services, processes and applications**

*Jorge Cardoso and Amit P. Sheth (Eds.)*

## **Semantic Web and semantic Web services**

*Liyang Yu*

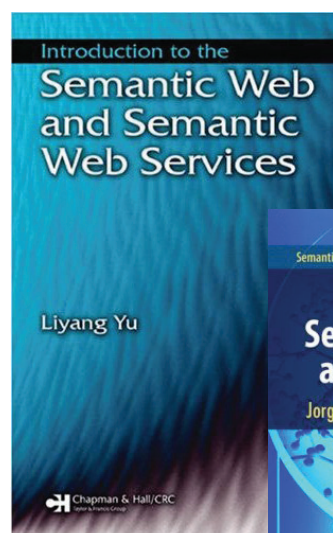
DOI: 10.3395/reciis.v3i1.246pt

*Laura Cristina Simões Viana*

Coordenação de Recursos Humanos em Pesquisa da Vice-Presidência de Pesquisa e Desenvolvimento Tecnológico da Fundação Oswaldo Cruz, Rio de Janeiro, Brasil  
laura@fiocruz.br

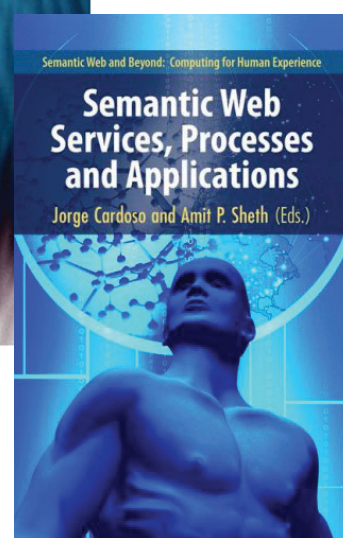
Tim Berners-Lee criou, em 1989, as ferramentas básicas necessárias ao funcionamento da Web, tal como a conhecemos atualmente: a linguagem HyperText Markup Language – HTML ou linguagem de marcação de hipertexto; o protocolo HyperText Transfer Protocol – HTTP ou protocolo de transferência de hipertexto, e o sistema de localização de objetos na Web, Universal Resource Locator – URL ou localizador universal de recursos. Dose anos após<sup>1</sup>, com a Web funcionando plenamente, Berners-Lee publicou um artigo original com uma proposta desafiadora de “Web semântica” (Berners-Lee, 2001).

De acordo com essa proposta, a Web semântica seria em uma ampliação da web atual, na qual a informação passaria a ter um significado definido, viabilizando o trabalho cooperativo entre homens e máquinas. Assim, a Web atual, que é formada, basicamente, por documentos apresentados por computadores e lidos pelo homem, passaria a incluir dados e informações que seriam automaticamente manipulados por agentes inteligentes e utilitários. Naquela época, Berners-Lee (2001) argumentou que o progresso da Web semântica necessitaria do desenvolvimento de uma linguagem que expressasse dados e regras de raciocínio sobre os dados, e que também permitisse a exportação para Web das regras de qualquer sistema de representação do conhecimento. Até então, encontravam-se desenvolvidas duas tecnologias funda-



*New York,  
Springer, 2006*

*ISBN  
978-0387302393*



*USA, Chapman & Hall/  
CRC, 2007*

*ISBN  
978-1584889335*

mentais para a realização da Web semântica: a linguagem de marcação eXtensible Markup Language – XML e a família de especificações Resource Description Framework – RDF, esta última para descrição ou modelagem de informação implementada em recursos da web.

Em 2006, embora os padrões Web que expressam significado compartilhado estivessem em franco desenvolvimento, Shadbolt (2006) publicou outro artigo intitulado “The semantic Web revisited”, no qual afirmava que a Web semântica permaneceria, basicamente, no plano das idéias até que esses padrões fossem bem estabelecidos e acordados. Shadbolt (2006) ressaltou que, a exemplo do protocolo HTTP, cujo uso pioneiro pela comunidade da área da física abriu caminho para o sucesso da Web, o uso crescente de ontologias pela comunidade científica eletrônica (“e-science”) também poderia levar a Web semântica ao sucesso tal como hoje conhecemos a Web. Segundo o autor, além de fatores sociais e de decisões de projeto, parte do sucesso da Web semântica reside na sequência de especificações (Universal Resource Identifier - URI, HTTP, RDF, ontologias, etc.) e de registros (esquema URI, conteúdos de Internet do tipo Multipurpose Internet Mail Extensions – MIME ou extensões multifunção para mensagens de Internet), os quais fornecem meios para que construções como uma ontologia derivem significado de um identificador URI.

