

Kátia Cardoso Coelho

**Aquisição de conhecimento especializado
para construção de ontologias: um estudo no
domínio das ciências da vida.**

Belo Horizonte

Escola de Ciência da Informação da UFMG

2012

Kátia Cardoso Coelho

**Aquisição de conhecimento especializado
para construção de ontologias: um estudo no
domínio das ciências da vida.**

Dissertação apresentada ao Programa de Pós-Graduação da Escola de Ciência da Informação, da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de mestre em Ciência da Informação.

Linha de Pesquisa: Organização e Uso da Informação (OUI)

Orientador: Prof. Maurício B. Almeida

Co-orientadora: Dra Anna Bárbara F. C. Proietti.

Belo Horizonte

2012

Coelho, Kátia Cardoso.

C672a Aquisição de conhecimento especializado para construção de ontologias [manuscrito] : um estudo no domínio das ciências da vida / Kátia Cardoso Coelho. – 2012.
193 f. : il., enc.

Orientador: Maurício B. Almeida.
Co-orientadora: Anna Bárbara F. C. Proietti.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Escola de Ciência da Informação.
Referências: f. 175-183
Anexos: f. 184-193

1. Ciência da informação – Teses. 2. Ontologias (Recuperação da informação) – Teses. 3. Aquisição de conhecimento (Sistemas especialistas) – Teses. 4. Biomedicina – Terminologia – Teses. I. Título. II. Almeida, Maurício Barcellos. III. Proietti, Anna Bárbara de Freitas Carneiro. IV. Universidade Federal de Minas Gerais, Escola de Ciência da Informação.

CDU: 025.4:61



UFMG

Universidade Federal de Minas Gerais
Escola de Ciência da Informação
Programa de Pós-Graduação em Ciência da Informação

FOLHA DE APROVAÇÃO

"AQUISIÇÃO DE CONHECIMENTO ESPECIALIZADO PARA CONSTRUÇÃO DE ONTOLOGIAS: UM ESTUDO NO DOMÍNIO DAS CIÊNCIAS DA VIDA "

Kátia Cardoso Coelho

Dissertação submetida à Banca Examinadora, designada pelo Colegiado do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Minas Gerais, como parte dos requisitos à obtenção do título de "**Mestre em Ciência da Informação**", Linha de Pesquisa: "**Organização e Uso da Informação - OUI**".

Dissertação aprovada em: 28 de junho de 2012.

Por:

Prof. Dr. Mauricio Barcellos Almeida - ECI/UFMG (Orientador)

Profa. Dra. Anna Bárbara Freitas Carneiro Proietti - Hemominas (Co-orientadora)

Prof. Dr. Renato Rocha Souza - FGV/RJ

Profa. Dra. Marta Macedo Kerr Pinheiro - ECI/UFMG

Aprovada pelo Colegiado do PPGCI

Profa. Gercina Ângela B. O. Lima
Coordenadora

Versão final Aprovada por

Prof. Mauricio Barcellos Almeida
Orientador



UFMG

Universidade Federal de Minas Gerais
Escola de Ciência da Informação
Programa de Pós-Graduação em Ciência da Informação

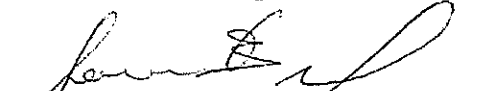
ATA DA DEFESA DE DISSERTAÇÃO DE **KÁTIA CARDOSO COELHO**, matrícula:
2010656118

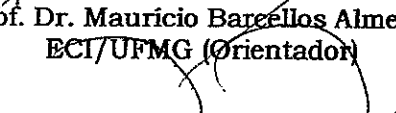
Às 14:00 horas do dia 28 de junho de 2012, reuniu-se na Escola de Ciência da Informação da UFMG a Comissão Examinadora aprovada pelo Colegiado do Programa de Pós-Graduação em Ciência da Informação em 28/05/2012, para julgar, em exame final, o trabalho intitulado ***Aquisição de conhecimento especializado para construção de ontologias: um estudo no domínio das ciências da vida***, requisito final para obtenção do Grau de MESTRE em CIÊNCIA DA INFORMAÇÃO, área de concentração: Produção, Organização e Utilização da Informação, Linha de Pesquisa: Organização e Uso da Informação - OUI. Abrindo a sessão, o Presidente da Comissão, Prof. Dr. Maurício Barcellos Almeida, após dar conhecimento aos presentes do teor das Normas Regulamentares do Trabalho Final, passou a palavra à candidata para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa da candidata. Logo após, a Comissão se reuniu sem a presença da candidata e do público, para julgamento e expedição do resultado final. Foram atribuídas as seguintes indicações:

Prof. Dr. Maurício Barcellos Almeida - Orientador	APROVADA
Profa. Dra. Anna Bárbara Freitas Carneiro Proietti - Co-orientadora	APROVADA
Prof. Dr. Renato Rocha Souza	APROVADA
Profa. Dra. Marta Macedo Kerr Pinheiro	APROVADA

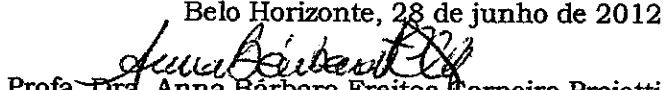
Pelas indicações, a candidata foi considerada APROVADA.

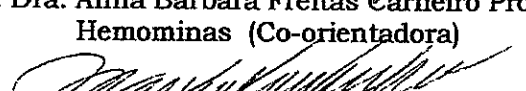
O resultado final foi comunicado publicamente à candidata pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a sessão, da qual foi lavrada a presente ATA que será assinada por todos os membros participantes da Comissão Examinadora.


Prof. Dr. Maurício Barcellos Almeida
ECI/UFMG (Orientador)

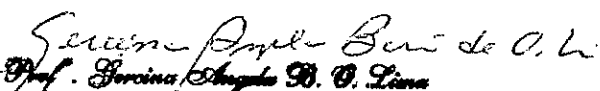

Prof. Dr. Renato Rocha Souza
FGV/RJ

Belo Horizonte, 28 de junho de 2012


Profa. Dra. Anna Bárbara Freitas Carneiro Proietti
Hemominas (Co-orientadora)


Profa. Dra. Marta Macedo Kerr Pinheiro
ECI/UFMG

Obs: Este documento não terá validade sem a assinatura e carimbo da Coordenadora.


Prof.ª Joviana Augusto B. O. Lima
Coordenadora do Programa de Pós-Graduação
em Ciência da Informação - ECI/UFMG

DEDICATÓRIA

A meus pais, Idalmo e Margarida, cuja distância não foi motivo de ausência nesta fase de grande aprendizado.

Aos meus irmãos, Adriano, Flaviano e Rodrigo, e meus sobrinhos Andressa e Arthur, por fazerem parte da minha vida e sempre entenderem os motivos pelos quais tantas vezes não nos encontramos.

Aos meus amigos, pela força e compreensão, pelas vezes que não estivemos juntos. Amizade é a família que escolhemos para a vida.

À amizade pura e verdadeira, que sempre me apoiou e esteve ao meu lado nesta jornada.

AGRADECIMENTOS

Ao meu mestre, Prof. Dr. Maurício Barcellos Almeida, por ter me ensinado que um pesquisador precisa: de disciplina para traçar os caminhos; de coragem para os desafios e de inquietude para buscar respostas às perguntas;

À Escola de Ciência da Informação, por abrir as portas à minha entrada na pesquisa;

Aos professores da Ciência da Informação, pela oportunidade e grande contribuição para o meu aprendizado.

À pesquisadora, Dra. Anna Bárbara Freitas Carneiro Proietti, pela convivência e oportunidade de aprender. Pesquisar: paciência e persistência;

Aos pesquisadores do Grupo Interdisciplinar de Pesquisa em HTLV (GIPH), pela troca de experiência, enriquecimento e importante contribuição para o meu trabalho;

À Fundação Hemominas, pelo incentivo e colaboração para a execução desta pesquisa;

À Fapemig, pelo incentivo à pesquisa.

À minha diretora, à minha gerente e colegas de trabalho, pela compreensão e respeito ao meu trabalho que exigia de mim dedicação e presença.

Nomes não foram citados, isso não quer dizer que não me lembre de cada um, que me ajudou e incentivou. Se o fizesse, correria o risco de deixar de fora aqueles que deveriam fazer parte desta luta, assim, a todos vocês, muito obrigada!

“É útil ter examinado a pluralidade de significados de um termo, tanto no interesse da clareza (pois um homem está mais apto, a saber, o que afirma quando tem uma noção nítida do número de significados que a coisa pode comportar), como para nos certificarmos de que o nosso raciocínio está de acordo com os fatos reais e não se referirá apenas aos termos usados. Pois, enquanto não ficar bem claro em quantos sentidos se usa um termo, pode acontecer que o que responde e o que interroga não tenham suas mentes dirigidas para a mesma coisa; ao passo que, depois de se haver esclarecido quantos são os significados, e também qual deles o primeiro tem em mente quando faz a sua asserção, o que pergunta pareceria ridículo se deixasse de dirigir seus argumentos a esse ponto.”

Aristóteles (384 a. C. – 322 a. C.)

COELHO, Kátia C. *Aquisição de conhecimento especializado para construção de ontologias: um estudo no domínio das ciências da vida*. Dissertação de mestrado – Universidade Federal de Minas Gerais, Escola de Ciência da Informação.

RESUMO

Obter conhecimento especializado de um dado domínio do conhecimento é um desafio para diversos campos científicos. Em áreas correlatas às ciências da vida a terminologia é abrangente e complexa sujeita a diferentes interpretações para termos especializados. Esse tipo de problema, dentre outros relacionados, tem sido abordado há anos pela Ciência da Informação. Seja para modelagem de sistemas, construção de tesauros, desenvolvimento de ontologias, ou mesmo para gestão do conhecimento, tal atividade exige planejamento, organização e emprego de método e técnicas apropriados. O objetivo geral do presente trabalho é buscar alternativas para minimizar a distância entre o que o especialista tem a oferecer e aquilo que é de fato registrado como o que ele sabe. Para tal, foi proposto um roteiro para elicitación de conhecimento, para obtenção de conhecimento especializado sobre Mielopatia associada ao HTLV I. Procurou-se levantar ainda as principais dificuldades envolvidas na aquisição de conhecimento, como forma de buscar melhorias nesta atividade. Para a realização do presente trabalho, procedeu-se à revisão de literatura, perpassando pelas diferentes áreas científicas que abordam a temática, situando-a na Ciência da Informação e Biomedicina. Buscou-se por contribuições provenientes das áreas pesquisadas, bem como pelas técnicas empregadas para o estabelecimento da metodologia. O universo empírico de dados corresponde à participação de pesquisadores do Grupo de Pesquisa em HTLV - GIPH, bem como a sua produção científica. Uma contribuição evidente da pesquisa é o roteiro metodológico descrito, que possibilitou que se levantassem os principais termos do domínio, candidatos à ontologia.

Palavras-chave: aquisição de conhecimento, expertise, conhecimento especializado, terminologia biomédica, construção de ontologias

COELHO, Kátia C. *Specialized knowledge acquisition for ontology construction: a study in the field of life sciences*. Degree of Master – Universidade Federal de Minas Gerais, Escola de Ciência da Informação.

ABSTRACT

Getting expert knowledge of a given domain of knowledge is a challenge for many scientific fields. In areas related to life sciences, terminology is broad and complex subject to different interpretations of specialized terms. This kind of problem, among others related, has been approached for years by Information Science. Whether for modeling systems, construction of thesauri, ontology development, or even for knowledge management, this activity requires planning, organization and use of appropriate methods and techniques. The overall objective of this study is to seek alternatives to minimize the distance between what the expert has to offer and what is actually recorded as what he knows. To this end, we propose a methodology for knowledge elicitation, to obtain specialized knowledge associated myelopathy HTLV I. We sought to further raise the main difficulties involved in acquiring knowledge, in order to seek improvements in this activity. For the present work, we proceeded to review the literature, spanning the different scientific fields that address the issue, placing it in the Information Science and Biomedicine. Was sought by contributions from the areas studied, as well as the techniques used for establishing the methodology. The empiric universe of data corresponds to the participation of researchers from the Research Group on HTLV - GIPH, as well as its scientific production. A clear contribution of the research methodology is described in the script, which made it possible to raise the main terms of domain ontology candidates.

Keywords: Knowledge acquisition, expertise, specialized knowledge, biomedical terminology, ontology construction

LISTA DE FIGURAS

FIGURA 1: AC baseada na extração da Informação no TIMS	33
FIGURA 2: Processo de curadoria do IEDB	36
FIGURA 3: Estrutura da PDBWiki: <i>Protein Data Bank</i> Fonte: Adaptado de Stehr <i>et al</i> (2010) ...	39
FIGURA 4: Técnicas de AC de acordo com o conhecimento elicitado.	44
FIGURA 5: Conjunto de objetos segundo os critérios "veículos" e "animais".	48
FIGURA 6: Conjunto 1 segundo os critérios "veículos terrestres" e "veículos aquáticos"	49
FIGURA 7: Resultado da BLO-Core na Internet	59
FIGURA 8: Esquema da relação de dependência das entidades continuantes.....	64
FIGURA 9: Taxonomia das entidades SNAP.....	66
FIGURA 10: Taxonomia das entidades SPAN.....	71
FIGURA 11: Representação gráfica do ciclo de vida da ontologia	76
FIGURA 12: Quadro clínico sob a perspectiva do curso da doença	107
FIGURA 13: Processo etiológico, desordem e doença (disposição)	107
FIGURA 14: Doença, processo patológico e condições anormais.....	108
FIGURA 15: Sinais, sintomas e processo interpretativo.....	108
FIGURA 16: Roteiro no <i>Protegé-Frames</i>	109
FIGURA 17: Tela da ferramenta Protégé com um fragmento da BLO	115
FIGURA 18: Publicação anual do GIPH	126
FIGURA 19: Numero de citações das publicações do GIPH.....	126
FIGURA 20: Fragmento de artigo analisado durante a Imersão de Literatura.....	128
FIGURA 21: <i>Template do Protege-Frames</i> com dados sobre a doença HAM/TSP	145
FIGURA 22: <i>Template do Protege-Frames</i> com dados sobre a doença HAM/TSP, após <i>teachback</i> com especialista.	148
FIGURA 23: <i>Template do Protege</i> com dados para HAM/TSP, na fase de refinamento.....	163

LISTA DE QUADROS

QUADRO 1: Base teórica para AC aplicada à Biomedicina.....	30
QUADRO 2: Métodos de AC conceitual	31
QUADRO 3: Consolidado de técnicas para Aquisição de Conhecimento.....	53
QUADRO 4: Classificação conforme os tipos de ontologias.....	56
QUADRO 5: Breve descrição das metodologias	74
QUADRO 6: Categorias dos níveis da <i>expertise</i>	87
QUADRO 7: Roteiro metodológico proposto	99
QUADRO 8: Termos de referência para o roteiro da entrevista	104
QUADRO 9: Formulário de agenda para elicitación de conhecimento	110
QUADRO 10: Formulário utilizado para levantamento dos especialistas.....	111
QUADRO 11: Formulário para síntese de entrevista	111
QUADRO 12: Exemplo de levantamento dos especialistas do GIPH.....	130
QUADRO 13: Matriz pesquisador X Áreas de pesquisa do GIPH, conforme cadastro CNPq. ...	132
QUADRO 14: Mapeamento dos especialistas participantes da atividade de elicitación de conhecimento	134
QUADRO 15: Etapas das atividades de Aquisição de Conhecimento	138
QUADRO 16: Agenda para elicitación de conhecimento	139
QUADRO 17: Síntese de entrevista não estruturada	142
QUADRO 18: Representação visual de fragmento registrado na <i>template do Protegé Frames</i>	146
QUADRO 19: Transcrição das sessões de elicitación de conhecimento	150
QUADRO 20: Curso da doença HAM/TSP.....	158
QUADRO 21: Validação dos termos pelo especialista	160

LISTA DE TABELAS

TABELA 1: Matriz resultante do uso da técnica de grade de repertório.....	47
TABELA 2: Distribuição mundial das publicações relacionadas ao HTLV	118
TABELA 3: Relação de periódico e respectivo fator de impacto	120
TABELA 4: Levantamento da produção científica do GIPH em periódicos com Fator de Impacto e Citações da produção científica.	122
TABELA 5: Extrato da produção científica dos especialistas entrevistados	136

LISTA DE SIGLAS

SIGLA – NOME COMPLETO

AC - Aquisição de Conhecimento

AD - Análise de Domínio

BFO - *Basic Formal Ontology*

BLO – *Blood Ontology*

CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico

DOLCE - *Descriptive Ontology for Linguistic and Cognitive Engineering*

Fapemig - Fundação de Amparo à Pesquisa de Minas Gerais

FI - Fator de Impacto

FIOCRUZ - Fundação Oswaldo Cruz

FRBR - *Functional Requirements for Bibliographic Records*

Fhemig - Fundação Hospitalar do Estado de Minas Gerais

GIPH - Grupo de Pesquisa em HTLV

GO - *Gene Ontology*

HAM/TSP - HTLV *associated myelopathy /Tropical spastic paraparesis*

HTLV - *Human T lymphotropic virus*

IEDB - *Immune Epitope Database and Analysis Resource*

ICI - Índice de Citação Imediata

IO - Índice de obsolescência

ISI - *Institute for Scientific Information*

ISK - *Informatics in the Scientific Knowledge*

JAD - *Joint Application Design*

JCR - *Journal of Citation Reports*

KR - *Knowledge Representation Ontology*

MV - Meia-Vida das citações

NCIT - *National Cancer Institute Term*

OBO - *Open Biological and Biomedical Ontologies*

PCT - *Personal Constructs Theory*

PDB - *Protein Data Bank*

PRO - *Protein Ontology*

RI - *Recuperação da Informação*

SciELO - *Scientific Electronic Library Online*

SCI - *Science Citation Index*

SI - *Sistemas de Informação*

SNAP - *Snapshot ontologies*

SPAN - *Spanning*

SRI - *Sistema de Recuperação da Informação*

SUMO - *Suggested Upper Merged Ontology*

TIMS - *Tag Information Management System*

UFMG - *Universidade Federal de Minas Gerais*

XML - *eXtensible Markup Language*

SUMÁRIO

1	INTRODUÇÃO	15
2	REVISÃO DE LITERATURA	20
2.1	Aquisição de Conhecimento: uma visão geral	21
2.2	Aquisição de conhecimento em Ciência da Informação	25
2.3	Aquisição de conhecimento no contexto biomédico.....	29
2.4	Classificação para as técnicas de AC.....	41
2.5	Aquisição de conhecimento na construção de ontologias	55
2.5.1	Ontologias: uma breve introdução	55
2.5.1.1	Ontologia de domínio	58
2.5.1.2	Ontologia de alto nível.....	61
2.5.2	Construção de ontologias: uma breve descrição de metodologias.....	73
2.5.3	Considerações complementares sobre AC para construção de ontologias	77
2.5.3.1	A análise do domínio	77
2.5.3.2	Especialista e expertise	84
3	METODOLOGIA DE PESQUISA.....	89
3.1	A instituição objeto da pesquisa	90
3.1.1	Histórico, missão e visão institucional.	90
3.1.2	A Hemominas como instituição de pesquisa	92
3.1.3	Motivação para a escolha do grupo de pesquisa na instituição	95
3.2	Condução e recorte da pesquisa	97
3.3	Descrição da metodologia de pesquisa	99
3.4	Instrumentos de pesquisa	100
3.4.1	Técnicas adotadas para coleta de dados	100
3.4.2	Instrumentos para coleta de dados	102
4	COLETA DE DADOS E ANÁLISE DE RESULTADOS	112
4.1	Levantamento	114
4.1.1	Escopo da Ontologia.....	114
4.1.1.1	Audiência	114
4.1.1.2	Ontologia de alto nível: fundamento da ontologia em desenvolvimento	115
4.1.1.3	Informações básicas sobre o contexto	116
4.1.2	Escopo básico do domínio.....	116
4.1.3	Expertise dos especialistas	128
4.2	Contato	139
4.2.1	Obter conhecimento	139
4.2.2	Esclarecer dúvidas e mal entendidos.....	147

4.2.3	Levantar e registrar termos.....	158
4.3	Validação.....	160
4.4	Refinamento.....	162
4.5	Apontando dificuldades	164
5	CONSIDERAÇÕES FINAIS.....	170
	REFERÊNCIAS.....	175

1 INTRODUÇÃO

Um mundo é dado ao homem; sua glória não é suportar ou depreciar
este mundo, mas sim enriquecê-lo construindo outros universos.
BUNGE

Estudar o modo como o conhecimento individual e o conhecimento produzido por grupos é traduzido em conhecimento especializado de um campo científico é essencial no âmbito da pesquisa em diversos segmentos. Neste contexto, a Ciência da Informação tem procurado cada vez mais buscar melhores formas para organização e representação (BATES, 1999). As tarefas inerentes a essa responsabilidade, entretanto, não são triviais uma vez que o conhecimento é de difícil apreensão e se encontra disperso em diferentes formatos, seja incorporado em pessoas, registrado em documento ou representado em sistemas de informação. A complexidade da tarefa é ainda maior ao considerar as dificuldades em compreender as terminologias especializadas e os problemas daí advindos

A preocupação com a organização do conhecimento não é recente. Iniciativas pioneiras foram, por exemplo, as propostas de *Linnaeus*¹ para a biologia e a *Sydenham*² para a medicina. Na Ciência da Informação, diferentes técnicas de classificação e de representação - oriundas da Biblioteconomia desde o século XIX e Documentação - têm sido utilizadas há bastante tempo: taxonomias, esquemas de classificação, vocabulários controlados, sistemas de classificação, para citar algumas. Estas técnicas representam a fundamentação básica e essencial para iniciativas de organização da informação, tanto do ponto de vista teórico quanto metodológico. No entanto, como o volume de informação a ser processado e armazenado tem aumentado significativamente nos últimos 20 ou 30 anos, representar o conhecimento e a informação para uso por pessoas e por sistemas de informação automatizados têm trazido desafios.

