

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 (10^a) 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.

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

Quadro 1 – Extraído da Classificação Internacional de Doenças, 10^a 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.

MeSH Heading	Leishmaniasis		
Tree Number	C03.752.700.500.508		
Tree Number	C03.858.560		
Tree Number	C17.800.838.775.560		
Annotation	protozoan infect; GEN or unspecified; prefer specifics; American leishmaniasis is LEISHMANIASIS, AMERICAN see LEISHMANIASIS, CUTANEOUS; tegumentary leishmaniasis = LEISHMANIASIS, CUTANEOUS		
Scope Note	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).		
Allowable Qualifiers	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		
Date of Entry	19990101		
Unique ID	D007896		
<table style="width: 100%; border: none;"> <tr> <td style="width: 50%; vertical-align: top;"> Parasitic Diseases [C03] Protozoan Infections [C03.752] Sarcomastigophora Infections [C03.752.700] Mastigophora Infections [C03.752.700.500] Leishmaniasis [C03.752.700.500.508] </td> <td style="width: 50%; vertical-align: top;"> Skin and Connective Tissue Diseases [C17] Skin Diseases [C17.800] Skin Diseases, Infectious [C17.800.838] Skin Diseases, Parasitic [C17.800.838.775] Leishmaniasis [C17.800.838.775.560] </td> </tr> </table>		Parasitic Diseases [C03] Protozoan Infections [C03.752] Sarcomastigophora Infections [C03.752.700] Mastigophora Infections [C03.752.700.500] Leishmaniasis [C03.752.700.500.508]	Skin and Connective Tissue Diseases [C17] Skin Diseases [C17.800] Skin Diseases, Infectious [C17.800.838] Skin Diseases, Parasitic [C17.800.838.775] Leishmaniasis [C17.800.838.775.560]
Parasitic Diseases [C03] Protozoan Infections [C03.752] Sarcomastigophora Infections [C03.752.700] Mastigophora Infections [C03.752.700.500] Leishmaniasis [C03.752.700.500.508]	Skin and Connective Tissue Diseases [C17] Skin Diseases [C17.800] Skin Diseases, Infectious [C17.800.838] Skin Diseases, Parasitic [C17.800.838.775] Leishmaniasis [C17.800.838.775.560]		

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).

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

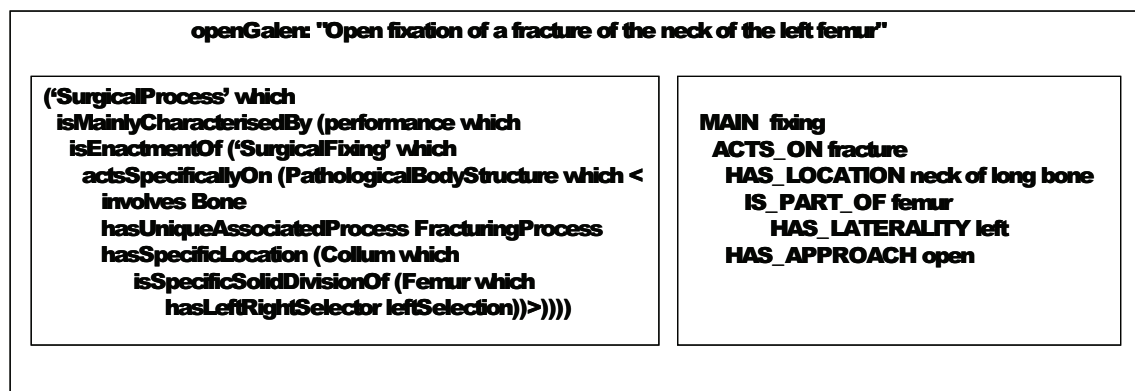
openGALEN

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.

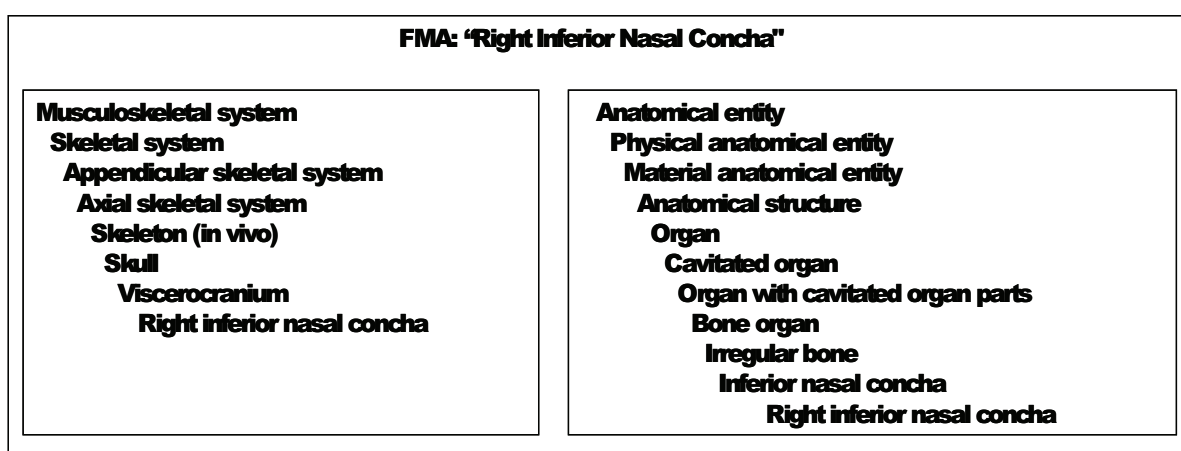
Foundational Model of Anatomy

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 envidaram 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 Metatesauro 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	www.who.int/classifications/apps/icd/
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®	www.pubmed.gov
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.	www.ihtsdo.org
GO	Componentes celulares, funções moleculares, processos biológicos	OBO/OWL	24,500 termos (2008)	Pesquisa de genes, proteínas	www.geneontology.org
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	www.opengalen.org
FMA	Conteúdo anatômico	Frames e (parcialmente) OWL	75.000 classes	Educação, pesquisa biomédica	http://sig.biostr.washington.edu/projects/fm/AboutFM.html
OBO	Bioinformática e Biologia Molecular	OBO/ OWL / OBO_XML / RDF	60 ontologias	Utilizado como repositório e esquema unificado para interoperar projetos biomédicos	www.obofoundry.org
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	http://www.nlm.nih.gov/research/umls/

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 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) 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. 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 BAOSW (*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).