O livro de Jorge Cardoso & Amit Sheth, publicado em 2006, e o livro de Liyang Yu, publicado em 2007, são textos práticos, que abordam a Web semântica com toda a riqueza dos padrões, linguagens, aplicações e processos que o tema provoca. Amit Sheth fez seu Doutorado em Ciências da Computação e Informação na Ohio State University e foi professor da Univesidade da Geórgia entre 1994 e 2006, tendo publicado mais de 250 artigos revisados pelos pares e oito livros; atualmente, é professor de Engenharia e Ciência da Computação e diretor do centro Kno.e.sis, ambos na Wrigt State University. Jorge Cardoso fez seu Doutorado em Ciências da Computação, com Amit Sheth, na Univesidade da Geórgia e trabalha atualmente no Departamento de Pesquisa da empresa alemã SAP, além de manter seu vínculo de professor com a Universidade da Madeira, em Portugal. Cardoso tem mais de 80 artigos publicados em periódicos revisados pelos pares nas áreas de sistemas de gestão de workflow, Web semântica e editou três livros sobre Web semântica e serviços Web. Liyang Yu fez seu Doutorado na Ohio State University e, atualmente, trabalha para a Delta Technologies, nos Estados Unidos.

O livro editado por Jorge Cardoso e Amit Sheth é uma coletânea de contribuições de pesquisadores de universidades e empresas sobre semântica, organizada em três partes: a primeira trata dos serviços Web semânticos, a segunda apresenta processos da Web semântica e a última parte do livro mostra aplicações da Web semântica. O conjunto dos capítulos aborda padrões que combinam semântica e tecnologia de serviços Web segundo três níveis de exigências de conhecimentos: iniciante, intermediário e avançado. Organizado em quatro partes, o livro de Liyang Yu apresenta os assuntos de

modo semelhante ao livro de Cardoso e Sheth. A primeira parte do texto de Liyang Yu é sobre o entendimento, as justificativas e o valor potencial da Web semântica; a segunda detalha aspectos técnicos e tecnologias da Web semântica, a terceira parte do livro apresenta exemplos reais e aplicações práticas de Web semântica e, ao final, mostra a evolução de serviços Web para serviços Web semânticos. A exemplo da edição de Cardoso e Sheth, o livro de Liyang Yu destina-se a desenvolvedores, estudantes de graduação e pós-graduação, e pesquisadores, atendendo às comunidades interessadas nas aplicações de tecnologia de serviços Web semânticos.

Ambos os livros almejam o posto de livro texto sobre Web semântica. No caso do texto editado por Cardoso e Sheth, ao final de cada capítulo constam questões para discussões e sugestões de leitura adicional. O livro de Liyang Yu, embora tenha um caráter introdutório ao mundo da Web semântica, a abrangência da abordagem exige do leitor conhecimentos básicos sobre a linguagem HTML e sólidos conhecimentos das linguagens XML, Java e de servidores Web. Ao longo dos capítulos do seu livro, cuja leitura não precisa ser linear, Liyang Yu apresenta exemplos de integração de sistemas, aplicativos e serviços Web; de máquinas de buscas e de aplicações de mineração de dados, que são as principais utilizações da Web, atualmente.

