

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, Freiburg, Alemanha
steschulz@uni-freiburg.de



Holger Stenzhorn

Instituto de Biometria Médica e Informática da Medicina, Centro Médico Universitário Freiburg, Freiburg, Alemanha
holger.stenzhorn@uniklinik-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 (Shortliffe 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. Afirmações 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 inconveniente¹, 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 :

$$Fígado \sqsubseteq \text{Órgão_do_Corpo}$$

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:

$$\text{Biópsia_Planejada} \sqsubseteq \text{Situação} \sqcap \exists \text{procedimento_associado.} \\ \text{Biópsia} \sqcap \dots$$

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):

$$\text{Prevenção_do_Abuso_de_drogas} \sqsubseteq \text{Procedimento} \sqcap \exists \text{tem_} \\ \text{foco.Abuso_de_Drogas}$$

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:

$$\text{Prevenção_ao_Abuso_de_Drogas} \sqsubseteq \text{Procedimento} \sqcap \\ \exists \text{tem_participante.Pessoa} \sqcap \exists \text{causas.} (\text{Estado} \sqcap \exists \text{tem_} \\ \text{participante.} (\text{Pessoa} \sqcap \exists \text{participa_de.} \neg \text{Abuso_de_Drogas}))$$

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.

Categories 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-dimensionais, 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 prototípais; 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 prototípico, 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 indecidíveis.

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 instanciarções 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:

$$\text{Lmo-2.7760102} \in \text{Porção_de_Lmo-2}$$

$$\text{Elf-2.776010} \in \text{Porção_de_Elf-2}$$

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

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

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

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> ∈ <i>Município</i>	<i>Polegar</i> ⊆ <i>Dígito</i>
<i>Condado de Orange</i> ∈ <i>Condado</i>	<i>Mão</i> ⊆ <i>Parte_do_Corpo</i>
<i>Flórida</i> ∈ <i>Estado</i>	<i>Extremidade_superior</i> ⊆ <i>Membro</i>
<i>EUA</i> ∈ <i>País</i>	<i>Corpo</i> ⊆ <i>Estrutura_Anatômica</i>
< <i>Orlando, Condado de Orange</i> > ∈ <i>parte_de</i>	<i>Polegar</i> ⊆ ∃ <i>parte_da.Mão</i>
< <i>Condado de Orange, Flórida</i> > ∈ <i>parte_de</i>	<i>Mão</i> ⊆ ∃ <i>parte_da.Extremidade_superior</i>
< <i>Flórida, EUA</i> > ∈ <i>parte_de</i>	<i>Extremidade_Superior</i> ⊆ ∃ <i>parte_do.Corpo</i>

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 Metatesauro 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 truísmos 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_sangüíneo	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_freqüente	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 I 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. 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.
- Kusniercyk 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.meddransso.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. 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.

Welty 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.

WHOCC (WHO Collaborating Centre for Drug Statistics Methodology). Anatomical Therapeutic Chemical Classification System (ATC), 2008. Available from: <http://www.whocc.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.