Nesse contexto, ontologias têm sido propostas como alternativa para criação de representações da realidade. De fato, desde os anos de 1980, tais estruturas de representação são objeto de estudo em campos como a Inteligência Artificial – como

¹ *Carolus Linnaeus* (1707-1778), botânico e naturalista sueco, criador do sistema taxonômico em biologia

² *Thomas Sydenham* (1624-1689), médico inglês, criador da estrutura de classificação de doenças

forma de apreender, formalizar e explicitar o conhecimento a ser expresso em sistemas especialistas. Na Ciência da Informação, pela tradição no uso de estruturas de categorização para organização da informação, as ontologias têm sido objeto de estudo desde os anos 90 (VICKERY, 1997). Estudos sobre ontologia em Ciência da Informação podem ser também entendidos como aqueles que buscam compreender os processos que envolvem princípios semelhantes aos já utilizados na construção de tesouros e taxonomias.

De forma simples, ontologias são compostas por regras que regulam a combinação entre *termos*, organizados em uma taxonomia e as *relações* entre esses termos em um dado domínio do conhecimento. Além disso, permitem a formulação de consultas a *instâncias* representativas de objetos reais (ALMEIDA; BAX 2003). No âmbito da Representação do Conhecimento, ontologias são especificações formais e explícitas de uma conceitualização compartilhada (GRUBER, 1993). Nesse contexto, *conceitualização* se refere a um modelo abstrato dos fenômenos no mundo, onde se identificam os conceitos relevantes daqueles fenômenos; *formal* explica que ontologia pode ser interpretada por máquinas; *explícito* sugere que os tipos e as restrições de uso são definidos explicitamente e *compartilhado* propõe que ontologia retém conhecimento de comunidades específicas. Tal definição, ainda que muito difundida, não é consensual.

Ontologias são também definidas a partir de um conjunto de atividades que compõem o processo de sua própria construção. Tratam-se de atividades que compreendem diferentes fases e envolvem o emprego de métodos e técnicas variados (USCHOLD; GRUNINGER, 1996; GÓMEZ-PÉREZ *et al*, 2004). Fernandez, Gómez-Pérez e Juristo (1997) apresentam: a *especificação*, a *conceitualização*, a *formalização*, a *integração*, a *implementação* e a *manutenção* como fases no desenvolvimento de ontologias. Acrescentam ainda atividades essenciais e complementares realizadas ao longo de toda a sua construção. Essas compreendem: *aquisição de conhecimento*, *documentação* e *avaliação da ontologia*. O estudo acerca de tais atividades é essencial para que se possa alcançar uma representação via ontologias e, portanto, de interesse científico para a Ciência da Informação.

A presente pesquisa se insere neste contexto, enfatizando a atividade de *Aquisição de Conhecimento*, a qual não se trata de uma etapa totalmente definida, mas simultânea a outras. Trata-se de uma atividade fundamental, ainda que considerada um “gargalo”. Carece ainda de estudos acerca de seus problemas e de uma metodologia

confiável para obter conhecimento de especialistas de domínios diversos (HAYES-ROTH; WATERMAN; LENAT, 1983; ZAF; MCNEESE; SNYDER, 1993; MASTELLA, 2005). Aquisição de Conhecimento é um termo amplo que compreende extrair, estruturar e organizar o conhecimento de uma ou mais fontes (TURBAN, 1992). É uma tarefa indispensável para a coleta de informações visando à organização, representação e validação do conhecimento (MASTELLA, 2005).

A expressão Aquisição de Conhecimento, daqui em diante referenciada como AC, foi empregada pela primeira vez na literatura da Engenharia do Conhecimento e orientava-se para a construção de bases de conhecimento, as quais correspondem ao componente que armazena dados em um sistema especialista (MILTON et al, 2006). No contexto do presente trabalho, referencia-se a AC executada ao longo da construção de ontologias, conforme exposto em Fernandez, Gomez-Perez e Juristo (1997), em Gómez-Pérez *et al.* (2004) e em Almeida (2006).

Cabe nesse momento, para maior clareza, explicar dois termos que aparecem na literatura abordada e dessa forma, empregados ao longo do trabalho: *engenheiro do conhecimento* é o agente responsável por compreender o domínio, analisar o problema e conceber a estratégia de AC (TUTHILL, 1990; CUNHA, 1995; MASTELLA, 2005); *especialistas do domínio* (ou especialistas no assunto) são pessoas com *expertise* e substancial experiência em um campo específico (MILTON, 2007). Segundo Mastella (2005), especialistas possuem a capacidade de aplicar habilidades cognitivas para resolver problemas e, portanto, considerados a melhor fonte de conhecimento para a AC. A AC apresenta-se sob duas abordagens principais: a primeira, voltada para obtenção do conhecimento de pessoas, em que é empregado o termo *elicitação do conhecimento*³, e a segunda, que busca obter conhecimento registrado em documentos, base de dados, banco de dados, literatura especializada, dentre outras.

A AC inclui as tarefas de coleta, análise, extração e validação do conhecimento (BOOSE; GAINES, 1989; GAINES, 2003; MILTON *et al* 2006; DELUGACH; WOLF, 1996; MILTON, 2007). Para o desenvolvimento dessas tarefas, diferentes técnicas de AC têm sido propostas, tanto para a obtenção de conhecimento proveniente das pessoas, quanto em documentos. Mastella (2005) considera que as técnicas de AC a partir de

³ Vale ressaltar que embora seja empregada no português como *elicitação do conhecimento*, não foi encontrada uma tradução do termo *elicitation* para o português, cuja denominação mais próxima seria *elicitar*, ou seja, extrair, fazer sair, expulsar (HOUAISS, 2003)

fontes documentadas, por ser de natureza explícita e registrada, encontram-se melhor estruturadas, embora sejam necessárias validações sistemáticas do especialista. Por outro lado, pessoas possuem diferentes visões de mundo e maneiras pessoais de solucionar problemas, tornando complexa a tarefa de AC.

Milton (2007) explica que as dificuldades em obter conhecimento de especialistas se devem à: dificuldades do especialista em se expressar claramente; dificuldades em entender o que o engenheiro de fato precisa saber; dificuldade em alcançar o nível de detalhes requerido; de apresentar ideias de modo claro e lógico; de explicar o jargão e a terminologia; de recordar o que é relevante. Do ponto de vista do engenheiro, as principais dificuldades seriam entender e registrar o que o especialista diz; manter o foco do especialista em questões relevantes; manter a concentração para compreender os novos conhecimentos que recebe.

Nesse contexto, caracterizado por limitações inerentes às atividades de AC, define-se o **problema** a ser abordado ao longo do presente trabalho: Que fatores geram dificuldades na AC em domínios especializados como a biomedicina? Do ponto de vista da Ciência da Informação, pretende-se enquanto estudo no campo do conhecimento e da informação, buscar melhorias na tarefa de AC, no processo de construção de ontologias. De fato, a Ciência da Informação é um campo voltado para o entendimento das manifestações da informação e suas formas de representação e organização (VICKERY, 1986) e para problemas inerentes ao processo de comunicação do conhecimento entre pessoas nos contextos social, institucional e individual (SARACEVIC, 1996). Entende-se assim, que as dificuldades decorrentes dos processos de interação do engenheiro e do especialista ao longo do processo de construção de uma ontologia, compõem um tema significativo para pesquisa.

Uma longa experiência em atividades de AC, disponível na literatura de diversas áreas pesquisadas, revela a existência de um *gap* entre o engenheiro e o especialista (BOOSE, 1990; GAINES, 2003; MILTON *et al* 2006; MILTON, 2007). O **objetivo geral** do presente trabalho é buscar alternativas para minimizar a distância entre o que o especialista tem a oferecer e aquilo que é de fato registrado como o que ele sabe. Para alcançar tal objetivo geral, propõem-se os seguintes **objetivos específicos**: i) levantar técnicas e métodos de AC, disponíveis nas disciplinas científicas de interesse para a presente pesquisa, distinguindo o que cada uma se propõe; ii) realizar uma AC com especialistas do domínio especializado, a partir de um roteiro proposto, de modo a buscar melhorias no processo de construção de ontologias biomédicas ou ciências da

vida; iii) identificar e registrar os principais problemas da atividade de AC, com vistas ao desenvolvimento de ontologias.

As considerações levantadas na revisão de literatura apresentam a AC como uma atividade importante. Tal atividade, entretanto, se encontra ainda pouco discutida, em especial no processo de construção de ontologias, que, embora tenha apresentado metodologias que exploram essa atividade, a apresenta de modo superficial. Além disso, mesmo que tenha sido apresentada como sendo uma atividade que apresenta problemas - principalmente quando se visa obter conhecimento de pessoas -, carece ainda de estudos que busquem identificar tais problemas com vistas a buscar melhorias na atividade, indicando, desse modo, a viabilidade da pesquisa.

Para a realização da pesquisa desenvolveu-se a metodologia baseada nas contribuições provenientes das áreas estudadas, bem como nas técnicas empregadas na atividade de AC, a partir de pessoas. O universo empírico de dados corresponde à participação de pesquisadores do Grupo de Pesquisa em HTLV - GIPH, na qual buscou elicitare o conhecimento dos especialistas do domínio do vírus HTLV, especificamente no contexto da Mielopatia Associada ao HTLV (HAM/TSP), doença com a qual se estabeleceu o recorte da presente pesquisa. O GIPH é um grupo de pesquisa em HTLV coordenado pela Fundação Centro de Hematologia e Hemoterapia de Minas Gerais – Hemomins e instituições parceiras.

O restante do presente trabalho está organizado como segue: o capítulo 2 apresenta a revisão de literatura sobre AC apresentando uma visão geral, seguida de uma abordagem em Ciência da Informação e Biomedicina; o capítulo 3 apresenta a metodologia de pesquisa, na qual apresenta a instituição objeto da pesquisa, a descrição da metodologia de pesquisa, os instrumentos de pesquisa adotados, além da condução e recorte da pesquisa. No capítulo 4 é apresentado como foi planejada e executada a coleta dos dados, bem como a análise dos dados obtidos. Por fim, no capítulo 5 encontram-se as considerações finais e os planos para trabalhos futuros.

2 REVISÃO DE LITERATURA

É por meio do conhecimento que o homem procura a compreensão e a explicação da realidade, estabelecendo conceitos para que outras pessoas também possam adquirir os mesmos conhecimentos sem ter que realizar experiências repetitivas. Para alguém definir o conceito de alguma coisa, não basta apenas repetir palavras já decoradas, mas sim manifestar o que se sabe sobre esta coisa e o que foi aprendido, sobretudo através de experiências (RUDIO).

O presente capítulo apresenta a revisão de literatura sobre o tema principal da presente pesquisa. Na Seção 2.1 é apresentada uma visão geral sobre AC, enquanto a Seção 2.2 é voltada para a AC em Ciência da Informação e a Seção 2.3 no contexto da Biomedicina. Na Seção 2.4 discorre-se sobre os tipos de classificação encontrados na literatura pesquisada para técnicas de AC, enquanto a Seção 2.5 procura contextualizar a AC na construção de ontologias, em que são apresentados, em detalhe, o domínio - como o domínio do conhecimento é discutido na abordagem da Análise de Domínios, e os modos de se identificar especialistas e a *expertise*.

Para tal, foram realizadas buscas de artigos científicos em periódicos, trabalhos apresentados em congressos, teses e dissertações nas fontes *Google Academics* e Portal de periódicos CAPES, no período de janeiro a abril de 2011 e julho e agosto de 2011. A estratégia de busca foi procurar por estudos que abordassem: aquisição de conhecimento e elicitación de conhecimento, inclusive em domínios especializados; dificuldades para elicitación de conhecimento; técnicas e métodos diversos empregados nos campos do conhecimento pesquisados. Este levantamento procurou ainda contextualizar a atividade AC, no âmbito da Ciência da Informação e da Biomedicina. Por fim, teve o propósito de situar o tema no contexto empírico e teórico.

O critério de busca se baseou em palavras-chaves simples e compostas envolvendo termos como: aquisição de conhecimento; *knowledge acquisition*, elicitación de conhecimento; *knowledge elicitation*, representação do conhecimento, *knowledge representation*, ontologia; *ontology*; *ontologies*; análise de domínio e *domain analysis*. A busca se concentrou nas áreas correlatas à Ciência da Computação/Engenharia do conhecimento, por ter sido a pioneira em estudos sobre AC; Ciência da Informação e

Biomedicina por serem áreas de principal interesse para o presente trabalho. A pesquisa no “Portal de Periódicos CAPES” não se limitou a uma base de dados específica como a *Web of Sciences*, a escolha aconteceu a partir da seleção de áreas, como as Ciências Sociais Aplicadas.

2.1 Aquisição de Conhecimento: uma visão geral

A AC, mesmo que sob diversas denominações, é identificada pela preocupação comum em diferentes campos científicos – Gestão do Conhecimento, Ciência da Computação, Ciência da Informação e Biblioteconomia, para citar alguns – em apreender conhecimento especializado para diferentes fins, no qual inclui o de representação.

A expressão AC foi empregada pela primeira vez na década de 1980 pela Engenharia do Conhecimento - área da Inteligência Artificial que se preocupa com princípios, métodos e ferramentas para aquisição de conhecimento e desenvolvimento de sistemas especialistas (BOOSE; GAINES, 1989; MILTON *et al*, 2006; MILTON, 2007). Boose e Gaines (1989) apresentam um conjunto de pesquisas relacionadas ao tema. Para esses autores, os primeiros encontros da área surgiram em 1986, em decorrência da necessidade de se discutir a temática, onde a literatura até então era esparsa, dispersa em diversas publicações e conferências, e de difícil acesso⁴.

Nos anos de 1990, em Gestão de Conhecimento, o conhecimento passa a ser visto como recurso, como inovação, como *commodity* e, portanto, tornou-se essencial às empresas desenvolver habilidades para adquirir, transmitir e materializá-lo em ações que resultem em produtos e serviços estratégicos (DAVENPORT; PRUSAK, 1998; NONAKA, 1994; NONAKA; TAKEUCHI, 199; CHOO, 2003). Sob essa perspectiva, a atividade de AC passou a integrar um conjunto de estratégias pioneiras que visavam apreender o conhecimento individual e convertê-lo em organizacional (CHOO, 2003).

No espaço organizacional, Almeida (2006) apresenta a AC como um processo em que recursos potenciais do conhecimento organizacional são identificados, capturados e disponibilizados, para um estágio seguinte, denominado organização dos

⁴A primeira publicação dedicada ao assunto: *First AAAI-Sponsored Knowledge Acquisition for Knowledge-Based Systems Workshop* (KAW86), Banff, Canada.

recursos. No contexto das Ontologias, a AC surgiu sob o enfoque da Engenharia de Ontologias (FERNANDEZ; GOMEZ-PEREZ; JURISTO, 1997), numa alusão ao processo de construção da ontologia, como sendo correspondente aos de sistemas baseados em conhecimento. Para estes autores, a AC é uma etapa do processo de representação.

Por sua vez, em Ciência da Informação e Biblioteconomia, a atividade de AC tem lugar de discussão na construção de vocabulários controlados para representar a informação e o conhecimento em sistemas de recuperação da informação, na avaliação de sistemas de navegação da internet, bem como em ambientes de pesquisa em que é necessária a identificação de conteúdos (NATIONAL, 2005). Considera-se ainda que AC acontece também da interação do bibliotecário com o usuário ou especialista para endosso sobre terminologia, como forma de prospectar novos termos e relacionamentos, bem como de confirmar o uso efetivo dos mesmos (LANCASTER, 1987).

Como regra geral, os campos envolvidos com a tarefa de organizar conhecimento consideram que, de modo geral, a AC inclui: coleta, análise, estruturação e validação do conhecimento com a finalidade de representação (GAINES, 2003; MILTON, 2007; PAYNE *et al*, 2007; HUA, 2008). Trata-se de atividade composta por um conjunto de tarefas que empregam técnicas manuais e baseadas em computador. Essas últimas se dividem em técnicas semiautomáticas, conhecidas também como interativas, e em automatizadas. Em ambos os casos, essa tarefa visa obter conhecimento, que pode acontecer desde a interação direta com o especialista do domínio até as fontes documentadas (BOOSE; GAINES, 1989; DELUGACH; WOLF, 1996; BURGE, 1998; SHADBOLT, 2005; GAINES, 2003; MILTON *et al* 2006).

Diversas definições para AC são encontradas. “[...] é a elicitación e captura do conhecimento em um formato estruturado” (SHADBOLT; BURTON, 1990); “processo de modelagem em que o conhecimento elicitado será um resultado do processo, e não o conhecimento já existente” (SHAW; GAINES, 1996); processo de “levar pessoas a descrever como elas fazem o que fazem” (SCOTT *et al*, 1991); “[...] é o processo de extrair, estruturar e organizar o conhecimento de uma ou mais fontes” (TURBAN, 1992); “[...] “é o processo total de apreender o conhecimento e transferi-lo” (REGOCZEI, 1994); “[...] é o processo de mapear as declarações de especialistas em modelos conceituais” (SHADBOLT *et al* , 2000).

As teorias e métodos que oferecem suporte à AC são obtidos em disciplinas acadêmicas diversas, bem como da prática (Payne *et al*. 2007). Formas de adquirir,

representar e verificar ou validar conhecimento são provenientes de áreas como: Ciência da Computação (NEWELL; SIMON, 1975), Ciência Cognitiva (HAWKINS, 1983), Linguística (ZELLIG HARRIS, 1976), Semiótica (CAMPBELL, 1998) e Psicologia (KELLY, 1955).

Em *Ciência da Computação*, os trabalhos de Newell e Simon (1975) e de Compton e Jansen (1989) são alguns exemplos. Sob esta abordagem, a capacidade de adquirir e representar o conhecimento em um formato computável é obtida através da hipótese de símbolos físicos. Essa hipótese postula que o conhecimento é constituído por símbolos que representam a realidade e pelas relações entre estes símbolos. Nesse contexto, a inteligência corresponde à habilidade de manipular logicamente os símbolos e relacionamentos entre eles.

Na *Psicologia*, as bases da atividade de AC se encontram em trabalhos seminais como a Teoria dos Construtos Pessoais de Kelly (1955). Essas bases assumem hipoteticamente que os seres humanos transferem sua *expertise* para sistemas computacionais, de forma que esses sistemas sejam capazes de replicar o desempenho humano do especialista.

A *Ciência Cognitiva*, representada pelo trabalho de Hawkins (1983), procura identificar caminhos de transferência de *expertise* como meio existente para a eliciação do conhecimento de especialistas de domínio. Esse modelo é ilustrado como o *modelo Hawkins* para a transferência de conhecimento cliente-especialista. As teorias da Ciência Cognitiva concentram-se na diferenciação entre tipos de conhecimento.

A Semiótica apresenta a perspectiva da tríade de Ogden-Richards (1923) e citada, por Campbell (1998), como uma base teórica para AC. Tal teoria consiste no “estudo dos signos”, de como o significado desses signos é transmitido e compreendido individualmente ou por um grupo de pessoas. A tríade semiótica é representada por um triângulo em que os vértices são: i) *Símbolos*: uma especialização de signos, de fato são convenções para representar um objeto ou entidade; ii) *referentes*: os próprios objetos ou entidades da realidade e; iii) *referências*: representações do entendimento do agente que utiliza o conhecimento. Payne *et al* (2007), entretanto, apontam deficiências de aplicação da semiótica em AC: carência de pesquisas empíricas demonstrando a eficácia das relações semióticas como forma de avaliação de conhecimento, além da necessidade de métricas para medir a consistência das relações entre os elementos da tríade semiótica.