A primeira parte do texto de Cardoso e Sheth, que aborda os serviços Web semânticos, se inicia com uma discussão sobre a Web semântica e suas aplicações, e mostra que a heterogeneidade na Web é conhecida de longa data da comunidade de sistemas distribuídos de bancos de dados. Segundo Cardoso e Sheth (2006), a heterogeneidade surge de discordâncias sobre o significado, a interpretação e o uso pretendido dos dados; a heterogeneidade semântica leva em conta o conteúdo de um item da informação e o significado pretendido. A solução proposta pelos autores para a heterogeneidade semântica é basear-se nos fundamentos tecnológicos da Web semântica que, em resumo, marca os documentos com metadados semânticos, os quais são compreendidos por máquinas, para posterior extração, por exemplo, através de ontologias. Os capítulos dois, três e quatro da primeira parte do livro de Cardoso e Sheth apresentam e discutem os caminhos da evolução de serviços Web para serviços Web semânticos: anotação semântica em serviços Web, ampliação semântica dos padrões de serviços Web, e esquemas para descoberta e publicação de serviços Web. O quinto capítulo do livro propõe uma metodologia para expandir a descrição semântica de um serviço com propriedades temporais, que permitam a inferência sobre o comportamento ao longo do tempo.

A primeira parte do livro de Liyang Yu introduz o leitor no mundo da Web semântica. O primeiro capítulo argumenta que a integração de sistemas, de aplicativos e de serviços Web, assim como as máquinas de buscas e a mineração de dados se ressentem de um agente de software que seja capaz de realizar o processamento em larga escala, na Web. Do mesmo modo que Cardoso e Sheth, Liyang Yu entende que é necessário introduzir se-



mântica na Web, destacando a importância de metadados para as máquinas. O segundo capítulo do texto de Liyang Yu mostra o funcionamento das máquinas de busca e as mudanças necessárias para adaptá-las ao mundo da Web semântica, além de destacar os benefícios desta evolução das máquinas de busca. A segunda parte do livro de Liyang Yu está estruturada em quatro capítulos e trata dos fundamentos técnicos da Web semântica: RDF, Resource Description Framework *Schema* – RDF *Schema*, Web Ontology Language - OWL, taxonomia e ontologia, e das ferramentas acessórias para a Web semântica, como a validação de ontologias OWL.

A segunda parte do livro de Cardoso e Sheth aborda os processos da Web semântica. Assim, o capítulo seis apresenta as idéias e os princípios que norteiam a coreografia, ou seja a modelagem do comportamento visível das interações entre serviços. O capítulo sete mostra como introduzir semântica em padrões de serviços Web, utilizando WSDL-S, que é uma extensão da especificação da linguagem Web Services Description Language – WSDL. Os capítulos oito e nove cobrem tópicos avançados e discutem a composição de serviços Web com base em propriedades não funcionais, utilizando técnicas de otimização multiobjetivo, além de apresentar um quadro genérico de harmonização e mapeamento de processos Web semânticos.

A terceira parte de ambos os livros apresenta aplicações e exemplos da Web semântica no mundo real. Cardoso e Sheth ilustram aplicações de Web semântica nas áreas do turismo, governo, bioinformática e serviços Web. O capítulo 10 do texto de Cardoso e Sheth ilustra e descreve a construção de uma ontologia para o turismo “eletrônico” (e-turismo) utilizando a linguagem OWL, e serve como ponto de partida para aquisição de conhecimentos avançados em OWL. O capítulo 11 apresenta um projeto piloto desenvolvido com o propósito de alcançar interoperabilidade semântica e integração de dados semânticos na área governamental. O décimo - segundo capítulo do livro discute a aplicação de serviços e processos Web, bem como o papel da semântica na bioinformática. Esse capítulo é uma leitura básica para a compreensão das aplicações de serviços Web semânticos nas ciências biológicas e a bioinformática associada. Segundo Sahoo (2006), embora existam numerosos serviços Web que oferecem acesso a recursos biológicos, muitos destes recursos são ferramentas computacionais isoladas, pois a interoperabilidade entre estes recursos é mínima. Esse capítulo descreve a genômica computacional, a proteômica computacional e a bioinformática estrutural, além de apresentar um estudo de caso em glicoproteômica. O penúltimo capítulo do livro de Cardoso e Sheth trata do projeto, desenvolvimento e implementação de sistemas orientados por serviços de negócios semânticos para o marketing de produtos agrícolas. O último capítulo do texto de Cardoso e Sheth apresenta estruturas que suportam o desenvolvimento programático de ontologias OWL, e discute aquelas mais utilizadas pela comunidade de desenvolvedores, como Jena, Protégé-OWL API e WonderWeb OWL API, todas disponíveis para a linguagem Java. Esse capítulo detalha como as aplicações

Web semântica podem ser desenvolvidas utilizando a estrutura Jena.

As aplicações e os exemplos de Web semântica apresentados no livro de Liyang Yu diferem daqueles relatados no livro de Cardoso e Sheth e mostram máquinas de busca e ontologias, com o objetivo de familiarizar o leitor com as ferramentas da Web semântica: RDF, RDF *schema* e OWL. Assim, o capítulo sete detalha o projeto de pesquisa Swoogle, da Universidade de Maryland, que é uma máquina de busca para a Web semântica, na Web. Swoogle visita um número grande de páginas na Web e abre todos ou quase todos os *hyperlinks* da página visitada. Liyang Yu mostra a arquitetura, o fluxo de dados e exemplos de como utilizar Swoogle para encontrar documentos semânticos na Web. O oitavo capítulo trata do projeto *Friend of a Friend* – FOAF, que se propõe criar uma Web de páginas que descrevem pessoas, os elos entre as pessoas e aquilo que essas pessoas criam e fazem, e que podem ser lidas por máquinas. Neste capítulo, Liyang Yu discute o conceito e as idéias do projeto FOAF, inclusive as ontologias que se relacionam com o projeto, e apresenta alguns exemplos, inclusive como criar e publicar seu próprio documento FOAF na Web e como inseri-lo no círculo de amigos. Com base nos exemplos práticos Swoogle e FOAF, Liyang Yu enfatiza a necessidade de conexão entre o mundo da semântica e o mundo da Web. Neste ponto, assim como Cardoso e Sheth (2006), Liyang Yu argumenta que esta conexão se dá através da marcação das páginas Web e apresenta, no nono capítulo, exemplos de como adicionar semântica aos documentos Web. No capítulo seguinte, o autor retorna às máquinas de busca Web semânticas, desta vez como um exemplo da utilização de metadados adicionados por marcação semântica.

Na última parte do seu livro, Liyang Yu investiga como os serviços Web podem beneficiar-se da Web semântica, concentrando-se na descoberta de serviços. Assim, o décimo - primeiro capítulo apresenta os serviços Web semânticos, junto com uma revisão detalhada dos padrões atuais para serviços Web: WSDL, Simple Object Access Protocol - SOAP e *Universal Description, Discovery, and Integration* - UDDI, e conclui que, para facilitar a descoberta automática, a composição e o monitoramento de serviços Web, a semântica deve ser adicionada aos padrões atuais. No décimo – segundo capítulo, o autor apresenta as idealizações e as características da linguagem de marcação OWL-S, que pode ser utilizada para expressar formalmente a semântica de um serviço Web. No décimo – terceiro capítulo, Liyang Yu mostra duas abordagens para introdução de semântica na descrição dos serviços Web: uma das abordagens consiste em adicionar semântica em serviços Web inserindo anotações semânticas nos padrões atuais de serviço Web, tais como WSDL-S ou UDDI; a outra solução é mais completa e utiliza uma ontologia de alto nível, OWL-S. A vantagem do caminho que adota WSDL-S é reutilizar padrões e ferramentas atualmente disponíveis, como WSDL. Com a utilização de OWL-S, um agente automático terá informação suficiente para descobrir, invocar, compor e monitorar o serviço, pois qualquer serviço pode ser des-



crito utilizando-se ontologias de alto nível. Este capítulo ainda mostra como mapear um documento OWL-S para estruturas de dados UDDI, resultando em um registro UDDI estendido semanticamente, que funciona como um repositório centralizado, facilitando a descoberta automática de serviços Web requisitados. Entretanto, com a indisponibilidade de registros UDDI públicos, surge a necessidade de uma proposta alternativa. Utilizando programação Java, junto com Jena Application Program Interfaces – API, Liyang Yu propõe, no décimo – quarto capítulo, a construção de uma máquina de busca de serviço Web semântico, que gerencia seus próprios registros e não depende de registros UDDI públicos. O último capítulo do livro de Liyang Yu resume os capítulos anteriores e sugere leitura adicional para aqueles que quiserem continuar os estudos sobre Web semântica e serviços Web semânticos.