No campo da *Linguística*, existem iniciativas para extração de conhecimento através da análise de fontes de coleções de textos. A análise de linguagem subjacente é uma das abordagens citadas na literatura para AC a partir de textos. Tal base teórica foi descrita por Zellig Harris (1976) em seu trabalho sobre a natureza do uso da linguagem em domínios altamente especializados, caracterizados pela estrutura gramatical regular e reprodutível. Esses recursos gramaticais podem ser descobertos através da aplicação de padrões de reconhecimento manuais ou automatizados para grande *corpus* linguístico de um domínio.

Cabe ainda citar que os estudos sobre AC fazem referência à existência de diferentes tipos de conhecimento em um domínio, conforme sugerem Berry e Broadbent (1984 *apud* MASTELLA, 2005). Apresenta-se a seguir algumas possíveis classificações elencadas por Gandon (2002), Almeida (2006), Mastella (2004), Milton (2007), Payne *et al*, (2007). Está além dos objetivos da presente pesquisa discorrer sobre cada tipo de conhecimento apresentado.

Almeida (2006) apresenta um levantamento de diversos tipos de conhecimento organizacional, a saber: conhecimento formal e informal (EUZENAT, 1996); conhecimento nas perspectivas cognitivista e construtivista (VON KROGH; ROOS, 1995); conhecimento tácito e explícito (NONAKA; TAKAEUCHI, 1997); *know-how* (MAHE; RIEU; BEICHEN, 1996; SIMON, 1996; KUNH; ABECKER, 1997); conhecimento declarativo e procedural (GANDON, 2002); conhecimento distribuído e centralizado (BROWN; DUGUID, 2001); conhecimento descritivo, dedutivo e documentário (POMIAN, 1996); conhecimento técnico e para gestão (GRUNSTEIN; BATHES, 1996); conhecimento tangível e intangível (GRUNSTEIN; BATHES, 1996); conhecimento descritivo, metodológico e comportamental (EUZENAT, 1999); conhecimento sugestivo, preditivo, metodológico, decisivo e sistêmico (SØRLI *et al*, 1999); conhecimento sugestivo, preditivo, decisivo e sistêmico, (STEIN, 1995); conhecimento tácito e focal (SVEIBY, 1997); conhecimento *hard e soft* (KIMBLE; HILDRETH; WRIGHT, 2001).

No contexto da Engenharia do Conhecimento, Mastella (2005) apresenta trabalhos de Harmon e Sawyer (1990), de Turban (1992), dentre outros, a saber: conhecimento superficial e profundo do domínio (HARMON; SAWYER, 1990; TURBAN, 1992); conhecimento declarativo, procedimental, semântico, episódico e meta-conhecimento (TURBAN, 1992); conhecimento tácito e conhecimento focal (POLANYI, 1974); conhecimento implícito e explícito (NONAKA; TAKAEUCHI,

1995); conhecimento compartilhado, conceitual, operacional e sistêmico. Por sua vez, Milton (2007), faz menção a: conhecimento procedural; conhecimento conceitual; conhecimento básico e explícito; conhecimento profundo e tácito.

Cada um dos campos apresentados tem contribuído para a formação, a compreensão e a evolução da atividade de AC e seu desdobramento em contextos distintos. Somando-se a esses estudos, é relevante apresentar como tal atividade vem sendo abordada pela Ciência da Informação, por se tratar da área com a qual a presente pesquisa se desdobra e a na biomedicina, por ser a área de enfoque do estudo.

Vale ressaltar que, no período em que foi feita a busca nas bases de dados mencionadas, não foi permitida, para a devida análise, a obtenção do arquivo completo de todos os trabalhos recuperados. Além disso, não foram considerados trabalhos cujos propósitos fossem diferentes ao da presente pesquisa, como, por exemplo, emprego do termo elicitación ou aquisição de conhecimento com diferentes finalidades. Procurou-se ainda, por trabalhos empíricos que fizeram uso de técnicas para elicitar conhecimento de especialistas. Desse modo, a seção 2.2 discorre sobre AC em Ciência da Informação e a Seção 2.3 aborda o assunto no contexto da biomedicina, ambas apresentam os trabalhos identificados na literatura que exploram a temática.

2.2 Aquisição de conhecimento em Ciência da Informação

Na Ciência da Informação, alguns trabalhos que abordam Aquisição de conhecimento ou elicitación de conhecimento o fazem sob diferentes enfoques e para fins diversos, conforme estratégias apresentadas em iniciativas de pesquisa nos seguintes campos: *Recuperação da Informação* (SPINK; SOLLENBERGER, 2004); *Classificação* (TANGMANEE, 1994; FIDEL, 1994); construção de *Tesauros* (KWASNIK, 1994), *Ontologias* (LEENHEER, 2008); *Estudo de Usuários* (PISANSKI; ZUMER 2010); *Gestão do Conhecimento* (GARCIA-PERES, 2007; JACKSON; WEBSTER, 2007; COFFEY, 2008; PITT, 2008; BEERS, 2009; KWONG; LEE, 2009; EL HARBI 2011), *Bibliotecas e Serviços de Informação* (MEI-MEI, 2003); *Representação do Conhecimento* (PREISS, 2000; JOHN, 2003).

Spink e Sollenberger (2004) apresentam um estudo empírico de elicitación relacionado às tarefas de busca durante a Recuperação da Informação (RI), que envolve: i) tarefa de modelagem de problemas com informação durante o estágio de busca pré-

online e online; ii) tarefa do Sistema de Recuperação da Informação (SRI), durante a busca online; iii) tarefa interativa para interpretar os resultados de busca; iv) tarefa social, em que pessoas que buscam informações e intermediários de busca, muitas vezes estão envolvidos em um diálogo cujos assuntos não estão diretamente relacionados à conduta de busca on-line.

Esses autores utilizaram um corpus de transcrição de áudio relativo à interação de 30 usuários de determinada universidade, com três profissionais treinados para busca intermediária. Os profissionais procediam a uma entrevista e, conforme o assunto da área, realizavam uma busca na base de dados online DIALOG. Spink e Sollenberger (2004) adotaram a elicitación como unidade de análise, na qual a definem como qualquer enunciado de uma pessoa A, que provoca diretamente (elicita) uma resposta de uma pessoa B, e para a qual foi a intenção do elicitante fazê-la. Ainda completam dizendo:

[...] “para se julgar a verdadeira intenção da elicitación, fortes evidências de tal intenção devem estar presentes no discurso transcrito, ou seja, um enunciado tem que gerar uma resposta do elicitado para ser contabilizado como uma elicitación” (SPINK; SOLLENBERGER, 2004)

Foram levantadas a elicitación de conhecimento dos profissionais e dos usuários e os propósitos da elicitación, por meio de um processo interativo de análise de buscas nas 30 transcrições.

O trabalho proposto por Spink e Sollenberger (2004) discute modelos das tarefas relacionados a RI, na qual constataram que a análise de elicitación fornece um modelo baseado em uma melhor compreensão dos usuários finais; das pessoas que buscam informação; as necessidades de informação dos intermediários de busca e dos propósitos durante uma interação de RI. Esses achados, de acordo com os autores, sugerem aspectos não esclarecidos com relação àqueles que buscam informação durante o processo de interação na recuperação da informação.

Pisanski e Zumer (2010) apresentam um estudo relacionado à aquisição de conhecimento, sob o enfoque de modelos mentais de usuários no universo bibliográfico, no qual verificam se os requisitos funcionais para registros bibliográficos (FRBR - *functional requirements for bibliographic records*) se comparam ao modelo conceitual do universo bibliográfico. Requisitos funcionais “são um modelo conceitual do universo bibliográfico” (PISANSKI; ZUMER, 2010, p. 668). Para o desenvolvimento deste

trabalho, os autores utilizaram algumas técnicas de elicitação do conhecimento, a saber: i) classificação de cartões (*card sorting*): elicitar modelos mentais da estrutura conceitual do universo bibliográfico; ii) mapeamento conceitual na extração de modelos mentais estruturados, embora a situação ainda fosse teórica; iii) entrevista em que os participantes classificaram os pares de itens de acordo com o FRBR.

Na primeira parte do trabalho de Pisanski e Zumer (2010) foram realizadas entrevistas, onde 30 usuários foram solicitados a descrever cada um dos itens dos 11 pares de publicações levantados pelos autores e afirmar se tais pares de itens seriam similares e se substituíveis. Nesse estudo, Pisanski e Zumer (2010) buscavam identificar os elementos dos livros percebidos pelos usuários, além de verificar se estes usuários estavam inclinados a descrever atributos diferentes daqueles das entidades FRBR e se prestavam atenção em todos os detalhes.

Segundo Pisanski e Zumer (2010), um dos maiores trunfos deste estudo é a riqueza e a variedade dos dados obtidos. Entretanto, consideram que ainda há espaço para uma maior análise, sugerindo, para tal, mais atenção aos tipos de cartões individuais e uma análise mais aprofundada das entrevistas. Além disso, propõem uma comparação dos modelos mentais elicitados de cada um dos participantes, em cada uma das tarefas, de modo a entender melhor se determinadas pessoas têm modelos mentais mais consistentemente com relação ao FRBR do que outras.

No trabalho de Kwong e Lee (2009), a aquisição de conhecimento é discutida no contexto da gestão do conhecimento, no qual propõem elicitar o conhecimento sobre gerenciamento de confiabilidade a partir de especialistas. Para tal, desenvolvem um método, cuja abordagem enfoca elicitar conhecimento tácito dos engenheiros de confiabilidade, por meio de narrativas e mapeamento cognitivo. De acordo com estes autores, os mapas cognitivos dos engenheiros revelaram certos padrões, que, agregados a um mapa cognitivo de todos os participantes, ajudam a resumir várias abordagens e procedimentos que poderiam ser tomados para lidar com diferentes problemas de gestão de confiabilidade.

A elicitação de conhecimento, através da narrativa, compreendeu um conjunto de técnicas e métodos para extrair conhecimento dos especialistas do domínio, utilizando para tal algum tipo de interação. [...] “tradicionalmente, as tarefas envolvem encontrar pelo menos um especialista do domínio em questão, em que são empregadas entrevistas em profundidade” (KWONG; LEE, 2009). A narrativa foi apresentada pelos autores como sendo um relato de uma sequência de eventos. Uma história que descreve

situações e personagens, de acordo com um cronograma, em um contexto de ficção ou não. Por sua vez, os mapas cognitivos foram aplicados pelos autores durante as entrevistas, para visualização dos modelos mentais dos entrevistados. Foram usados como um método de anotações, empregados alternativamente para registrar as transcrições das entrevistas ou dos outros dados documentais.

Segundo Kwong e Lee (2009), o trabalho fornece um exemplo real para apoiar as fases de aprendizado do indivíduo e de um grupo no nível organizacional. De acordo com estes autores, é o primeiro estudo no setor aéreo que se propôs capturar o *know-how* e a experiência dos engenheiros de confiabilidade, reunindo-os em mapas cognitivos, de modo a facilitar o aprendizado em equipe e na construção da memória organizacional. O modelo foi aplicado na Divisão de Engenharia de uma companhia aérea, que visava lidar com suas questões de gestão de confiabilidade.

Também no contexto da gestão de conhecimento, Jackson e Webster (2007) apresentam um trabalho que enfoca elicitación de conhecimento dos formuladores de políticas de uma agência do governo australiano. A proposta do trabalho era elicitar conhecimento para a criação de um inventário dos tipos de conhecimento necessários para se desenvolver políticas complexas para o gerenciamento dos recursos naturais. De acordo com os autores, este inventário foi usado para investigar soluções de gestão de conhecimento para lidar com a perda de informações, desgaste de pessoal e eficácia operacional. Para tal, desenvolveram uma metodologia que se propunha combinar metodologia de sistemas flexíveis, mapeamento cognitivo causal e *brainstorming* para fornecer um conjunto de requisitos de conhecimento.

O conhecimento do negócio, de acordo com Jackson e Webster (2007), foi feito através da análise de negócios e mapeamento de processos de negócios e visava identificar as atividades de trabalho, bem como os fluxos de informação, os papéis e responsabilidades. No estudo foram aplicados o mapeamento de processos de negócios e a metodologia de sistemas flexíveis. Em seguida, procedeu-se à construção de um mapa cognitivo que tinha por finalidade representar o conhecimento das pessoas envolvidas no processo.

Após a conclusão do mapa cognitivo, os integrantes participaram de uma oficina de modo que pudessem discutir os elementos identificados como pré-requisitos para o trabalho e anotar aqueles considerados problemas ou obstáculos. Após a informação ter sido provocada por parte do pessoal, ela foi transferida para um diagrama de classes de

alto nível da UML. O resultado do trabalho foi um mapa de conhecimento de alto nível para a organização.

Segundo Jackson e Webster (2007), o componente substancial da pesquisa foi o método de elicitação do conhecimento e da ontologia resultante, que contém elementos de contextualização social e raciocínio para a ação organizacional e de decisões. A ontologia básica fornece a base para identificar as necessidades de conhecimento, armazenada em um repositório de informação, que poderia ser a estrutura de uma taxonomia organizacional e um conjunto detalhado de questões que podem ser abordadas por decisão de gestão.

Na perspectiva das ontologias, o trabalho de Almeida (2006) apresenta a atividade de AC como parte do processo de construção de ontologias, em que são empregadas técnicas como *brainstorming*, entrevistas não estruturadas e estruturadas, análise informal e formal de documentos.

No referido trabalho, tais técnicas tinham as seguintes funções, a saber: i) entrevistas não estruturadas com especialistas, que tinham por objetivo obter termos candidatos a conceitos do modelo para a memória organizacional; ii) análise de documentos, em que foi empregada a técnica de Análise de Assunto, para análise dos documentos como documentos internos, documentos externos, documentos auxiliares, registros, anotações, entre outros, manipulados pelos funcionários para desempenho de suas atividades; iii) análise de cenários; iv) técnica JAD - *Joint Application Design*, com vistas a obter consenso e interatividade, uma vez que tal técnica possibilita discussões em grupo, onde as decisões são baseadas em consenso; v) análise de cenários, pela facilidade de obtenção de dados sobre as necessidades e atividades dos funcionários das empresas, através de narrativas (ALMEIDA, 2006).

Tanto a técnica JAD quanto a análise de cenários, de acordo com Almeida (2006), são provenientes das áreas correlatas ao desenvolvimento de sistemas. Por sua vez, a análise de assunto é um método reconhecidamente válido e amplamente utilizado na Ciência da Informação.

2.3 Aquisição de conhecimento no contexto biomédico

A organização de terminologias na área biomédica é um desafio em função da amplitude do assunto e da multiplicidade de interpretações existentes para os seus dados. Desse modo, a presente seção apresenta uma visão geral das práticas adotadas de AC na biomedicina. A literatura científica sobre AC aplicada à biomedicina apresenta diferentes propostas para representação do conhecimento médico e biológico, em que são empregados tipos, técnicas e abordagens variadas para AC (TSUMOTO, 1998; VITA *et al*, 2006; PAYNE *et al* 2007; HOEHNDORF *et al*, 2009; STEHR *et al*, 2009; VAN AUKEN *et al*, 2009).

Payne *et al*. (2007) apresentam iniciativas de AC aplicadas à biomedicina, considerando que a atividade se refere ao processo de gerar conhecimento para uso em bases de conhecimento integrantes de sistemas especialistas (por exemplo, sistemas de apoio à decisão clínica). Entretanto, se apóiam no trabalho de Payne e Starren (2005), para argumentar que a definição do AC no domínio da informática biomédica deveria ser ampliada para englobar modelos de banco de dados compartilhados, ontologias e modelos de interação homem-máquina. O QUADRO1 a seguir representa o conjunto de iniciativas sugeridas por Payne *et al* (2007) para abordar AC no contexto da Biomedicina.

QUADRO 1: Base teórica para AC aplicada à Biomedicina

Tópico	Autor(es)	Ano	Título	Fonte
Definição de tipos de conhecimento	McCormick	1997	<i>Conceptual and Procedural Knowledge</i>	Educação
	Patel, Arocha	2001	<i>A primer on aspects of cognition for medical informatics</i>	Ciência Cognitiva
	Zhang	2002	<i>Representations of Health concepts: a cognitive perspective</i>	Biomedicina
Framework da Engenharia do Conhecimento	Liou	1990	<i>Knowledge acquisition: issues, techniques, and methodology</i>	Ciência da Computação
	Menzies e Harmelen	1999	<i>Evaluating knowledge engineering techniques</i>	Ciência da Computação
Representação computacional do conhecimento	Newell e Simon	1981	<i>Computer science as empirical inquiry: symbols and search In Mind Desing</i>	Ciência da Computação
	Compton eJansen	1990	<i>A philosophical basis for knowledge</i>	Psicologia
	Hereth, Stumme	2000	<i>Conceptual Knowledge Discovery and Data Analysis</i>	Ciência da Computação

Framework da AC	Brachman e McGuinness	1988	<i>Knowledge representation, connectionism and conceptual retrieval</i>	Ciência da Computação
	Gaines e Shaw	1989	<i>Social and Cognitive Processes in Knowledge Acquisition</i>	Ciência Cognitiva
	Compton e Jansen	1990	<i>A philosophical basis for knowledge acquisition</i>	Psicologia
	Liou	1990	<i>Knowledge Acquisition: issues, techniques, and methodology</i>	Ciência da Computação
Base filosófica para AC	Kelly	1955	<i>The psychology of personal constructs</i>	Psicologia
Base semiótica para AC	Campbell, Oliver	1998	<i>Representations of health concepts: a cognitive perspective</i>	Biomedicina
Métodos de AC conceitual	Liou	1990	<i>Knowledge acquisition: issues, techniques, and methodology</i>	Ciência da Computação
Métodos de validação de acervo de conhecimento	Preece	2001	<i>Evaluating Verification and Validation Methods in Knowledge Engineering</i>	Ciência da Computação

Fonte: adaptado de Payne *et al*, 2007

Do ponto de vista dos métodos de conhecimento aplicados à biomedicina, Payne *et al* (2007) apresentam uma taxonomia agrupada em três categorias para AC: *unidade de elicitação do conhecimento*: refere-se às técnicas de elicitação de unidades de conhecimento; *elicitação de relacionamento do conhecimento*: diz respeito a técnicas para elicitação de relacionamento entre as unidades de conhecimento; *elicitação combinada*: para técnicas que provocam tanto unidades de conhecimento e as relações existentes entre eles. O QUADRO 2 mostra os tipos de técnicas para cada categoria, conforme sugerido pelos autores.

QUADRO 2: Métodos de AC conceitual

Unidades de elicitação de de conhecimento	Elicitação de Relacionamento de conhecimento	Elicitação combinada
Entrevista	Classificação categórica	Análise de Protocolo
Entrevista Estruturada	Análise de Grades de Repertório	Análise de Discurso
Observação	Análise do conceito formal	Análise de sub-linguagens
		<i>Laddering</i>
		Técnicas de Grupo

Fonte: Adaptado de Payne *et al*, 2007

Outra abordagem apresentada para AC no âmbito da biomedicina é apresentada por Tsumoto (1998). Trata-se de um método denominado *método de indução de regras*

para AC automática em base de dados clínicos. Esse método visa ao desenvolvimento de um sistema especialista capaz de fazer diferentes diagnósticos para doenças congênitas. Segundo o autor, o desempenho de regras induzidas é relatado desde o início da década de 1980. Além disso, o método de indução de regras é proposto para AC automática em bases de dados para resolver gargalos da AC.

Segundo Tsumoto (1998), o método consiste de quatro passos. O *primeiro passo* enfoca dois tipos de raciocínio médico: raciocínio positivo e negativo e em um *framework* de conjuntos aproximados são definidas duas regras correspondentes, ou seja, regras positivas e regras negativas. Além disso, foram introduzidas generalizações orientadas a atributos, que convertem os atributos de um banco de dados para generalizá-los como conhecimento de domínio. No segundo passo foi desenvolvido um sistema de indução de regras, denominado *Probabilistic Rule Induction Method based on Rough Sets for Rules of Expert System*, o qual foi aplicado a uma base de dados clínicos de malformações congênitas, a fim de obter conhecimento para diagnóstico. No terceiro passo, foi desenvolvido um sistema especialista chamado *Rule based system for Congenital Malformation*, que gera diagnóstico a partir da entrada de observações feitas por usuários. Por fim, o sistema foi avaliado na prática clínica e os resultados indicaram uma performance aproximada ao do especialista médico

No âmbito na biomedicina, devido à enorme quantidade de conhecimento registrado em diversas fontes, uma forma difundida e empregada para AC é a curadoria. Segundo Nenadic *et al* (2002), a curadoria faz parte do processo de anotação, por exemplo, de grandes projetos de sequenciamento de genes, dado o grande volume de informações, e geralmente feito de forma automatizada. Contudo, anotações mais confiáveis ainda requerem uma análise manual, feita pelo especialista da área, em cima dos dados gerados automaticamente. Anotações que foram inspecionadas e editadas manualmente são ditas curadas, pois sofrem uma curadoria.

Nenadic *et al* (2002) descrevem um sistema de gestão de anotações denominado *Tag Information Management System* (TIMS), um sistema integrado de gestão de conhecimento no domínio da biologia molecular e da biomedicina, em que a mineração da literatura é orientada pela terminologia, pela AC, pela integração do conhecimento e recuperação da informação baseada em XML, os quais são combinados usando-se a gestão de anotação e ontologias de inferência. Trata-se de um sistema para mineração de literatura em grandes fontes de conhecimento.

Segundo Nenadic *et al* (2002), essa proposta surgiu em decorrência da enorme quantidade de *papers* publicados e pela dificuldade das pessoas em localizar a informação de interesse. Para os autores, o principal objetivo da mineração de textos é recuperar o conhecimento contido em um texto e apresentá-lo aos usuários. A vantagem, em comparação com os processos manuais, é baseada na suposição de que os métodos automáticos são capazes de processar grande quantidade de textos. Por estas razões, a mineração de textos auxilia cientistas na obtenção, manutenção, interpretação e curadoria da informação.

O processo de AC, baseado na extração da Informação no TIMS, tem o seguinte curso: em primeiro lugar, uma coleção de documentos é linguisticamente processada e terminologicamente analisada, ou seja, termos relevantes específicos do domínio são automaticamente reconhecidos e estruturados (classificados ou incorporados em uma ontologia). Na etapa seguinte, um usuário (em geral um especialista em biomedicina) formula uma consulta, a qual é executada na coleção. Os resultados são apresentados em uma tabela e o texto relevante é realçado. Essas atividades encontram-se descritas na FIG. 1 a seguir:

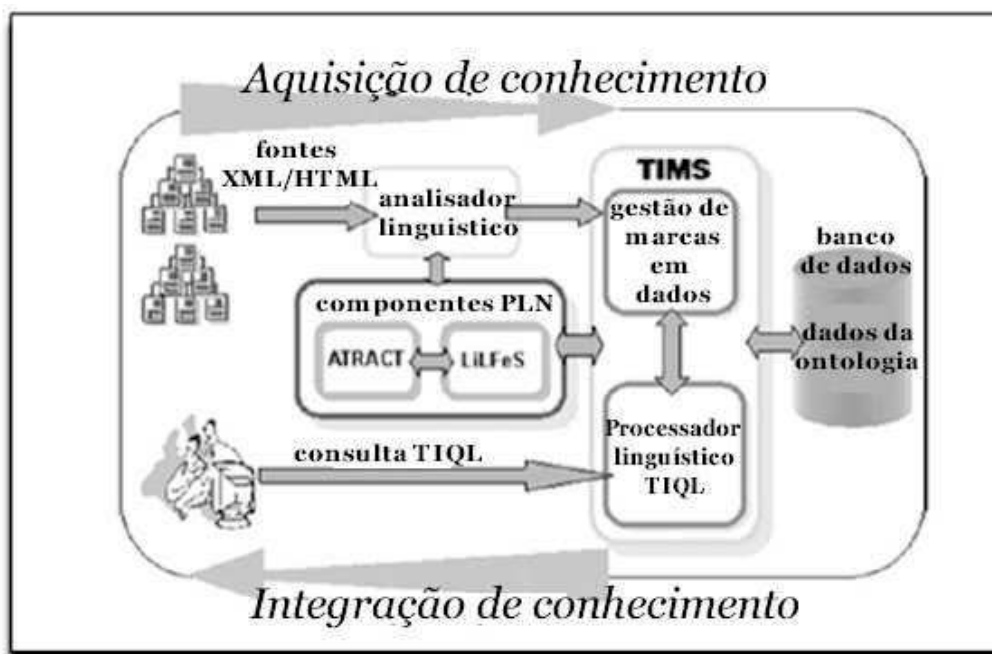


FIGURA 1: AC baseada na extração da Informação no TIMS
 Fonte: Adaptado de Nenadic *et al* (2002)

O TIMS permite que os especialistas do domínio pesquisem e combinem informações de várias fontes. A extração da informação dentro do sistema é orientada à

terminologia, em que as informações são fornecidas automaticamente no formato XML. A recuperação baseada em anotações é implementada através de operações, em combinação com um algoritmo de casamento de padrão hierárquico baseado em ontologias, o qual oferece formas de executar a AC através da mineração literatura.

Por sua vez, Vita *et al* (2006) apresentam um trabalho voltado para extração de dados imunológicos dependentes de contexto, que exige elevado nível de curadoria especializada. Uma ontologia formal específica⁵ fornece a plataforma sobre a qual se baseia a curadoria. A curadoria permite o processamento de grande volume de dados para inserção em bases de conhecimento disponíveis para a comunidade científica. O primeiro passo consiste em selecionar artigos a partir de pesquisa no PubMed⁶. A consulta é composta de palavras-chave e operadores lógicos para selecionar referências relacionadas ao epitope⁷. Palavras-chave são usadas para selecionar as referências em combinação com uma palavra-chave filtro usada para limitar a seleção de referências por data de publicação, idioma e tipo de publicação, com exclusão de artigos de revisão, editoriais, meta-análise e comentários.

Na sequência, conforme sugerido por Vita *et al* (2006), a consulta é modificada com a entrada de especialistas no assunto, que analisam os resultados para garantir a inclusão de todas as referências relevantes. Após a seleção pela consulta automatizada, os resumos são formalmente analisados quanto à sua relevância para o escopo do *Immune Epitope Database and Analysis Resource* (IEDB), que é uma base de dados dedicada à captura, armazenamento e análise de dados complexos relacionados aos epítopes. O IEDB contém dados relacionados com epítomos de células T e de anticorpos para seres humanos, primatas, roedores e outras espécies animais.

O processo de seleção utiliza uma combinação da consulta baseada em computador e avaliação ativa feita por especialistas no assunto. Os resumos dos artigos são classificados de acordo com a possibilidade ou impossibilidade de aplicação de um processo de curadoria, ou seja, são classificados em “*curatable*” e “*uncuratable*” (VITA *et al*, 2006). Além disso, para um conjunto de dados, o processo de varredura de

⁵ Um relatório detalhado desta ontologia encontra-se disponível no endereço <http://www.ncbi.nlm.nih.gov/pubmed/16305755?dopt=AbstractPlus&holding=f1000,f1000m,isrcn>

⁶ PUBMED é um banco de dados de livre acesso à base de dados de citações, resumos e artigos em texto completo em ciências da vida e assuntos biomédicos.

⁷ É uma região localizada na superfície de um antígeno capaz de induzir uma resposta imune e de combinar com um anticorpo específico para combater essa resposta. Para mais informações: <http://www.immuneepitope.org/>

resumos e categorização é manual, mas são usadas ferramentas de classificação de texto.

De acordo com Vita *et al* (2006), a curadoria manual e a ontologia subjacente ao banco de dados fornecem orientações para a identificação de dados relevantes do epitope. A fim de reforçar estas orientações e maximizar a precisão da curadoria, o banco de dados contém um número de ferramentas de curadoria. Uma aplicação é o uso de *agentes de busca*⁸, que são amplamente utilizados em todo o IEDB. Agentes de busca traduzem a entrada do curador em um vocabulário controlado, como aqueles mantidos em bancos de dados externos sobre domínios relacionados. A FIG. 2 a seguir representa um modelo visual para os fluxos das atividades desenvolvidas pelo IEDB.

Segundo Vita *et al* (2006), o desafio do projeto está em capturar as características complexas dos epitopes imunológicos necessários para o desenvolvimento e implementação de um novo processo de curadoria de dados. Sabe-se, entretanto, que se trata de uma ação altamente dependente dos especialistas no assunto. Além disso, é necessária a tradução exata das informações contidas na literatura para o formato do banco de dados estruturado.

⁸ Do inglês “*finders*” (tradução nossa)



FIGURA 2: Processo de curadoria do IEDB
Fonte: Adaptado de Vita *et al* (2006)

Van Auken *et al* (2009) fazem referência a outro modelo de curadoria, numa abordagem para AC em textos, feito por meio de um *software* de mineração textual, para inserção no *Gene Ontology*⁹. De acordo com esses autores, as bases de conhecimento biológicas dependem fortemente da curadoria manual para extração e entrada de dados, mas ao mesmo tempo, a curadoria manual dos dados experimentais da literatura biomédica é um empreendimento caro e demorado. Desse modo, a curadoria automática é feita por um *software* denominado *Textpresso*¹⁰ e utilizado para dar suporte à curadoria manual.

Segundo Van Auken *et al.* (2009), o *Textpresso* é uma ferramenta de mineração de textos que funciona como um motor de busca simples, baseado em padrões, que

⁹ Disponível na Internet em: <<http://www.geneontology.org/GO.doc.shtml>>. Acesso em 22 de Julho de 2010.

¹⁰ Disponível na Internet em: <<http://www.textpresso.org>>. Acesso em 23 de Julho de 2010.

empregam categorias de palavras relacionadas conceitualmente às marcações semânticas para marcar o texto completo em documentos. Pesquisas no *Textpresso* podem ser feitas por palavras-chave ou categorias. Os resultados são apresentados numa lista de frases que contém termos correspondendo aos critérios da pesquisa, numericamente ordenados de acordo com o número de termos na sentença que correspondam à sequência de pesquisa. Exemplos de categorias incluem *doença*, *fenótipo* e *regulação* que contenham expressões como *Síndrome de Bardet-Biedl*, “*a transformação das células*” e “*downregulate*”, respectivamente. Existem 19 implementações do *Textpresso* em todo o mundo, incluindo neurociência, leveduras, *drosophila*, *arabidopsis*, e *coli*. Ainda existem mais de 100 categorias, tais como área do cérebro e doenças humanas, partes do corpo e os genes de *arabidopsis*, para ajudar na busca e extração.

Stehr *et al* (2010) abordam outra proposta de curadoria realizada por meio de projetos comunitários como a *Wikipedia*. De acordo com esses autores, existem atualmente várias propostas para as “*ciência-wikis*”, que visam à obtenção de conhecimento especializado em recursos centralizados. No entanto, argumentam que ainda não há consenso sobre como atingir esse objetivo. [...] “*não está claro qual a melhor forma de integrar os dados das estabelecidas e centralizadas bases de dados, com o fornecido pelas comunidades de anotações*” (STEHR *et al*, 2010, p. 2)¹¹.

Os autores apresentam a PDBWiki¹², uma *wiki* científica para a comunidade de anotação de estruturas de proteínas. A wiki é composta por uma página estruturada para cada entrada no *Protein Data Bank* (PDB)¹³, que permite aos usuários anexarem comentários classificados para as entradas. Cada página inclui uma lista de referências cruzadas a recursos externos editável pelo usuário. Assim, como em um banco de dados é possível produzir relatórios tabulares e estruturas baseadas em consultas definidas pelos usuários ou listas de entradas.

A PDBWiki é executada em paralelo com o PDB, separando-se o conteúdo original do banco de dados das anotações do usuário. A PDBWiki demonstra como os recursos de colaboração podem ser integrados aos dados primários de um banco de dados biológico. O sistema combina uma grande base estática de dados biológicos

¹¹ *It is not clear how to best integrate data from established, centralized databases with that provided by ‘community annotation’.*

¹² Disponível na Internet em: <http://www.pdbwiki.org>. Acesso em 20 de Setembro de 2010

¹³ Disponível na Internet em: <http://www.pdb.org/pdb/home/home.do> Acesso em: 24 de setembro 2010

heterogêneos com um sistema wiki para a anotação da comunidade. O banco de dados resultante serve como um recurso para eliminar erros e fórum de discussão para as estruturas no PDB.

O banco de dados é executado em paralelo ao repositório de arquivo central e incorpora os comentários do usuário em uma forma semi-estruturada, permitindo a possibilidade de incorporar as anotações de volta para a PDB. Os dados da PDB são automaticamente sincronizados com a PDBWiki, a cada semana. Em resumo, as novas entradas da PDB são adicionadas como novas páginas, entradas atualizadas são recriadas e entradas obsoletas são eliminadas. Comentários do usuário para entradas obsoletas permanecem no sistema e são automaticamente vinculadas à entrada atualizada da PDB quando aplicável.

A PDBWiki é centrada em torno das estruturas macromoleculares depositadas no PDB. Cada estrutura no PDB tem uma página separada na wiki, resultando em um total de mais de 60.000 páginas de estrutura. Uma página de estrutura é dividida em três seções principais: dados, comentários de usuários e links:

- *Dados:* contém informações básicas sobre a macromolécula, o que inclui o título, autor, data de inserção e sequência para cada cadeia única; as sequências são ligadas às suas entradas adequadas na *UniProt*¹⁴ ou *GenBank*¹⁵; os dados são obtidos diretamente do PDB e atualizados quando os dados subjacentes são alterados;
- *Comentários do usuário:* o usuário pode criar ou atualizar anotações usando a mesma sintaxe *wiki* ou usando um formulário. A abordagem semi-estruturada da PDBWiki permite comentários em texto livre, sem restrições de formato, e fornece um sistema de classificação de anotações hierárquica. Por exemplo, o usuário pode marcar um comentário com a categoria “erro de anotação na estrutura secundária”. Posteriormente, esse rótulo pode ser usado como um filtro para localizar ou excluir estruturas com esta anotação;
- *Links:* a terceira seção contém links para fontes externas de informações sobre a estrutura; os links são agrupados em quatro subseções: bancos de dados, ferramentas de visualização, ferramentas de análise e estrutura

¹⁴ Disponível na Internet em: <http://www.uniprot.org/> Acesso em 24 de Setembro de 2010

¹⁵ Disponível na Internet em: <http://www.ncbi.nlm.nih.gov/genbank/>

quaternária (para sites que fornecem dados sobre estrutura quaternária da proteína).

Sempre que possível, os links apontam diretamente para as páginas de dados relevantes ou resultados da análise para a estrutura vista. A estrutura quaternária foi considerada como importante o suficiente, para merecer sua própria seção na lista de recursos externos. A subseção da estrutura quaternária contém links para vários recursos com informações sobre a suposta unidade biológica, dando aos usuários acesso fácil a esse tipo de informação. A FIG. 3 a seguir mostra a estrutura da PDBWiki. Há uma estrutura de página para cada estrutura no *Protein Data Bank*, com três seções principais, dados (A), comentários de usuários (B) e links (C).

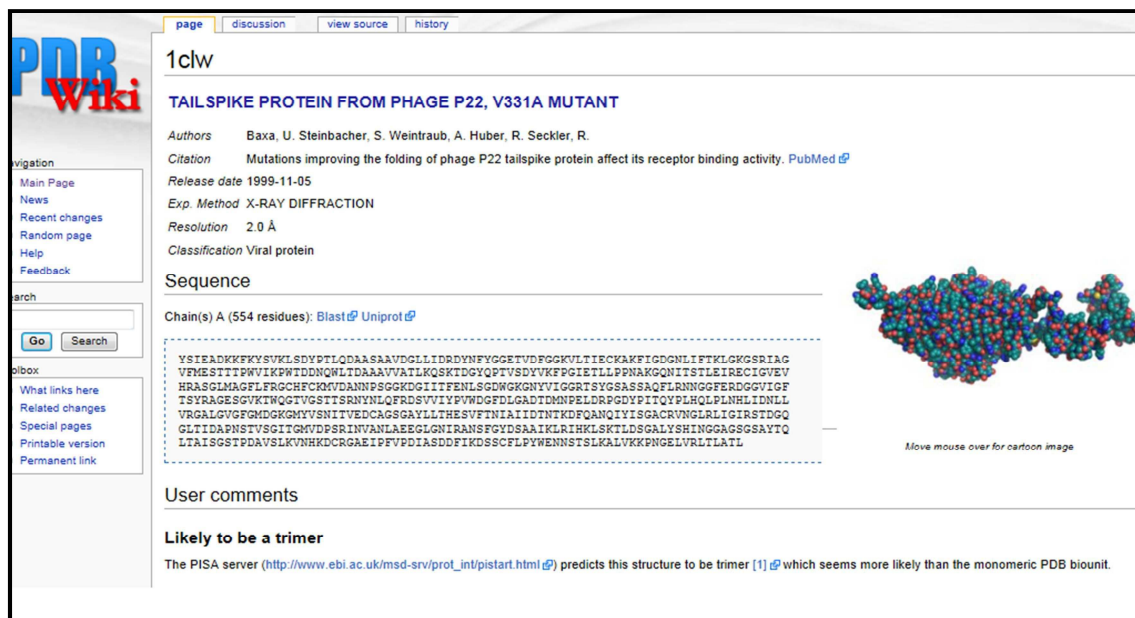


FIGURA 3: Estrutura da PDBWiki: *Protein Data Bank*
Fonte: Adaptado de Stehr *et al* (2010)

Por fim, Hoehndorf *et al* (2009) apresentam uma ferramenta *wiki*¹⁶ baseada em ontologias para *aquisição semi-automática* de conhecimento, desenvolvida para anotação de dados e integração de conhecimento biológico. Para os autores, ontologias biológicas são cada vez mais adotadas para anotação e classificação de dados biológicos, entretanto, tais atividades são um exercício demorado, que exige esforço de curadores. Além disso, o desenvolvimento e a manutenção dessas estruturas exigem criação, eliminação e correção manual de conceitos e definições da ontologia. Desse

¹⁶ Disponível na Internet em <<http://www.bowiki.net>>. Acesso em: 22 julho 2010.

modo, Hoehndorf *et al* (2009) acreditam que a AC baseada em sistemas *wiki* são viáveis para facilitar a aquisição colaborativa de conhecimento.

De acordo com Hoehndorf *et al* (2009), a AC baseada em *wiki* identifica muitas fontes de erro, como, por exemplo, uma entrada a uma base de conhecimento que não corresponde à realidade. A título de exemplo, os autores demonstram, por meio da declaração, que “todos os elefantes africanos têm cor púrpura” é factualmente errada. Esse erro pode ser detectado por uma pessoa que revê informações sobre elefantes africanos e sabe se este fato na base de conhecimento está incorreto. Esses autores argumentam que a detecção automatizada, desse tipo de problema, exige que grande parte do conhecimento biológico esteja formalizado e representado de forma que possa ser usado para inferências automatizadas. No entanto, é um resultado longe ainda de ser alcançado: “[...] *acreditamos que a detecção automatizada do conhecimento é de fato inviável*¹⁷ [...] *a detecção manual pode ser suportada pelo fornecimento de acesso fácil às inferências obtidas a partir das afirmações*¹⁸” (HOEHNDORF *et al*, 2009, p.2).

Outra fonte de erro, apresentada por Hoehndorf *et al* (2009), é o conflito no uso de ontologias por agentes que colaboram com a construção da base de conhecimento. Segundo esses autores, os termos em ontologias podem se referir a conceitos diferentes e, portanto, à realidade em diferentes perspectivas. Por exemplo, há pelo menos cinco conceitos distintos para se referir à membrana celular em subdomínios biológicos. Todos os conceitos são definidos de forma diferente e podem estar corretos em contextos apropriados, mas nenhum é preferível *a priori*. Nesse caso, uma ontologia formal pode ser usada para fazer tais distinções e, assim, restringir o significado do vocabulário àquele pretendido.

Apresentou-se até aqui enfoques diversos à AC, enfatizando a complexidade da atividade; os diferentes agentes envolvidos; a exigência por recursos intensivos para a sua consecução, bem como as dificuldades inerentes a AC, conforme exposição de alguns autores. Desse modo, torna-se relevante selecionar métodos apropriados de acordo com as características inerentes ao contexto. Na tentativa de organizar os métodos e os critérios para sua adoção, diversas tentativas de classificação para as

¹⁷ “*We believe that completely automated detection of factually inaccurate knowledge is unfeasible*”

¹⁸ “*Alternatively, manual detection can be supported by providing easy access to the inferences drawn from asserted knowledge*”.

técnicas de AC são encontradas na literatura. É fundamental selecionar métodos apropriados, bem como selecionar fontes do conhecimento adequadas. Fontes de conhecimento assumem muitas formas, incluindo textos, bases de dados e especialistas no domínio. Descreve-se a seguir as técnicas, ferramentas e métodos discutidos e empregados pelas diferentes áreas no contexto científico pesquisado.

2.4 Classificação para as técnicas de AC

A presente pesquisa explorou diferentes técnicas numa tentativa de identificar aquelas que pudessem se adequar ao contexto da AC a partir de especialistas para representação do conhecimento em ontologias. É relevante considerar que propõe identificar técnicas aplicáveis a cada uma das fases e que possibilitem interação com o especialista do domínio. Apresentam-se, a seguir, algumas propostas de classificação baseadas nos autores investigados, como forma de esclarecer a aplicação de cada uma. Não foi encontrado consenso sobre qual a melhor maneira de classificar técnicas de AC.