Ambos os livros fornecem uma visão abrangente da Web semântica e preparam o leitor para a próxima rodada de padrões, linguagens e especificações, a exemplo do desenvolvimento da linguagem de consulta a dados em RDF SPARQL Protocol and RDF Query Language – SPARQL. Esta linguagem é considerada uma inovação relevante para Web Semântica, pois pode ser utilizada em consultas diretas, ou via mediação, a diversas fontes de dados.

Antes de concluir, convém lembrar que, embora a disputa por padrões, linguagens, especificações e demais técnicas e tecnologias necessárias ao próximo estágio evolutivo da Web seja favorável à proposta contida na Web semântica, outros especialistas discordam. Assim, Lévy (2006), entende que a interoperabilidade semântica equivale ao desenvolvimento da inteligência coletiva de base digital, e argumenta que a Web semântica não resolve a questão da interoperabilidade semântica porque a notação de conceitos em linguagem natural é arbitrária e também porque as inúmeras ontologias são incompatíveis.


Concluindo, recentemente, Berneres-Lee (2008) admitiu que a visão da Web semântica apresentada em 2001 era de alguma forma, ficção científica, pois estava baseada em um futuro ainda muito distante. Imaginava-se que a Web semântica seria implementada e as pessoas fariam tudo que os sistemas do tipo inteligência artificial poderiam fazer. Para Berneres-Lee(2008), a prática da Web semântica acontece a partir da interoperabilidade, a exemplo da integração de dados intra e entre empresas e da integração de dados científicos, além da possibilidade

de consultas aos dados integrados, a exemplo da iniciativa Linked Data<sup>2</sup>. Entretanto, como a Web semântica é um conjunto variado de tecnologias, que deverá ser capaz de realizar tarefas diferentes para comunidades diferentes, os desenvolvimentos necessários são igualmente diferentes. Assim, as necessidades da comunidade de ciências da vida para que ela seja capaz de utilizar seus dados sobre proteínas em um ambiente de Web semântica deverão ser diferentes dos passos necessários para se alcançar a interoperabilidade entre repositórios de dados de bibliotecas e de museus.

## Notas

1. Shadbolt (2006) informa que em 1994, Tim Berners-Lee já havia articulado uma visão da Web Semântica.
2. <http://linkeddata.org>

## Referências bibliográficas

- Berners-Lee T, Hendler J, Lassila O. The semantic web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American. Mai. 2001: 34-43
- Shadbolt N, Berners-Lee T, Hall W. The semantic Web revisited. IEEE Intelligent Systems. 2006; 21(3): 96-101.
- Cardoso J, Sheth AP. The semantic Web and its applications. In: Semantic Web services, processes and applications. Nova Iorque: Springer; 2006: 3-33
- Sahoo SS, Sheth A. Bioinformatics applications of Web services, Web process and role of semantics. In: Cardoso J, Sheth AP (editors), Semantic Web services, processes and applications. Nova Iorque: Springer; 2006: 306-22.
- “Sir Tim Berners-Lee Talks About the Semantic Web”, transcrição de entrevista em Podcast com Paul Miller. Fev. 2008. Disponível em: [http://talismodcasts.s3.amazonaws.com/twt20080207\\_TimBL.html](http://talismodcasts.s3.amazonaws.com/twt20080207_TimBL.html). Acessado em: 21 Jan. 2009.
- Lévy P. IEML: computational semantics in the service of collective intelligence. Ottawa: CRC/FRSC; 2006: 9 (translated by Michele Healy). Disponível em: <<http://www.ieml.org/IMG/pdf/visionieml-Initiativeen.pdf>>. Acessado em: 8 Apr. 2007. 