De modo geral, as classificações apresentadas a seguir foram definidas pelos autores pesquisados da seguinte maneira: i) sob o ponto de vista de aplicação da técnica (BOOSE; GAINES, 1989; BOOSE, 1990); ii) pelo tipo de interação: direta ou indireta (BURGE, 1998; MILTON, 2007); iii); em relação ao tipo de conhecimento obtido (SHADBOLT; SWALLOW, 1993; SHADBOLT, 2005); iv) sob o ponto de vista dos métodos (SHADBOLT; SWALLOW, 1993). A descrição de cada grupo será apresentada a seguir:

▪ Sob o ponto de vista da aplicação

Boose (1990) classifica como *técnicas manuais* e *técnicas baseadas em computador*. As primeiras possuem raiz na psicologia, principalmente, nos estudos de Kelly¹⁹. As chamadas técnicas baseadas em computador fazem uso de ferramentas semiautomáticas, também conhecidas como interativas, e ferramentas do processamento de linguagem natural, em geral identificadas pela expressão *aprendizado de máquina*. Nessa abordagem, um sistema comanda o processo de AC no domínio. Hua (2008),

¹⁹ *Psicologia dos Construtos Pessoais* é um sistema construtivista desenvolvido por George Kelly na obra KELLY, G. *Princípios da Psicologia dos Construtos Pessoais*. New York: Norton, 1955.

entretanto, sugere que, mesmo assim, para a obtenção de bons resultados existe ainda a necessidade de validar com especialistas os resultados obtidos automaticamente.

De acordo com Boose e Gaines (1989) e Boose (1990), as técnicas *manuais* são utilizadas a partir da interação do engenheiro do conhecimento com o especialista. Boa parte dessas técnicas tem raiz na Psicologia. Por sua vez, aquelas *baseadas em computador* se subdividem ainda em *automatizadas* e *interativas*. Nessa última, o especialista utiliza o computador com supervisão do engenheiro do conhecimento. Nas técnicas automatizadas, a exemplo da técnica de *aprendizado da máquina*, uma aplicação de computador detém todo o processo de obtenção de conhecimento do domínio (WOLF; DELUGACH, 1996; BOOSE; GAINES, 1989).

Wolf e Delugach (1996) explicam que nas primeiras abordagens automatizadas, baseadas em técnicas de aprendizado da máquina, a AC consistia de um conjunto de exemplos fornecidos a um programa que seria capaz de generalizá-los para extrair conhecimento sobre um dado fenômeno. Aplicavam-se a áreas relativamente simples e estruturadas, mas não em áreas complexas, a exemplo da linguagem natural. Segundo os autores, a rede neural é um tipo aplicado à aquisição de tais tipos de conhecimento.

Segundo Wolf e Delugach (1996), na aquisição interativa um sistema auxilia no processo de extração de conhecimento. Shadbolt (2005) lembra que uma série de sistemas foi desenvolvida para esse propósito, as quais são divididas em três categorias principais: i) ferramentas dependentes de domínio, desenvolvidas para domínios de tarefas específicas; ii) ferramentas independentes de domínio, que são implementações de uma técnica particular de elicitación de conhecimento; iii) integração dos sistemas de aquisição e suporte à elicitación.

De acordo com Milton *et al* (2006), os trabalhos realizados pela engenharia do conhecimento transformaram essas técnicas de AC tornando-as mais estruturadas, eficientes e sistemáticas. A título de exemplo, esses autores citam as ferramentas de modelagem como o Protegé, os editores de base de conhecimento como o CYC, além do PCPACK, KADS e MOKA, que surgiram desses trabalhos e são amplamente utilizadas por diferentes áreas. Vale ressaltar que algumas dessas ferramentas também são utilizadas para implementar métodos manuais de AC (MASTELLA, 2005).

Segundo Boose e Gaines (1989), as técnicas manuais apresentam algumas limitações: i) dificuldades de envolver uma amostra suficiente de especialistas para dar significância estatística; ii) dificuldade de se avaliar e validar o conhecimento adquirido dada a falta de padrões para comparação; iii) necessidade de descobrir uma variedade de

tarefas e domínios para que tais técnicas sejam úteis; iv) dificuldade de isolar qual valor agregado uma técnica oferece, quando as técnicas trabalham simultaneamente.

Por outro lado, segundo Mastella (2005), as ferramentas de AC automáticas e semiautomáticas apresentam desvantagens, sendo a principal delas a generalidade, ou seja, a possibilidade de aplicação a qualquer domínio. A aquisição resultante gera informações superficiais sobre o domínio, pois a ferramenta não se adapta conforme os construtos são processados e compreendidos. Além disso, segundo a autora, ferramentas automáticas só devem ser aplicadas quando existe uma fonte tratável pelo computador. Além disso, no caso da ferramenta semiautomática ser projetada especificamente para um domínio particular, é necessário efetuar um processo de AC prévio, de forma a identificar a melhor maneira de coletar informações dos usuários.

▪ **Pelo tipo de interação**

Burge (1998) e Milton (2007) consideram que as técnicas para AC se dividem em diretas ou indiretas, ou seja, conforme o tipo de interação com os especialistas. Para esses autores, as técnicas diretas são usadas para elicitare conhecimento por meio de questionamentos diretos ao especialista de modo que este possa expressar sua perícia. Por sua vez, as técnicas indiretas são usadas para viabilizar questionamentos indiretos, ou seja, em situações em que o conhecimento especializado não pode ser facilmente expressado pelo especialista ou encontra-se registrado em outras fontes.

Entrevistas estruturadas, não estruturadas ou semiestruturadas, estudo de caso, *teachback* são exemplos de técnicas diretas. Por sua vez, *card-sorting*, *laddering*, análise de documentos são tipos de técnicas indiretas. Tais técnicas também recebem a denominação de naturais ou elaboradas (*contrived*) e dizem respeito ao grau de familiaridade do especialista com a técnica (BURGE 1998; MASTELLA, 2005).

▪ **Em relação ao conhecimento obtido**

A suposição que a partir de diferentes métodos de elicitação derivam-se de diferentes tipos de conhecimento é conhecida na literatura científica como a *hipótese do acesso diferenciado*²⁰, uma denominação atribuída a Hoffman. A organização de

²⁰ No original, *Differential Access Hypothesis*

técnicas baseadas nessa hipótese tem o objetivo de facilitar a escolha da técnica apropriada conforme as características do domínio (HOFFMAN, 1992).

Para Hoffman (1992), assim como para Shadbolt e Swallow (1993), a hipótese do acesso diferenciado se baseia no fato de que especialistas possuem tipos diferentes de conhecimento e várias técnicas são necessárias para acessá-los. A FIG. 4 é um modelo visual para apresentar as técnicas de AC conforme os tipos de conhecimento. A descrição de tais técnicas é apresentada a seguir.

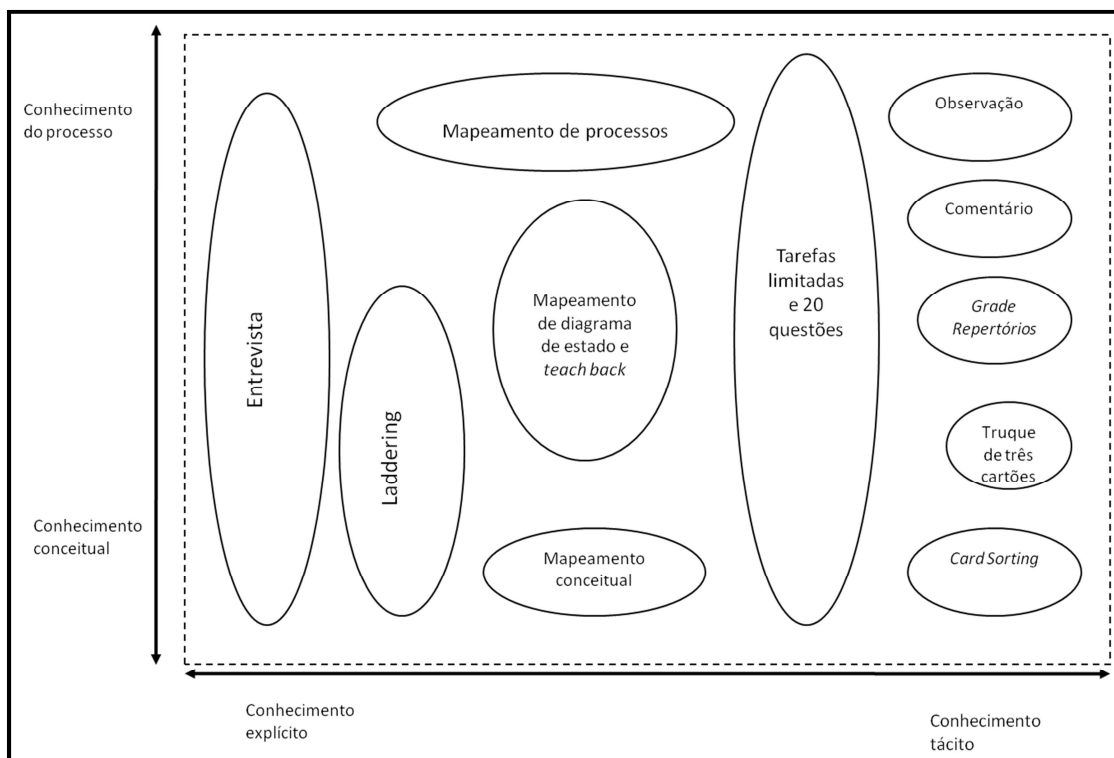


FIGURA 4: Técnicas de AC de acordo com o conhecimento elicitado.

Fonte: Adaptado de Shadbolt e Swallow (1993)

▪ Do ponto de vista dos métodos de aplicação

A classificação apresentada a seguir foi feita principalmente partir de Shadbolt e Swallow (1993), bem como de propostas complementares de outros autores. A utilização das técnicas pode variar de acordo com a dimensão do conhecimento que se objetiva elicitar.

- **Geração de protocolos**

A geração de protocolos corresponde a um conjunto de técnicas que inclui *entrevistas* (não estruturadas, semiestruturadas e estruturadas), *técnicas de comunicação*, como o *auto relato*, e de *técnicas de observação*. O objetivo dessas técnicas é a produção de um protocolo, ou seja, um registro de comportamentos que podem ser registrados em áudio, vídeo ou qualquer outra mídia eletrônica. A gravação em áudio é o método mais usual de transcrição (SHADBOLT; SWALLOW, 1993; SHADBOLT 2005; HUA, 2008).

De acordo com Delugach e Wolf (1996) e Almeida (2006), as entrevistas são um processo em que um engenheiro de conhecimento, formal ou informalmente, interroga um especialista sobre determinado assunto. A observação é um processo em que um observador determina a informação útil e observa uma tarefa ser completada pelo especialista. Para Burge (1998), o sucesso de uma entrevista depende das perguntas a serem feitas e da capacidade do especialista em articular aquilo que conhece.

Uma técnica de comunicação utilizada é conhecida por *teachback*, na qual engenheiro do conhecimento descreve uma parte do conhecimento adquirido em sessões anteriores ou de outras fontes. A partir daí, os especialistas ajudam o engenheiro descrevendo o que entendem sobre o assunto e esclarecendo mal-entendidos (SHADBOLT; SWALLOW, 1993).

- **Análise de Protocolo**

Segundo Shadbolt e Swallow (1993) e Shadbolt (2005), a análise de protocolo é usada, na transcrição de entrevistas ou de outros registros textuais, para analisar o registro gerado pelo especialista ou para identificar diferentes tipos de conhecimento. Atua também como uma ponte entre o uso das técnicas de protocolo e as técnicas de modelagem do conhecimento, além de envolver a identificação dos objetos do conhecimento no âmbito de um protocolo.

A análise de protocolo, segundo Delugach e Wolf (1996), envolve a inspeção de registros verbais de especialistas, em busca de informações sobre como eles resolveriam um determinado problema ou como eles definiriam certos conceitos.

- **Baseadas em matriz**

Segundo Shadbolt (2005) e Hua (2008), essa técnica envolve a construção de “*grades*”, onde são indicados, por exemplo, problemas e possíveis soluções. Nesse caso, matrizes são utilizadas para representar as propriedades de conceitos e o *repertory grid* para eliciar, calcular, analisar e categorizar as propriedades dos conceitos. Envolvem a construção e preenchimento de uma matriz de duas dimensões (grades e tabelas). Por exemplo, conceitos *vs* propriedades (atributos e valores), problemas *vs* soluções, hipóteses *vs* técnicas de diagnósticos, tarefas *vs* recursos, dentre outros.

Segundo Mastella (2004), essa técnica tem sua origem na Teoria dos Construtos Pessoais (*Personal Constructs Theory - PCT*), desenvolvida por George Kelly em 1955. Para adotar essa técnica, é necessário que o engenheiro do conhecimento já tenha identificado muitos dos principais componentes de conhecimento associados à tarefa. Os passos de aplicação são os seguintes:

1. Selecionar um conjunto de conceitos relacionados à tarefa;
2. Apresentá-los em grupos de três para o especialista;
3. O especialista deve escolher dois conceitos como ‘similares’ e um como o ‘diferente’ do grupo;
4. Após isso, ele deve indicar a razão pela qual diferenciou os três conceitos. Essa dimensão será utilizada como um construto;
5. O especialista estabelece uma escala bipolar de valores para esse construto (em geral, uma característica e seu oposto);
6. Os conceitos restantes são cotados em relação a esse construto nas características associadas a eles (MASTELLA, 2004, p.26).

A título de exemplo, apresenta-se a seleção de uma linguagem de programação (MASTELLA, 2004). O especialista seleciona as linguagens LISP, PROLOG, COBOL e C (1). O engenheiro pede que as três primeiras sejam analisadas (2). O especialista indica que LISP e PROLOG são similares (3) em relação à orientação (4) e indica que uma linguagem pode apresentar orientação simbólica ou numérica (5). As duas primeiras linguagens são simbólicas, enquanto COBOL é numérica. LISP e PROLOG são cotados com nota 3 e COBOL é cotado com nota 1.

Conforme exposto por Mastella (2004), a escala utilizada para dar notas aos conceitos é aplicada pelo especialista dependendo do objetivo da cotação. O construto

Tempo de Treinamento, por exemplo, recebe nota 1 se o tempo for alto e nota 3 caso contrário. Isso quer dizer que uma linguagem que possa ser rapidamente dominada é mais importante para o objetivo final do sistema que utilizará as informações coletadas.

O processo é repetido com diferentes tríades de elementos até que o especialista não consiga pensar em outros construtos discriminantes. O resultado é uma matriz que relaciona elementos e seus construtos através de medidas de similaridade. Um exemplo de construtos, de suas características e das cotações resultantes da elicitación de conhecimento para a tarefa de seleção de uma linguagem de computador é mostrado na TABELA 1.

TABELA 1: Matriz resultante do uso da técnica de grade de repertório

Atributo	Orientação	Facilidade de Programação	Tempo de Treinamento	Disponibilidade
Característica Oposto	Simbólica (3) Numérica (1)	Alta (3) Baixa (2)	Alta (1) Baixa (3)	Alta (3) Baixa (1)
LISP	3	3	1	1
PROLOG	3	2	2	1
C	2	3	2	2
COBOL	1	2	1	2

Fonte: Adaptado de Mastella (2004)

- **Classificação de cartões (*sorting*)**

Trata-se de uma técnica na qual as entidades de um domínio são classificadas para determinar como o especialista classifica o seu conhecimento (HUA, 2008). De acordo com Shadbolt e Swallow (1993) e Shadbolt (2005), essa técnica é utilizada no intuito de capturar o modo como especialistas comparam e ordenam conceitos. A forma mais simples é conhecida por classificação de cartões (*cardsorting*), na qual são utilizados cartões constando nomes de determinados conceitos. O especialista tem a tarefa de classificar os cartões em pilhas que tenham algo em comum.

De acordo com Mastella (2004), a técnica de classificação de fichas visa à identificação e organização de termos ou conceitos e seus relacionamentos num domínio particular, tais como a classificação dos objetos, as hierarquias e outras descrições estáticas dos objetos do domínio, segundo a visão do especialista. Uma lista

de termos do domínio é obtida a partir das entrevistas iniciais ou extraída da análise de protocolos. O método de aplicação consiste em: escrever o nome dos objetos em fichas; separar em grupos, seguindo algum critério, repetidas vezes; selecionar um dos grupos e repete-se o procedimento; incluir ou eliminar objetos se achar necessário (MASTELLA, 2004, p 24).

Mastella (2004) explica que os conceitos são apresentados em cartões ao especialista, e este é instruído a classificar os cartões que possuem alguma similaridade ou que são do mesmo tipo em pilhas, formando categorias de conceitos e nomeando-as sempre que possível. As cartas são embaralhadas novamente e o especialista deve agrupá-las novamente de forma a criar novas categorias.

A FIG 5 a seguir mostra um exemplo fictício em que o especialista precisa separar os objetos carro, pardal, barco, gato, ônibus e trem. As fichas são novamente unidades e é solicitado novamente que separe em dois grupos utilizando outro critério, até que não existam mais critérios de separação. Nesse caso, um dos grupos é selecionado e o processo é reiniciado.

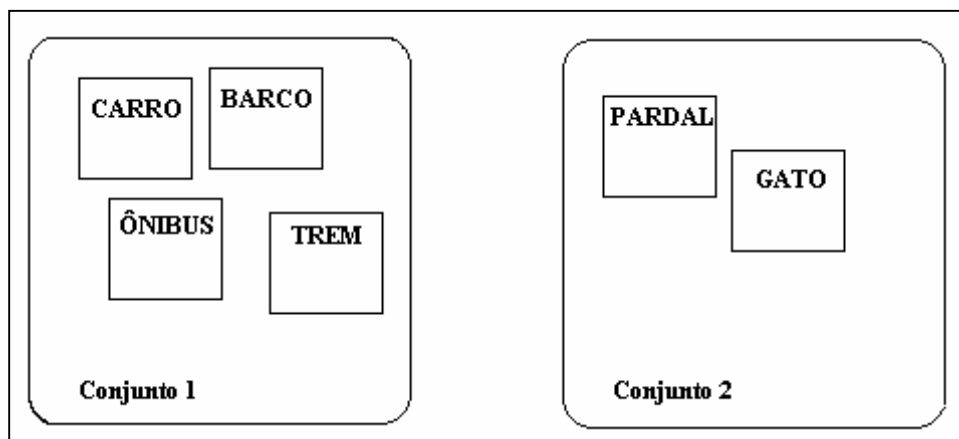


FIGURA 5: Conjunto de objetos segundo os critérios "veículos" e "animais".

Fonte: Adaptado de Mastella (2004)

Durante a classificação, o especialista retira objetos sinônimos ou irrelevantes e inclui outros que, segundo seu critério, estão ausentes, como representado pela FIG. 6

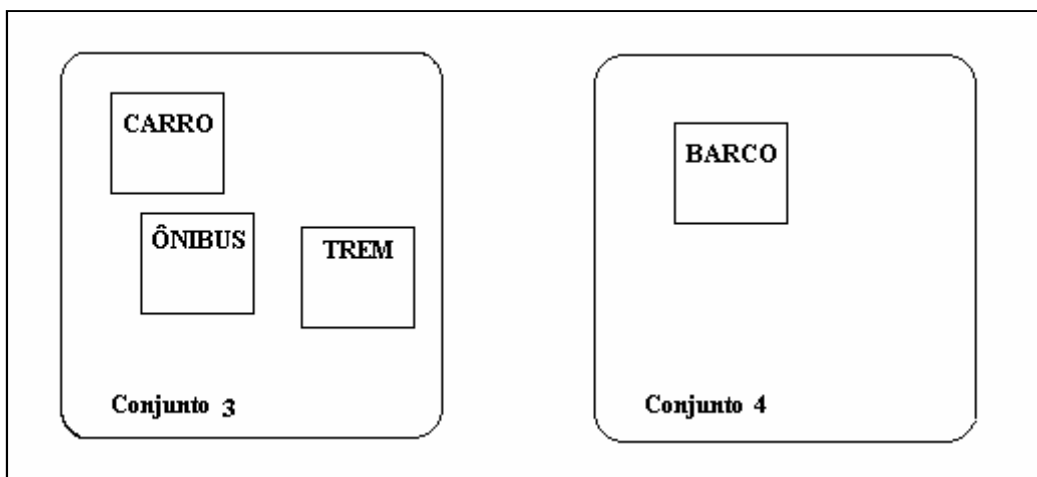


FIGURA 6: Conjunto 1 segundo os critérios "veículos terrestres" e "veículos aquáticos"
Fonte: Adaptado de Mastella (2004)

Mastella (2004) apresenta algumas considerações sobre esta técnica, a saber: i) processo que se repete muitas vezes e podem ser feitas algumas variações; ii) resulta em um grupo de componentes que compartilham atributos comuns; iii) possibilita a obtenção de termos não evidenciados através de entrevistas, o reconhecimento de conceitos que são sinônimos.

De acordo com Mastella (2004), a aplicação dessa técnica torna-se difícil nos domínios muito complexos onde um número excessivamente grande de termos pode inviabilizar o uso de cartões. De acordo com esta autora, a prática aponta que os melhores resultados são obtidos quando a técnica é aplicada sobre menos do que uma centena de conceitos. Resultados medíocres são também obtidos em domínios excessivamente procedimentais, onde a explicitação das hierarquias de termos não parece fazer senso ao especialista.

- **Limitação de Informação e tarefa de restringir processamento**

De acordo com Shadbolt (2005) e Hua (2008), essas técnicas são utilizadas quando se pretende limitar o tempo e a informação a disposição do especialista, quando este desempenha uma tarefa que normalmente exigiria recursos para desempenhá-la. É um meio para estabelecer tarefas e informações chave a serem utilizadas.

Uma variante é a “técnica de 20 questões”, cujo objetivo é adivinhar o que o engenheiro do conhecimento está pensando a respeito de algo. Ao especialista é permitido fazer perguntas ao engenheiro, que apenas responde sim ou não. As perguntas e a ordem na qual elas são proferidas sugerem aspectos importantes e que devem ser

considerados para fins de aquisição, como, por exemplo, propriedades-chave ou categorias em uma ordem de prioridade.

- **Diagrama**

Diz respeito à geração e ao uso de redes de diagramas, tais como mapas conceituais, redes de transição de estado e mapa de processos (HUA, 2008). O engenheiro do conhecimento suscita ideias no especialista, através de referência em um diagrama em papel ou computador. Tal referência também é empregada quando se utiliza outros métodos, como no caso da *laddering*. Mapas conceituais são muito empregados quando se pretende adquirir conhecimento para o desenvolvimento de sistemas orientados a objeto. Shadbolt (1994) explica que *laddering* é uma técnica utilizada quando se pretende tornar clara a relação entre os construtos eliciados pela técnica *repertory grid*, para organizá-los em relações hierárquicas.

Sob o ponto de vista do emprego dessas técnicas, Wang *et al* (2006) apresentam uma metodologia e um *plug-in* do Protégé para a elicitação do conhecimento a partir de documentos, que utilizam o *cardsorting* e o *laddering*. O objetivo do trabalho é integrar as técnicas de elicitação do conhecimento ao processo de construção de ontologias, de modo a tornar esse processo mais eficiente e menos propenso a erros. A proposta dos autores é baseada em princípios de desenvolvimento centrados no usuário, permitindo capturar requisitos para construção de ontologias.

Exemplo de como a técnica de *laddering* pode ser empregado é apresentado por Mastella (2004). O engenheiro suscita no especialista a geração de um ponto de início, um conceito inicial. Pode-se utilizar o domínio das técnicas de aquisição como um exemplo. O conceito inicial seria “técnicas de aquisição de conhecimento”. É feita uma tentativa de mover a hierarquia abaixo fazendo uma pergunta que extraia exemplos de diferentes tipos de técnicas, tais como “técnicas manuais” e “técnicas auxiliadas por computador”. A hierarquia poderia ser estendida pedindo exemplos de “técnicas manuais” e assim por diante. Para mover-se um nível abaixo na hierarquia, utiliza-se a questão “Você pode fornecer EXEMPLOS dessa CLASSE?”. Perguntas similares são usadas para o engenheiro mover-se dentro do mesmo nível hierárquico (“Quais outros exemplos dessa CLASSE existem para o ITEM 1?”) ou para cima (“O ITEM 1 e o ITEM 2 são exemplos de que CLASSE?”). Essa última questão precisa ser empregada

com cuidado, para que itens de níveis muito diferentes não sejam incluídos na mesma pergunta (MASTELLA, 2004, p.27).

Alguns tipos de *ladders* que podem ser coletados empregando essa técnica são apresentados por Shadbolt e Swallow, (1993):

- **De conceitos:** mostra conceitos e instâncias e as classes e subclasses às quais eles pertencem. Também conhecida como taxonomia, é vital para representar conhecimento na maior parte dos domínios;
- **Composicional:** mostra a maneira como um objeto é composto de suas partes. Também chamada partonomia, ela é útil para entender entidades de organização complexa como máquinas, documentos e empresas;
- **De decisões:** mostra os caminhos de ação alternativos para uma tomada de decisão. É útil para representar conhecimento procedimental;
- **De atributos:** mostra atributos e seus valores como sub-nodos. É uma maneira de representar todas as propriedades associadas a um conceito.
- **De processos:** Mostra processos no domínio (tarefas, atividades) e seus subprocessos, também representando conhecimento procedimental.

Por fim, Hoffman (1992) apresenta uma proposta de classificação para as técnicas de AC subdivida em três categorias: i) *análise de tarefas*, em geral atribuídas aos engenheiros (imersão na literatura, análise de tarefas, relatórios verbais); ii) *tipos de entrevista*: estruturadas, não estruturadas e semi-estruturadas; iii) *técnicas elaboradas*, como análise de decisão, tomada de decisão em grupo, tarefas de classificação e de gradação, limitação de informação, construção de grafos, dentre outras. Ainda segundo o autor, tal divisão busca responder o que os especialistas fazem, o que os especialistas dizem que fazem e o que eles fazem quando são limitados de alguma maneira em suas ações.

A organização apresentada no QUADRO 3 a seguir apresenta um conjunto de técnicas discutidas na literatura que aborda a temática em: Ciência da Informação, Engenharia do conhecimento e Biomedicina, bem como de propostas complementares de outros autores e campos científicos.

Em ambos os casos, verificou-se que o método escolhido pode variar conforme o contexto, ou seja, como fazendo parte do processo de desenvolvimento de sistemas,

para construção de ontologias, tesauros, em projetos de gestão de conhecimento, entre outros. Vale ressaltar que a literatura pesquisada apresenta uma gama extensa e variada de técnicas de AC.

QUADRO 3: Consolidado de técnicas para Aquisição de Conhecimento

CATEGORIA	SUBCATEGORIA	OBJETIVO	INSTRUMENTO	INTERAÇÃO	RESULTADO
Entrevista	Estruturada, não estruturada e semi-estruturada	Visão geral do domínio; Elicitar conhecimento básico (facilmente verbalizado) Detalhar o conhecimento sobre conceitos, suas propriedades e relações.	Manual	Direta	Varia de acordo com a pergunta. Elicitar principais termos e conceito representativos do domínio
	Questionário	Questões que podem ser aplicadas a partir de uma amostragem significativa e representativa de especialistas	Manual	Direta	Varia de acordo com o objetivo da tarefa
Classificação de conceitos	Classificação de cartões (<i>Card Sorting</i>)	Revelar o conhecimento em relação a classes, propriedades de prioridades. Organizar termos ou conceitos, seus relacionamentos num dado domínio.	Manual	Indireta	Classificação de entidades, objeto, hierarquias e outras descrições estáticas dos objetos do domínio
Elicitação de Construtos	<i>Repertory Grid</i> (Grades de Repertório)	Obter informações sobre como o especialista discrimina as entidades no domínio.	Manual	Indireta	Atributos (entidades fornecidas) Mapeamento conceitual do domínio
	<i>Laddering</i>	Clarificar relações entre os construtos eliciados pela técnica de repertory grid e organizá-los em relações hierárquicas. Tipos de Ladders: Conceitos : mostrar conceitos e instancias, classes e subclasses às quais pertencem. Composicional : mostrar a maneira como um objeto é composto de suas partes; Decisões : mostrar caminhos de ação alternativos para uma tomada de decisão. Apresentar conhecimento procedimental Atributos : mostrar atributos e seus valores como sub-nós. Mostrar propriedades associadas a um conceito. Processo : Mostrar processos no domínio e seus subprocessos, também representa conhecimento procedimental.	Manual	Indireta	Mapa hierárquico das tarefas do domínio.
Observação	Observação focada	Fornecer impressão global do domínio.	Manual	Direta	Identificação de conhecimento implícito dos especialistas
	Participação ativa		Manual	Direta	Conhecimento e habilidade necessários para desempenhar uma tarefa
Feedback	<i>Teachback</i>	Obter <i>feedback</i> dos especialistas a partir da descrição de uma parte do conhecimento adquirido em sessões anteriores.	Manual	Direta	Correção de equívoco conceitual Esclarecimento de mal entendidos

Maapeamento	Maapeamento conceitual	Ajudar especialistas explorar os relacionamentos entre conceitos	Interativo	Indireta	Mapa conceitual
Análise de Documentos	Imersão na Literatura	Levantar os conceitos básicos o domínio em questão; Compreender minimamente o domínio do conhecimento.	Manual	Indireta	Obtenção de principais conceitos do domínio para conhecimento mínimo do domínio. Compreender um mínimo do domínio de trabalho, para que as entrevistas possam começar de um ponto mais adiantado.
	Análise de Assunto	Extrair conceitos que traduzem a essência do conhecimento registrado em um documento	Manual	Indireta	Conceitos que melhor expressam o assunto do documento
	Análise Informal	Estudar principais conceitos no domínio, obtidos em livros, manuais, publicações, entre outros	Manual	Indireta	Varia dependendo dos documentos disponíveis e se haverá ou não interação com os especialistas
Análise de Domínio	Bibliometria	Mapear campos do conhecimento a partir do domínio	Manual	Indireta	O que é o domínio em estudo Quem são os especialistas e a literatura Identificar os especialistas a partir das publicações
	Corpus linguístico	Definir o domínio do conhecimento	Automático	Indireta	O que é o domínio em estudo
Técnicas de Grupo	Brainstorming,	Elicitar conhecimento baseado em consenso	Manual	Direta	Geração de consenso em grupo de especialistas
	JAD- Joint Application Design		Manual	Direta	Discussões em grupo, onde as decisões são baseadas em consenso.
Análise de cenários	Narrativas	Elicitar conhecimento por meio de narrativas	Manual	Direta	Dados sobre as necessidades e atividades dos funcionários das empresas
Formação de consenso	Ferramenta Wiki	Obter consenso	Interativa	Indireta	Obtenção de conhecimento especializado em recursos centralizados Permite aos usuários anexarem comentários

Fonte: dados da pesquisa.

2.5 Aquisição de conhecimento na construção de ontologias

A presente seção contextualiza a AC no processo de construção de ontologias. Embora estudos relacionados a ontologias sejam discutidos a fundo pela literatura científica, abordando a temática em diferentes campos do conhecimento e sob diferentes perspectivas, o presente trabalho enfoca sua tipologia e metodologias para construção. Apesar da preocupação em cobrir questões relevantes não se pretendeu esgotar a discussão, pois tal ação deveria ser realizada de modo exaustivo, o que não fez parte dos propósitos do presente trabalho, uma vez que a temática principal é a própria atividade de AC. Vários autores têm estudado ontologias, tanto em Ciência da Computação (GENESERETH; NILSSON, 1987; GRUBER, 1993; GUARINO, 1995; GUARINO; GIARETTA, 1995; GUARINO, 1998; SOWA, 2000; SMITH, 2003), quanto em Ciência da Informação (SØERGUEL, 1997; VICKERY, 1997; WAND; STOREY; WEBER, 1999; GILCHRIST, 2003; ALMEIDA, 2006).

A seção 2.5.1 apresenta uma breve introdução e conceituação sobre ontologias, abrangendo aspectos relevantes para a presente pesquisa; a seção 2.5.2 lista as principais metodologias para construção de ontologias onde a AC tem lugar; a seção 2.5.3 traz considerações complementares a AC na construção de ontologias, especificamente sobre o domínio do conhecimento e a expertise de especialistas.

2.5.1 Ontologias: uma breve introdução

O termo ontologia é originário da filosofia e tem sido utilizado desde os anos 80 em Ciência da Computação e em Ciência da Informação, para designar uma estrutura de organização da informação baseada em um vocabulário representacional. Segundo Almeida (2006), a origem é a palavra aristotélica “categoria”, utilizada para designar o ato de classificar e caracterizar alguma coisa. Diz respeito à definição de categorias para as “coisas” que existem em um mesmo domínio de conhecimento. É uma estrutura que organiza termos, conceitos e suas relações em um domínio específico, gerando modelos com variados níveis de formalidade.

De modo geral, ontologias são apresentadas pela Ciência da Computação como sendo alternativas propostas para a modelagem do conhecimento - atividade que é

parte do processo de construção de sistemas especialistas, onde o tema é discutido desde a década de 1960. Nos anos 80, surgiram pesquisas em Engenharia de Software aplicando ontologias à fase de modelagem conceitual no desenvolvimento de sistemas de informação (SI).

Na Ciência da Informação, a temática é abordada como forma de representação e organização do conhecimento, propondo meios de estruturá-lo e torná-lo disponível para uso das pessoas ou legível por máquinas. Vickery (1997) destaca que os primeiros trabalhos desenvolvidos sob o tema na Ciência da Informação foram apresentados em 1996 na conferência da *Informatics in the Scientific Knowledge – ISK*.

Conceitualmente, a literatura apresenta diversas definições para ontologias, seus tipos, propostas para aplicação em diferentes áreas do conhecimento, bem como para a sua construção - metodologias, ferramentas e linguagens. Uma definição simples, completa e adotada no presente trabalho foi apresentada por Borst (1997): “*uma ontologia é uma especificação formal e explícita de uma conceitualização compartilhada*”. “Formal” por ser legível para computadores; “especificação explícita” diz respeito a conceitos, propriedades, relações, funções, restrições, axiomas, explicitamente definidos; “compartilhado” quer dizer conhecimento consensual e, “conceitualização” diz respeito a um modelo abstrato de algum fenômeno do mundo real.

Considerando a variedade de tipos de ontologias existentes, Almeida (2002); Almeida e Bax (2003) e Almeida (2006) apresentam um quadro consolidado, conforme classificação, a saber: função, grau de formalismo, aplicação, estrutura e conteúdo, como demonstrado no QUADRO 4 a seguir:

QUADRO 4: Classificação conforme os tipos de ontologias

Abordagem	Classificação	Descrição
Quanto à função Mizoguchi, Vanwelkenhuysen e Ikeda (1995)	Ontologias de domínio	Reutilizáveis no domínio fornecem vocabulário sobre conceitos, seus relacionamentos, sobre atividades e regras que os governam.
	Ontologias de tarefa	Fornecem um vocabulário sistematizado de termos, especificando tarefas que podem ou não estar no mesmo domínio.
	Ontologias gerais	Incluem um vocabulário relacionado a coisas, eventos, tempo, espaço, casualidade, comportamento, funções etc.

Quanto ao grau de formalismo Uschold e Gruninger (1996)	Ontologias altamente informais	Expressa livremente em linguagem natural.
	Ontologias semi-informais	Expressa em linguagem natural de forma restrita e estruturada.
	Ontologias semi-formais	Expressa em uma linguagem artificial definida formalmente.
	Ontologia rigorosamente formal	Os termos são definidos com semântica formal, teoremas e provas.
Quanto à aplicação Jaspere Uschold (1999)	Ontologias de autoria neutra	Um aplicativo é escrito em uma única língua e depois convertido para uso em diversos sistemas, reutilizando-se as informações.
	Ontologias como especificação	Cria-se uma ontologia para um domínio, a qual é usada para documentação e manutenção no desenvolvimento de <i>softwares</i> .
	Ontologias de acesso comum à informação	Quando o vocabulário é inacessível, a ontologia torna a informação inteligível, proporcionando conhecimento compartilhado dos termos.
Quanto à estrutura Haav e Lubi (2001)	Ontologias de alto nível	Descrevem conceitos gerais relacionados a todos os elementos da ontologia (espaço, tempo, matéria, objeto, evento, ação etc.) os quais são independentes do problema ou domínio.
	Ontologias de domínio	Descrevem o vocabulário relacionado a um domínio, como por exemplo, medicina ou automóveis.
	Ontologias de tarefa	Descrevem uma tarefa ou atividade, por exemplo diagnósticos ou compras, através da inserção de termos especializados na ontologia.
Quanto ao conteúdo Van-Heijst, Schreiber e Wielinga (2002)	Ontologias terminológicas	Especificam termos que serão usados para representar o conhecimento em um domínio (por exemplo, os léxicos).
	Ontologias de informação	Especificam a estrutura de registros de bancos de dados (por exemplo, os esquemas de bancos de dados).
	Ontologias de modelagem do conhecimento	Especificam conceitualizações do conhecimento, têm uma estrutura interna semanticamente rica e são refinadas para uso no domínio do conhecimento que descrevem.
	Ontologias de aplicação	Contêm as definições necessárias para modelar o conhecimento em uma aplicação.
	Ontologias de domínio	Expressam conceitualizações que são específicas para um determinado domínio do conhecimento.
	Ontologias genéricas	Similares às ontologias de domínio, mas os conceitos que as definem são considerados genéricos e comuns a vários campos.
	Ontologias de representação	Explicam as conceitualizações que estão por trás dos formalismos de representação do conhecimento

Fonte: Adaptado de Almeida (2002), Almeida e Bax (2003)

No consolidado apresentado pelos autores foi possível notar que não há consenso na classificação de tipos de ontologias. De acordo com Almeida (2006), muitas vezes, as categorias apresentadas se sobrepõem; um mesmo nome de categoria é utilizado por mais de um autor para significados distintos e nomes diferentes se

referem ao mesmo tipo de ontologia. Na presente pesquisa, destaque particular é dado às ontologias de domínio e à ontologia de alto nível BFO. Ambas descritas a seguir.

2.5.1.1 Ontologia de domínio

Com relação às ontologias de domínio, Guizzardi (2000) explica que são tipos de ontologias que fornecem vocabulário sobre conceitos, seus relacionamentos, atividades e regras que os governam. Descrevem o vocabulário relacionado a um domínio, como: medicina, música, direito, entre outros, e expressam conceitualizações específicas para um determinado domínio do conhecimento. Os objetivos principais de se desenvolver uma ontologia de domínio são: compartilhar informação, reusar elementos do domínio, tornar suposições do domínio explícitas, separar conhecimentos de domínio de conhecimento operacional e analisar o conhecimento do domínio. (NOY, N. F., MCGUINNESS, 2001).

Ontologias de domínio apresentam algumas finalidades, a saber: *i) Comunicação*: auxilia as pessoas a se comunicarem de formas variadas acerca de um determinado conhecimento; ajudam as pessoas entenderem o domínio do conhecimento, estabelecendo os parâmetros para a obtenção do consenso, numa comunidade sobre o vocabulário a ser usado nas suas interações; *ii) Formalização*: a especificação do domínio elimina contradições e inconsistências envolvendo restrições, resultando numa especificação sem ambiguidades. *iii) Representação e reuso*: um vocabulário consensual é criado pela ontologia, que representa o conhecimento de um domínio de forma explícita. Esse conhecimento formalizado pode ser utilizado em diferentes aplicações, atendendo a propósitos distintos e por diferentes equipes de desenvolvimento (GUIZZARDI, 2000, p.41).

Para o presente trabalho, relevante considerar a ontologia de domínio denominada BLO, em desenvolvimento no âmbito do *BloodProject*. A BLO compreende um conjunto de ontologias co-relacionadas, cada uma delas englobando um grupo de questões relevantes na área de hematologia e hemoterapia (ALMEIDA *et al*, 2011; ALMEIDA; PROIETTI; SMITH, 2011). Conforme exposto pelos autores, a BLO foi concebida considerando as melhores práticas para ontologias e fundamentada nos princípios preconizados pela BFO. As seguintes ontologias foram planejadas para a BLO, a saber: *BLO-Core*; *BLO-Management*; *BLO-Products* e *BLO-Administratives*.

Cabe nesse momento descrever cada uma destas ontologias (ALMEIDA *et al*, 2011; ALMEIDA; PROIETTI; SMITH, 2011).

BLO-*Core* enfatiza aspectos fisiológicos do sangue e apresenta informação básica essencial para a pesquisa e prática em hematologia. O vocabulário objetiva fornecer informações básicas sobre substâncias químicas, como, por exemplo, aspectos químicos e estruturais de DNA e RNA, genes e cromossomos, proteínas, aminoácidos, bases, enzimas, vetores e outros; bem como das bases moleculares, imunológicas e celulares do sangue, como por exemplo, hematopoese, tipos e funções de células sanguíneas, homeostase e outros (ALMEIDA *et al*, 2011; ALMEIDA; PROIETTI; SMITH, 2011).