**RECIIS**

Revista Eletrônica de Comunicação  
Informação & Inovação em Saúde

[www.reciis.cict.fiocruz.br]

ISSN 1981-6278

*Resenhas*

## **Semantic Web Technologies: trends and research in ontology-based systems**

*John Davies, Rudi Studer & Paul Warren*

DOI: 10.3395/receis.v3i1.245pt

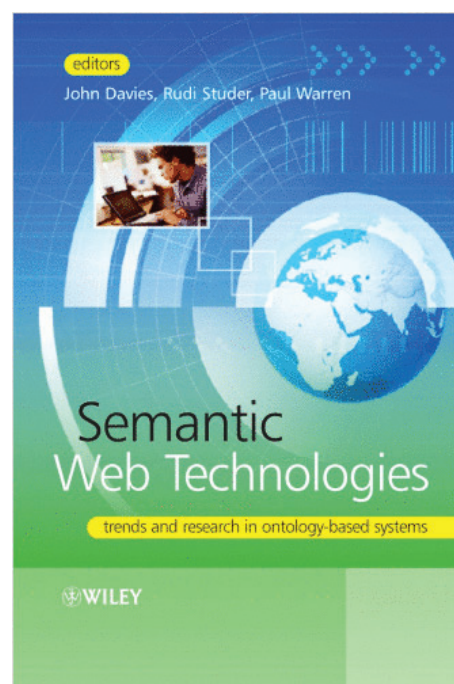
*Karin Breitman*

Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brasil  
karin@inf.puc-rio.br

Atualmente a maior parte da informação disponível na Internet está em linguagem natural, e só pode ser interpretada exclusivamente por seres humanos. Nos deparamos com esta situação diariamente quando realizamos buscas na Web e somos forçados a “filtrar” informações, que mesmo dentro dos critérios de busca solicitados, pertencem a diferentes contextos. Um exemplo simples é fazer uma pesquisa escolar sobre árvores, carvalhos em particular. O resultado desta busca, além de páginas referentes a este tipo de árvore, também traz pessoas cujo sobrenome é Carvalho, empresas de guindastes e escritórios de advocacia.

Tim Berners Lee, aclamado como um dos pais da Internet, aposta no aparecimento de uma Web Semântica no futuro próximo. Nesta Web, a informação estaria disponível para o consumo humano mas também seria formatada de modo a permitir o processamento automático das fontes de informação por parte de computadores. Para ele a *Web Semântica pode ser definida como uma EXTENSÃO da web atual na qual é dada a informação um SIGNIFICADO bem definido, permitindo com que computadores e pessoas trabalhem em cooperação.*” (Berners-Lee, Hendler e Lassila). Esta nova Web vai permitir que os computadores sejam capazes de interpretar e processar estas informações, estimada na casa de bilhões de páginas.

De forma a viabilizar esta situação, será necessário combinar uma série de tecnologias já estabelecidas e uma série de outras emergentes. O livro *Semantic Web Technologies* se propõe a discutir estas tecnologias com



*USA, Willey; 2006*

*ISBN: 978-0470025963*

o foco no ciclo de vida de ontologias: criação, utilização e gerência.

O livro é estruturado como uma coleção de artigos, organizados em capítulos. Este fato não fica muito claro para o leitor, a princípio, uma vez que o sumário suprime os nomes dos autores. Apesar de contar com uma constelação de autores reconhecidos na área, apresenta diferenças sensíveis de qualidade entre os capítulos. Os pontos fortes do livro são as discussões acerca da descoberta de informações e de inferência na presença de inconsistência. Muito interessante e original é o capítulo sobre evolução de ontologias, um assunto muito pouco discutido na literatura. O capítulo de anotação semântica também vale um destaque, pela qualidade e amplitude da revisão realizada. Apesar de apresentar as informações de forma muito resumida, apresenta referências bibliográficas atuais e será de grande valor para os interessados no assunto.

Outros assuntos tais como metodologias para construção de ontologias e mediação, técnicas de alinhamento e geração automática de ontologias já apareceram exaustivamente na literatura e só servem para engrossar o volume.

A parte final do livro apresenta casos de estudo onde algumas das tecnologias discutidas no livro foram utilizadas na prática. São três os domínios de aplicação. O primeiro, e bastante óbvio, é o de bibliotecas digitais. O segundo apresenta um inexpressivo protótipo de sistema baseado em ontologias para responder questões ligadas ao sistema judiciário espanhol. O último capítulo descreve situações mais interessantes de utilização de mediação semântica na indústria de telecomunicações.

Em minha experiência particular, cruzei com poucos livros do tipo compilação de artigos em que todos os capítulos traziam alguma contribuição. O Semantic Web Technologies não é exceção, mas apresenta um balanço final positivo. 