Além desses fundamentos, a BLO-*Core* abrange distúrbios provenientes diretamente de questões genéticas e moleculares relacionadas ao sangue (SCHALLER *et al*, 2008). Nesse grupo, consideram-se doenças relacionadas às hemácias (hereditárias e adquiridas), desordens da hemóstase e coagulação, doenças hematológicas não malignas e malignas. Um fragmento dos resultados da BLO-*Core* em um navegador da internet é apresentada na FIG.7

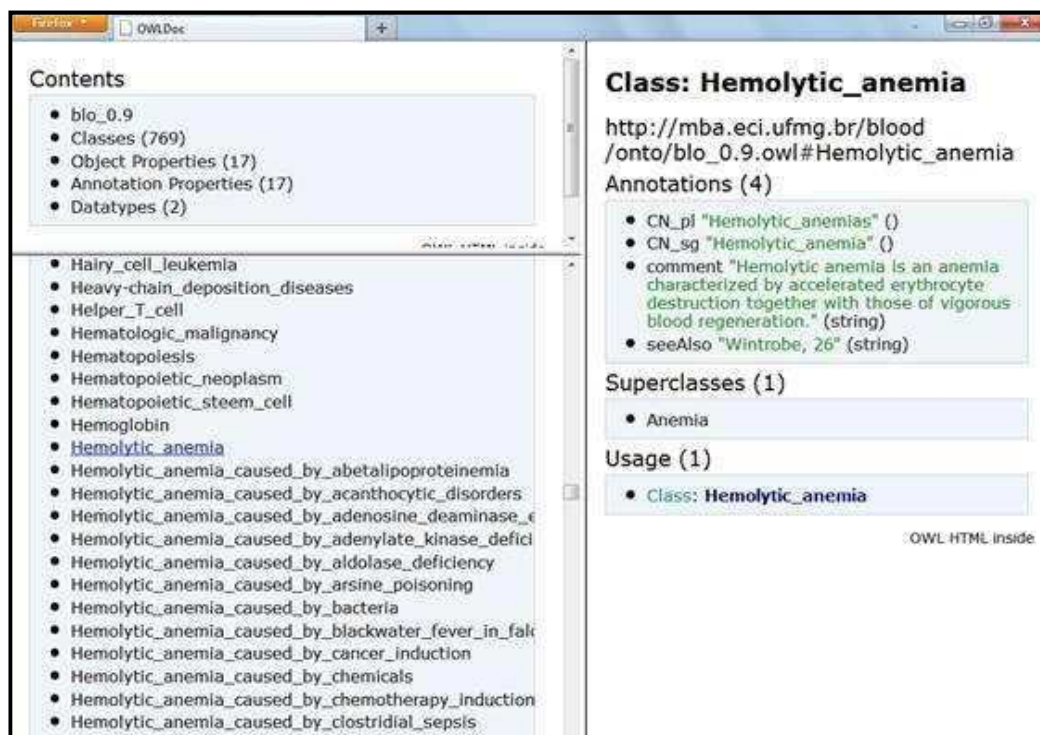


FIGURA 7: Resultado da BLO-Core na Internet
Fonte: Obtido da BLO

A *BLO-Management* abrange os processos relevantes envolvidos no processamento de sangue em bancos de sangue e serviços relacionados. O objetivo é representar a miríade de processos e eventos que compõem o campo da transfusão de sangue. (ALMEIDA *et al*, 2011; ALMEIDA; PROIETTI; SMITH, 2011). Os processos relacionados à manipulação do sangue ⁽¹⁸⁾ que compõem a *BLO-Management* são:

- Sistemas de qualidade para bancos de sangue: i) gestão organizacional; ii) recursos humanos, responsabilidades e competências; iii) políticas e procedimentos técnicos; iv) gestão para acreditação; v) gestão de processos; vi) não-conformidades;
- Processo de seleção de doadores: i) registro de candidatos; ii) determinação da elegibilidade; iii) registro da história médica do doador; iv) exame físico.
- Coleta de sangue: i) identificação de recipientes para armazenar sangue; ii) preparação da punção venosa; iii) flebotomia e coleta de amostras; iv) cuidado ao doador após a flebotomia;
- Procedimentos de teste sanguíneo: i) determinação do grupo ABO e tipo Rh ii) sífilis; iii) ácidos nucleicos do HIV; iv) ácido nucleico do vírus da hepatite C ; v) anti-HIV-1 e anti-HIV-2; vi) anticorpos para o antígeno nuclear da hepatite B; vii) anti-HCV; viii) os anticorpos para o HTLV (vírus T-linfotrópico humanos);
- Estudos de sorologia de grupos sanguíneos: i) aglutinação; ii) hemólise; iii) precipitação;
- Transfusão de sangue.

A *BLO-Products* manipula a multiplicidade de produtos derivados do processamento de sangue em escala mundial. Em geral, uma descrição de rótulos padronizados identifica os produtos de sangue nos diversos locais para os quais eles possam ser enviados. Nesse contexto, uma descrição comum é essencial para garantir a correta utilização do produto final em diferentes situações, como, por exemplo, em diferentes países (ALMEIDA *et al*, 2011; ALMEIDA; PROIETTI; SMITH, 2011).

A *BLO-Products* lida com questões de integração de dados, ainda não totalmente resolvidas em padrões na área de hematologia, como o ISBT-128²¹. O *International Council for Commonality in Blood Banking Automation* (ICCBBA) define o ISBT-128 como “uma norma internacional para a transferência de informações associadas ao transplante de tecidos humanos, a terapia celular, e transfusão de sangue” (ICCBBA, 2010). O ISBT descreve um esquema de código de barras para derivados do sangue, o que permite identificar uma amostra de sangue.

2.5.1.2 Ontologia de alto nível

Embora exista uma diversidade de ontologias de alto nível, como a *KR* – *Knowledge Representation Ontology* - Ontologia de Sowa (SOWA, 2006); Ontologia CYC (CYC PROJECT, 2005); Ontologia SUMO - *Suggested Upper Merged Ontology* (SUMO, 2004) e; Ontologia DOLCE - *Descriptive Ontology for Linguistic and Cognitive Engineering* (GUARINO *et al.*, 2003), dentre outras. O presente trabalho discorre apenas a Ontologia Formal Básica, cuja sigla em inglês a denomina como BFO, por se tratar da base de fundamento da BLO. O restante da presente seção se fundamenta nos trabalhos de GRENON, 2003; GRENON; SMITH, 2004; SMITH; GRENON, 2004; GRENON *et al.*, 20004; GRENON, 2004; dentre outros.

A Ontologia Formal Básica- *Basic Formal Ontology* (BFO)²² é uma ontologia de alto-nível criada em 2002 para apoiar pesquisas científicas. Ontologias de alto nível descrevem conceitos gerais como espaço, tempo, matéria, objeto, evento, ação, para citar alguns, os quais não dependem do problema ou domínio (ARP; SMITH, 2008). É chamada de ontologia básica por ser concebida como uma ontologia dos constituintes básicos da realidade. Por esse motivo, não define classes, nem entidades ficcionais. É concebida para ser muito pequena e capaz de representar as categorias de alto nível comuns às ontologias de domínio, desenvolvidas pelos cientistas em diferentes

²¹Padrões de terminologia para Sangue, Terapia Celular e Descrição de produtos dos tecidos. Disponível na Internet em <http://iccbba.org/>. Acesso em 24 de julho de 2010

²²Disponível na Internet em: <<http://www.ifomis.org/bfo/>>. Acesso em: 20/03/2011

domínios e a diferentes níveis de granularidade (GRENON; SMITH, 2004; SMITH; GRENON, 2004).

De acordo com Arp e Smith (2008), muitos membros das Ontologias Biológicas e Biomédicas Abertas - *Open Biological and Biomedical Ontologies* (OBO) utilizam a BFO para auxiliar na categorização de entidades e relacionamentos em seus respectivos domínios de investigação.

No escopo da OBO, uma das mais importantes iniciativas que fazem uso da BFO é a *Gene Ontology* (GO). A GO é um esforço colaborativo, considerado o maior repositório de dados genômicos do mundo. De acordo com o Consórcio GO (2010)²³, a GO descreve os produtos dos genes em termos de processos biológicos, componentes celulares e respectivas funções moleculares. A principal realização do GO é permitir consultas em diversas bases de dados, promovendo a integração entre outros recursos e em níveis distintos.

Outra iniciativa relevante é a *Protein Ontology* (PRO), uma ontologia para proteínas que representa avanços em termos de descobertas biomédicas. Ela contém um conjunto enorme de dados relacionados à proteína humana e de ratos na forma de representações explícitas de produtos de gene, que dão apoio a inferências computacionais. Entre outras, a principal contribuição da PRO é permitir uma comparação de informações sobre organismos relacionados em ontologias correlacionadas (DARREN, A. *et al.*, 2006).

Ontologias como essas são submetidas à verificação e aquelas consideradas “certificadas” são fontes relevantes para obtenção de termos e relações, uma vez que passam por testes e discussões ao longo de anos. O uso dessas abordagens facilita o gerenciamento da complexidade do campo biomédico, promove o reuso dos dados e evita ambiguidades, resultando em redução de custos.

Segundo Grenon e Smith (2004), a BFO adota uma visão da realidade e compreende entidades continuantes e ocorrentes. Continuantes são entidades que continuam ou persistem através do tempo, como objetos, qualidades e funções. Ocorrentes, por sua vez, são eventos ou acontecimentos onde participam as continuantes. Para esses autores, continuantes e ocorrentes existem no tempo de

²³ Gene Ontology Consortium, disponível em: <http://www.geneontology.org/>. Acesso: maio 2011

diferentes maneiras. Continuantes estão elas mesmas, sujeitas a constantes mudanças, já os ocorrentes dependem dos objetos continuantes para serem seus portadores.

Para Smith (2006), a melhor forma de identificar se uma entidade é um continuante ou um ocorrente é pela metáfora da máquina fotográfica e da câmara de vídeo: “*nós só podemos fotografar continuantes enquanto que os ocorrentes só podem ser captados em vídeo*”. Smith (2006) entende que tudo o que existe pertence a uma dessas duas categorias. E, por esta razão, esse autor defende que uma ontologia científica deve conter pelo menos estas duas categorias. Elas correspondem aos dois níveis superiores de uma ontologia, em que todos os outros elementos de uma ontologia deveriam ser submetidos.

Relevante para o entendimento acerca de tais entidades é a relação de dependência. Nesse caso, somente continuantes, segundo Smith (2003), podem ser dependentes ou independentes. Esse autor afirma que todas as entidades caracterizadas como ocorrentes são, necessariamente, entidades dependentes de um continuante dependente. A diferença central é que enquanto os continuantes independentes existem por si mesmos, os continuantes dependentes precisam dos continuantes independentes para existir.

Exemplos para esses casos seriam: i) continuantes dependentes: peso, uma determinada doença, altura, cor; ii) continuantes independentes: organismos, células e proteínas. Além disso, esse autor lembra que continuantes podem ser materiais (uma célula) ou imateriais (a cavidade do coração). A FIG. 8 ilustra a relação de dependência considerada pelo autor.

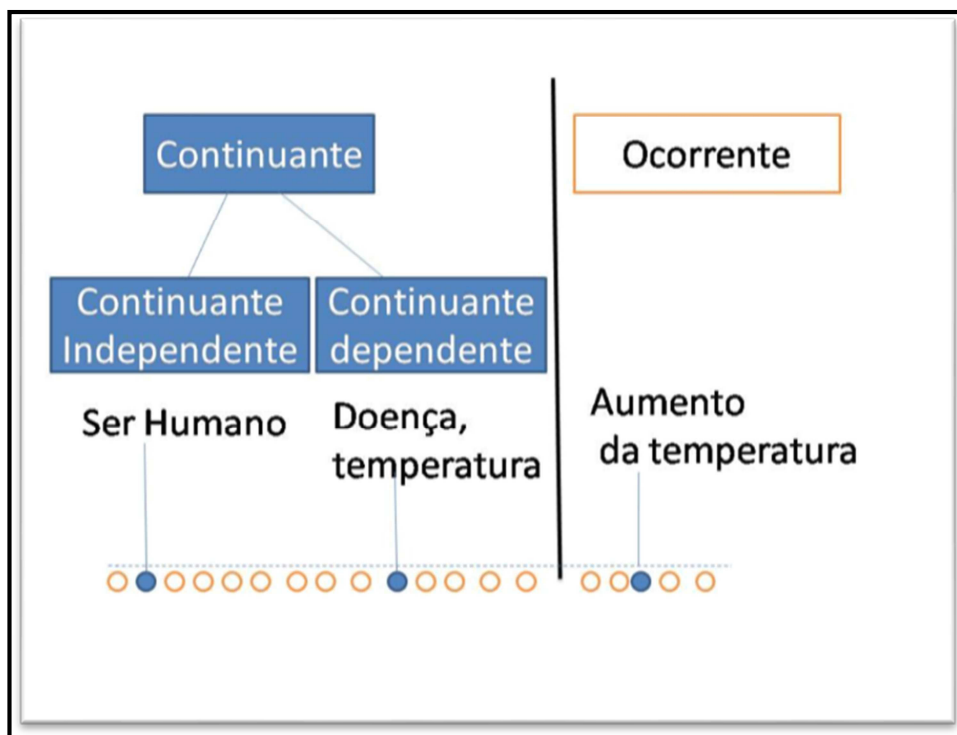


FIGURA 8: Esquema da relação de dependência das entidades continuantes
Fonte: Obtido Grenon e Smith (2004)

Acerca das principais entidades existentes, a BFO desenvolveu um *framework* de entidades denominadas SNAP e SPAN. A arquitetura SNAP-SPAN aborda tanto as entidades, quanto as ontologias através das quais as entidades são apreendidas (GRENON, 2003; SMITH; GRENON, 2004; GRENON; SMITH, 2004).

Para Smith (2003), entidades são habitantes da realidade, ou seja, tudo o que existe ou ocorre no mundo espaço-temporal é uma entidade. Uma ontologia, por sua vez, segundo esse autor, compreende as entidades que existem dentro de uma determinada parte do mundo em um dado nível de generalidade. A ontologia inclui uma taxonomia dos tipos de entidades e as relações existentes na medida em que parte do mundo é vista de dentro de uma determinada perspectiva.

Na estrutura SNAP encontram-se as entidades que se enquadram nas categorias continuantes e na ontologia SPAN, as entidades que se enquadram nas categorias ocorrentes. A ontologia SNAP pode refletir uma série de achados, por exemplo, no modo como um gráfico de temperatura é usado como base para inferir mudanças na saúde de um paciente, mas não pode capturar alterações, por exemplo, processos envolvidos com o crescimento e desenvolvimento. É a ontologia SPAN que é capaz de

capturar essas alterações. A ontologia SNAP está relacionada com o que existe em um instante único e a ontologia SPAN refere-se a uma extensão total ou a um intervalo de tempo.

Uma consideração importante apresentada por Smith e Grenon (2004) é que, tanto as ontologias SNAP quanto SPAN são indexadas não apenas pelo tempo, mas também pelo domínio e pelo nível de granularidade. Um exemplo de um domínio SNAP apresentado pelos autores seria: este organismo humano ou (o típico organismo humano, de certa idade e sexo). Já de um domínio SPAN: a vida desse organismo humano.

As relações entre as entidades SNAP e SPAN são apresentadas por Smith e Grenon (2004) e incluem: dependência, participação, início, término, criação, destruição, deterioração, sustentação, facilitação e impedimento. Para esses autores, um tipo de relação é de especial importância na integração dos componentes SNAP e SPAN: a relação entre cada entidade SNAP à entidade SPAN única que é a vida. Essa relação é difícil de definir, no entanto, pelo menos, satisfaz as seguintes condições: que a vida de um organismo é dependente deste organismo (a vida não pode existir sem seu portador) e que um organismo participa em sua vida. A vida de um organismo (continuante) é a soma de todos os ocorrentes em que participa.

Taxonomia das entidades SNAP

As principais categorias das entidades SNAP são apresentadas a seguir (FIG 9), cujas respectivas definições se encontram descritas em seguida. Ambas as definições são apresentadas conforme definido pelos autores (GRENON, 2003; GRENON; SMITH, 2004; SMITH; GRENON, 2004; GRENON *et al.*, 20004; GRENON, 2004) citados.

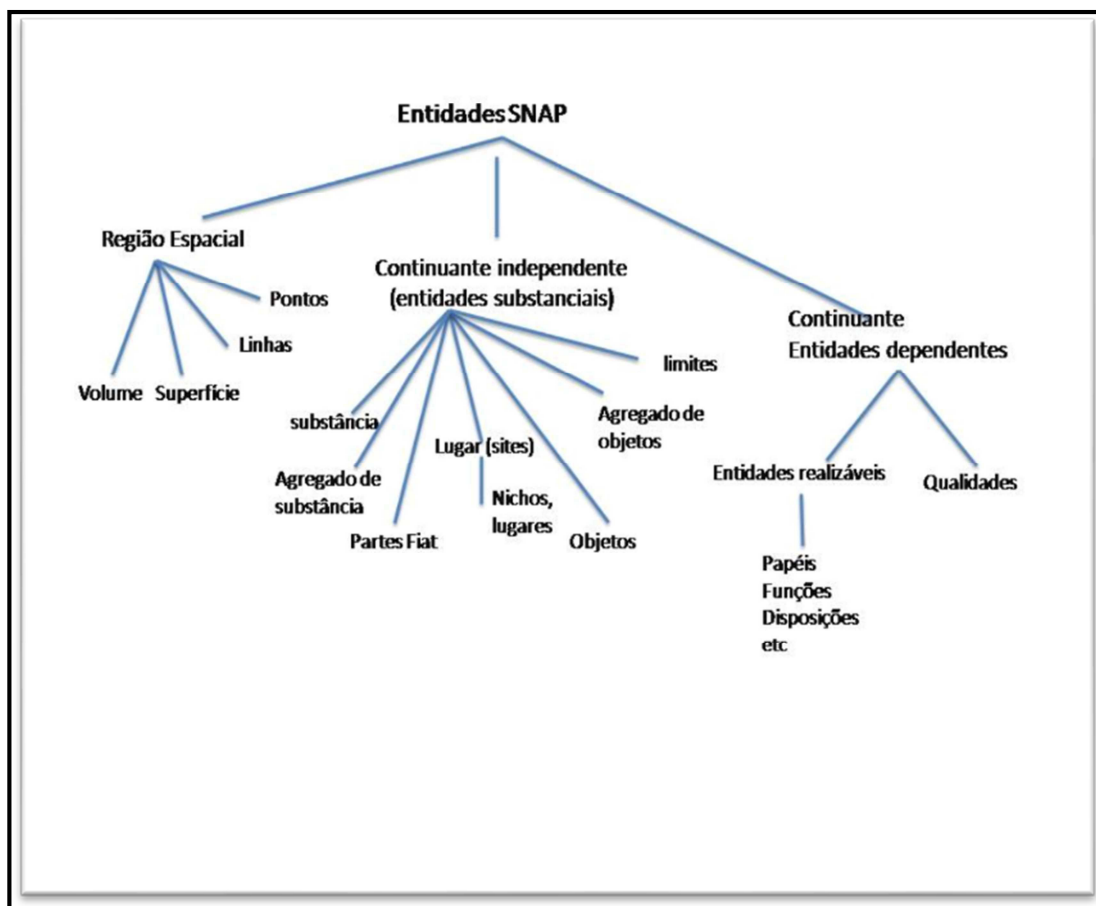


FIGURA 9: Taxonomia das entidades SNAP

Fonte: Adaptado de Grenon e Smith (2004) e Smith e Grenon et al (2004)

Regiões espaciais

As regiões espaciais são entidades em seus próprios direitos. O espaço, por exemplo, é a totalidade de tais regiões espaciais. São endurantes, de modo que outras entidades SNAP podem estar localizadas nela. Existem regiões espaciais distintas e todas as regiões espaciais são partes do espaço. Nas regiões espaciais existem as relações espaciais adquiridas em determinado momento, em virtude das relações que existem entre as próprias regiões. Assim, o cérebro está localizado em uma cavidade cerebral, um tumor cerebral pode estar localizado na matéria cinzenta intermediária do lado esquerdo da medula espinhal entre os níveis torácicos T3 e T4.

Entidades substanciais

Além de regiões espaciais, as entidades SNAP são divididas em duas grandes famílias de entidades dependentes e independentes. A categoria de entidades SNAP independentes, também chamadas de "substanciais" compreendem uma série de subcategorias, como segue:

Substâncias

Substâncias podem ser vistas, na maioria dos casos, como sinônimo de “objeto”, “coisa”, “massa”. São entidades que gozam de integridade natural. Incluem, por exemplo, não apenas os organismos, mas também as células, as pedras, os planetas, a água e uma série de artefatos autônomos como uma agulha, um saco plástico, para citar alguns. Substâncias têm as seguintes características principais: i) são independentes: não dependem de serem sustentadas ou mantidas em existência a outras entidades; ii) são portadoras de qualidades e estão sujeitas à mudança qualitativa; iii) são duradouras: preservam sua identidade ao longo do tempo e por meio de mudanças de vários tipos; iv) têm localização no espaço; v) são conjuntos auto-ligados com limites físicos que as separam de outras substâncias.

Partes Fiat

Partes fiat são partes não demarcadas por discontinuidades físicas como, por exemplo, os lobos superiores e inferiores do pulmão esquerdo, as superfícies dorsal e ventral do corpo, a faixa da superfície do cérebro demarcada como o córtex motor primário. As partes fiat podem, em alguns casos, serem transformadas em substâncias no seu próprio direito, por exemplo, nos casos em que determinado segmento do rim é demarcado como canceroso para fins de remoção cirúrgica. Outro exemplo que se aplica a esse tipo de entidade são as partes fiat da Terra, como a demarcação do Polo Norte ou do Equador; a parcela de terra demarcada em um mapa cadastral, certo volume da superfície da Terra demarcado para fins de exploração mineral.

Limites (Fronteiras)

Limites são as partes de mais baixa dimensão das entidades espaciais. Exemplos são: o limite do organismo, a superfície interna do estômago, o limite da Terra. O entendimento desses limites é crucial quando se tratar de entidades biomédicas, nesse caso, para lidar com as superfícies das substâncias e não apenas com as regiões em que essas superfícies estão localizadas. Limites podem ser divididos em dois tipos: o fiat, o físico ou bona fide. A fronteira entre lóbulos do pulmão é um limite fiat (depende das demarcações fiat), a superfície de todo o organismo é um limite bona fide (ele existe independentemente de qualquer demarcação que se queira).

Agregados

Agregados de substâncias são somas mereológicas que compreendem substâncias separadas como partes. Em oposição às substâncias, agregados não têm limites conectados. Agregados em um nível de granularidade (por exemplo, um conjunto de células) podem corresponder às substâncias (por exemplo, um organismo inteiro) em outro nível de granularidade.

Lugar (site)

Cada lugar está associado a qualquer momento com alguma região específica do espaço. São análogas às substâncias, e como as substâncias podem se mover através do espaço. Mas, não são idênticos às regiões espaciais. Exemplos de entidades do tipo Lugar (site) são: buracos, cavidades e entidades similares, como os buracos formados por agregados de células e estruturas ósseas, juntamente com os limites associados fiat, chamados seios, canais, ventrículos etc.

Nichos ou Ambientes são tipos especiais de lugares ocupados por organismos. Nichos são normalmente feitos de um meio (de ar ou água) delimitados por uma mistura fiat e limites físico. A cavidade nasal é um lugar na medida em que é habitado por microorganismos. Grenon e Smith (2004) sugerem que lugares, por vezes, atuam como espaços em torno de outras entidades e fazem isso de modos diferentes. Os endereços onde as pessoas vivem e trabalham são rótulos para certos lugares que

ocupam. No ambiente ao redor do organismo, também podem ser esculpidas porções fiat, por exemplo, correspondentes às zonas de diferentes temperaturas ou toxicidade.

Fronteira livre e lugares fisicamente limitados são lugares cujos limites sejam parciais ou totalmente sólidos são chamados de “fisicamente limitados”, outros são chamados “fronteira-livre”. A narina é um lugar, em parte, fisicamente limitada. A cavidade oral, quando os lábios estão fechados e quando a parte posterior da língua está elevada, é um lugar totalmente fisicamente limitado.

Entidades dependentes

Entidades SNAP dependentes são entidades que perduram no tempo e são inerentes às entidades substanciais. Smith e Grenon (2004) afirmam que inerência é uma relação intra-ontológica entre uma entidade SNAP dependente e seu portador substancial. A vermelhidão do glóbulo vermelho é inerente à célula sanguínea, a circunferência da cintura é inerente à cintura, a forma de um nariz é inerente ao nariz. Tal inerência é uma forma de dependência existencial no sentido de que a cor especial do sangue arterial não existiria sem o sangue arterial.

Entidades SNAP dependentes incluem: *as qualidades particularizadas* (a elasticidade da pele, a temperatura normal do corpo), *as funções* (a função do canal de nascimento para transportar o feto), *disposições* (dos dentes, que são propensos a enfraquecer), *poder* (de escovar os dentes, para evitar a sua deterioração), *passividade* (dos dentes, de se desgastarem pela escovação), *desordens* (dor de cabeça de uma pessoa por ter passado duas horas sem comer), *características físicas* (um sorriso).

Outros tipos de entidade dependente SNAP incluem: estados ou condições, planos, normas, tarefas, regras, algoritmos, procedimentos e doenças. No entanto, se a resistência e dependência são condições necessárias para entidades SNAP dependentes, não são condições suficientes. A característica distintiva destas entidades é que elas são inerentes às substâncias.

Universais

Universais formam taxonomias de vários tipos. Cada taxonomia é uma árvore cujos nós correspondem aos universais de generalidade maior e menor. Existem dois tipos principais de universais SNAP: universais substanciais e universais instanciados pelas entidades dependentes SNAP. Os primeiros incluem: *substâncias universais*: a célula, os membros, o corpo; *lugares universais*: a cavidade, o sulco do seio; universais instanciados por *agregados de substâncias* como: a família, a equipe de enfermagem, a população. Universais instanciados pelas entidades dependentes SNAP incluem: *papéis* (pai ou médico); *funções* (bombear o sangue, transportar o ar); e *qualidade* (cor ou forma), instanciada pela cor e pela forma específica do rim.

Ao combinar substâncias universais e de qualidade (ou seus associados predicados definidos), pode-se construir mais predicados construídos, como por exemplo, “um cabelo grisalho”. Não corresponde a qualquer universal, mas, a todos os cabelos que estão cinza, ou seja, a todos os cabelos cuja superfície há inerência, uma instância cinza de qualidade universal.

Taxonomia das entidades SPAN

Entidades processuais

Entidades processuais são ocorrentes ou acontecimentos, que existem no tempo pela ocorrência e envolvem os participantes de um tipo SNAP e dependem destes participantes. Estão localizadas em regiões espaço-temporais. Incluem não apenas processo, mas partes fiat arbitrárias, agregados e limites dos mesmos.

As entidades processuais podem ter estruturas mais ou menos complexas, mas ao contrário das entidades substanciais, há poucas articulações nítidas. Isso porque se fundem a outras de maneiras variadas de se constituir o todo maior, na medida em que o reino SPAN é em grande parte, um reino de fluxo (um reino de transições, caracteristicamente contínuo). Grenon e Smith (2004) explicam que essas duas entidades processuais se fundem de modo que exista alguma contrapartida no reino de substâncias SNAP.

A FIG. 10 apresenta a taxonomia das entidades SPAN. A definição para cada uma dessas entidades é apresentada a seguir.

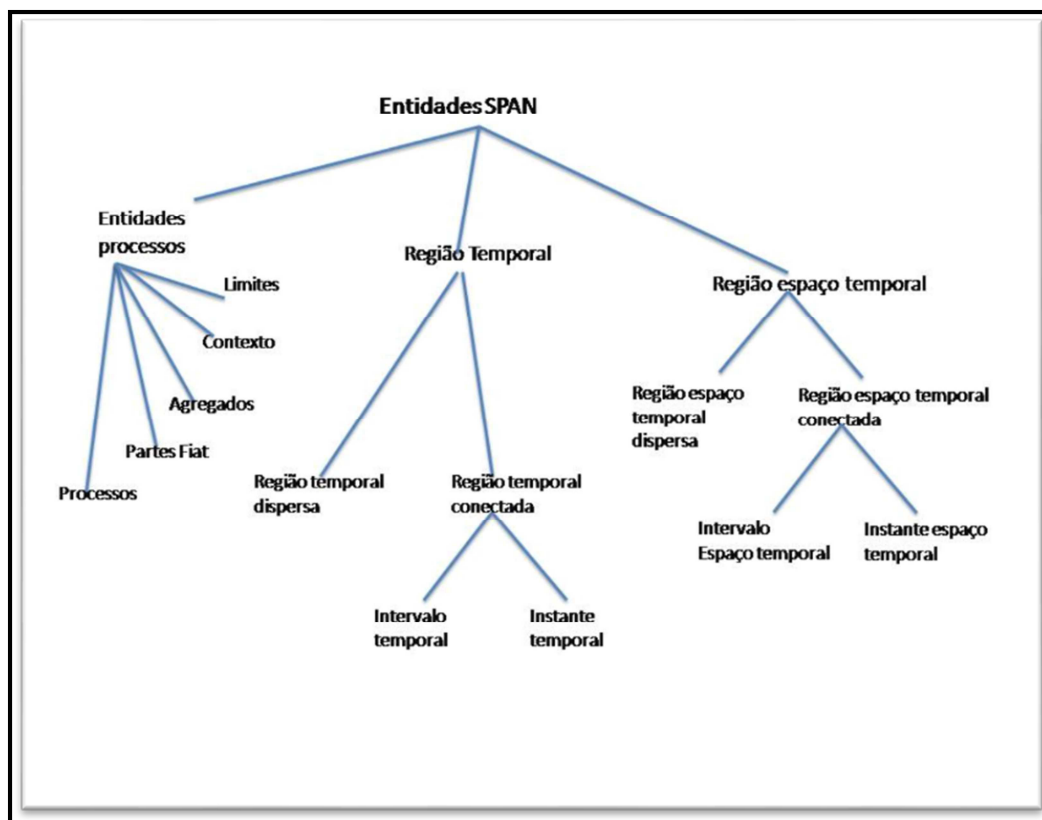


FIGURA 10: Taxonomia das entidades SPAN

Fonte: Adaptado de Grenon e Smith (2004) e Smith e Grenon et al (2004)

Processos

Unidades naturais onde têm começos e fins *bonafide* correspondentes às discontinuidades reais na ordem processual. Exemplos: a vida de um organismo ou o processo de dormir (demarcado pelo limite de adormecer e acordar). Assim como as substâncias em SNAP são definidas de tal forma a ser a totalidade máxima espacialmente conectada, então, os processos SPAN são obrigados a ser totalidades máximas espaço-temporalmente conectados. Isso significa antes de tudo, que os processos não envolvem lacunas temporais.

Um determinado processo não pode ocorrer em dois momentos distintos, sem que ocorram também em todo o tempo de intervalo entre eles. Da mesma forma, um processo não pode ocupar duas regiões espaços-temporais separadas ao longo da dimensão espacial, sem ocupar também uma sequência contínua de regiões espaço-temporal, formando uma ponte entre eles. A linha entre processos e processuais de outros tipos, no entanto, é claramente difícil de desenhar. O que é classificado como

processo e como o conjunto de processos, muitas vezes depende do nível escolhido de granularidade.

Partes Fiat

As partes de processos em um determinado nível de granularidade são partes Fiat. Exemplos de partes fiat de processos são: o processo individual de mastigar no curso de uma refeição, a passagem do bolo alimentar da boca ao estômago. Quando se passa para um nível mais baixo de granularidade, no entanto, muitas vezes torna-se possível reconhecer as parte bona fide: as contrações coordenadas dos músculos individuais no ato de comer.

Eventos

Eventos, conforme definido por (Grenon (2003) e Grenon e Smith (2004), são os limites temporais instantâneos fiat e bona fide de processos. Exemplos: a formação de uma sinapse, o início do sono REM, o descolamento de um dedo em um acidente industrial. Estes autores sugerem que eles são análogos temporais em SPAN dos limites espaciais de substâncias no SNAP. Para Grenon e Smith (2004), algumas entidades processuais são apreendidas como mudanças, porque envolvem súbitos qualitativos ou outros saltos que parecem dividir a realidade temporal em duas partes. Eventos instantâneos, por exemplo, a passagem do 13 ° para o 14 ° ano de vida, são muitas vezes criados via Fiat.

Agregados de Processos

Exemplos de agregados de processos são: o agregado de todos os disparos neuronais envolvidos na formação de um pensamento; o agregado de todas as medidas tomadas no decurso de uma maratona; o agregado de todas as guerras e conflitos étnicos na África Central em um determinado ano; o agregado de todos os movimentos de tropas na selva em determinada noite.

Configurações (ambiente espaço-temporal).

Exemplos de configurações são: uma operação cirúrgica como cenário para uma infecção nasocomial e uma rotina de *check-up* como um ajuste para a descoberta de um tumor.

Regiões Espaço-Temporal

A totalidade das regiões espaço-temporal reflete a totalidade das demarcações *fiat* possíveis daquela máxima região denominada espaço-tempo. Incluem-se nas regiões espaço-temporais, as extensões espaço-temporal das entidades processuais. Entidades processuais são para as regiões espaço-temporais em ontologias SPAN, de forma análoga à maneira pela qual substanciais estão para as regiões espaciais na ontologia SNAP.

Regiões espaço são entidades que existem em seu próprio direito, no sentido de que elas existem independentemente de quaisquer entidades processuais que podem ser localizadas no/ou nos mesmos.

2.5.2 Construção de ontologias: uma breve descrição de metodologias

A presente seção se valeu dos estudos apresentados pelos autores, no qual procurou verificar, nas metodologias estudadas, a atividade de AC na construção de ontologias. Diferentes metodologias para construção de ontologias têm sido propostas nos últimos 20 anos. Almeida (2002), Almeida e Bax (2003) e Almeida (2006) apresentam propostas identificadas na literatura científica das áreas estudadas. Esses autores se basearam nos estudos de Corcho, Fernández-López e Gomez-Pérez (2001) e apresentam diferentes metodologias, proporcionando uma visão geral de seu funcionamento. O QUADRO 6 a seguir, apresenta uma breve descrição de tais metodologias. (ALMEIDA, 2002; ALMEIDA; BAX, 2003; ALMEIDA, 2006).

QUADRO 5: Breve descrição das metodologias

Metodologia	Breve descrição
<i>Cyc</i>	Codifica manualmente o conhecimento implícito e explícito das diferentes fontes e, quando já se tem conhecimento suficiente na ontologia, um novo consenso pode ser obtido por ferramentas que utilizam linguagem natural (LENAT; GUHA,1990).
USCHOLD e KING	Identifica o propósito, os conceitos e relacionamentos entre os conceitos, além dos termos utilizados para codificar a ontologia e, em seguida, documentá-la (USCHOLD;KING, 1996).
GRÜNINGER e FOX	Método formal que identifica cenários para uso da ontologia utiliza questões em linguagem natural para determinação do escopo da ontologia, executa a extração sobre os principais conceitos, propriedades, relações e axiomas, definidos em PROLOG (GRÜNINGER; FOX, 1995).
<i>KACTUS</i>	Método recursivo que consiste em uma proposta inicial para uma base de conhecimento; quando é necessária uma nova base em domínio similar, generaliza-se a primeira base em uma ontologia adaptada a ambas aplicações; quanto mais aplicações, mais genérica a ontologia (BERNARAS; LARESGOITI ; CORERA, 1996).
<i>Methontology</i>	Constrói uma ontologia utilizando-se o conhecimento do domínio, cujas atividades principais são especificação, conceitualização, formalização, implementação e manutenção, (FERNÁNDEZ-LÓPEZ <i>et al.</i> , 1999).
<i>Sensus</i>	Constrói ontologias a partir de outras ontologias, identificando os termos relevantes para o domínio e ligando-os à ontologia mais abrangente (<i>Sensus</i> , com 50.000 conceitos); um algoritmo monta a estrutura hierárquica do domínio (SWARTOUT <i>et al.</i> , 1996)
<i>On-to-knowledge</i>	Auxilia a administração de conceitos em organizações, identificando metas para as ferramentas de gestão do conhecimento e utilizando cenários e contribuições dos provedores / clientes de informação da organização (STAAB <i>et al.</i> , 2001).

Fonte: Adaptado de Almeida (2002), Almeida e Bax (2003), Almeida (2006).

Ainda não há uma metodologia aceita consensualmente, mas sim tentativas de explicitar ou documentar uma variedade de técnicas (ALMEIDA, 2002; ALMEIDA; BAX, 2003; ALMEIDA, 2006). Para verificar a utilidade das metodologias e compará-las seria necessário avaliar a ontologia resultante da aplicação de cada metodologia. Além disso, de acordo com Almeida e Bax (2003), não parece provável a unificação das propostas em uma única metodologia.

Uma vez que o presente trabalho se localiza no contexto de ontologias biomédicas, relevante considerar as observações de Castro *et al* (2006), para quem um dos principais obstáculos no desenvolvimento de ontologias é a falta de uma metodologia unificada. Diferentes metodologias têm sido propostas para cenários diferentes, mas não existe uma metodologia com padrão acordados para construção de

ontologias. Para estes autores, o envolvimento de especialistas de domínio geograficamente distribuídos, a necessidade de especialistas de domínio para liderar o processo de desenvolvimento, a aplicação das ontologias e os ciclos de vida de bio-ontologias estão entre os recursos não considerados por metodologias anteriormente propostas.

Considerando tais abordagens, a atividade de AC é apresentada como parte de um conjunto de atividades que compõem o próprio processo de construção da ontologia, conforme abordado por Fernandez, Gomez-Perez e Juristo (1997). Neste contexto, esses autores fazem referência à METHONTOLOGY. Trata-se de uma metodologia baseada no padrão IEEE (1995), como sendo um *framework* testado, que permite definir estágios e atividades a serem realizadas na construção de uma ontologia a partir do zero. Tais estágios e atividades que compõem a *Methontology* constituem o ciclo de vida da ontologia e encontram-se descritas a saber:

- *especificação*: objetiva a elaboração de um documento escrito em linguagem natural, que contém informações a respeito da ontologia;
- *conceitualização*: tratada estruturação do domínio do conhecimento, em um modelo conceitual. Descreve problemas e soluções relativos ao vocabulário do domínio, que é identificado na atividade de especificação;
- *formalização*: o modelo conceitual criado é transformado em um modelo formal, representado por meio de uma linguagem formal;
- *integração*: integração da ontologia em construção às outras já existentes;
- *implementação*: o modelo conceitual gerado é implementado de forma computável;
- *avaliação*: trata da avaliação da ontologia em si e deve considerar os processos de verificação e validação;
- *aquisição de conhecimento*: busca obter conhecimento em possíveis fontes de conhecimentos, tais como especialistas do domínio ou a partir de textos, base de dados, entre outros.
- *documentação*: auxilia na possível manutenção e facilita a reutilização;
- *manutenção*: constituem as alterações quando necessárias, para possíveis melhorias ou correções (FERNANDEZ; GOMEZ-PEREZ; JURISTO, 1997, p.35).

A FIG. 11 representa o ciclo de vida da ontologia, conforme apresentado pelos autores.

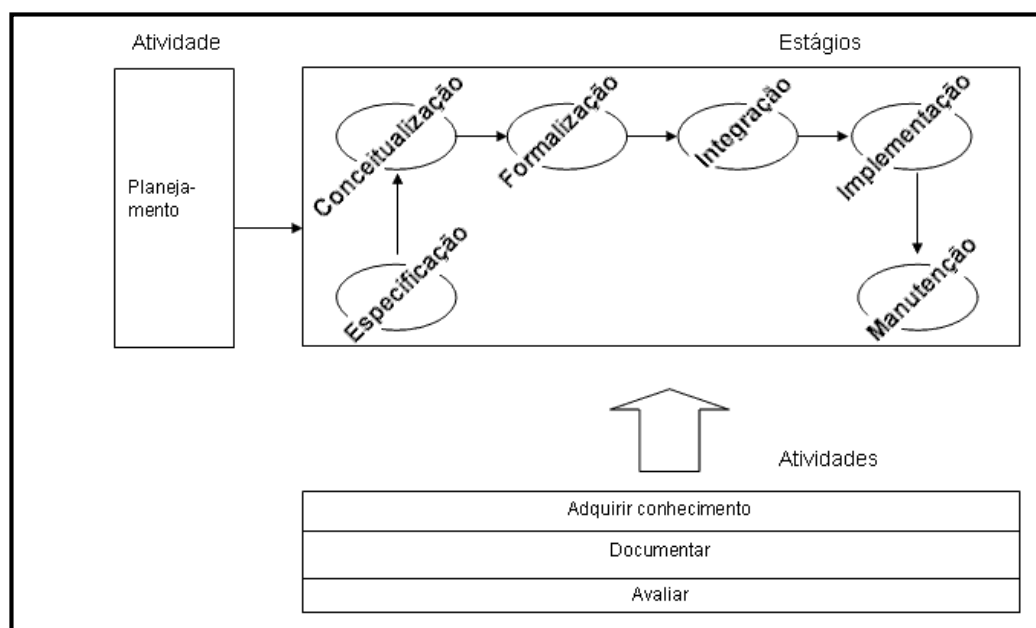


FIGURA 11: Representação gráfica do ciclo de vida da ontologia
Fonte: Adaptado de Fernandez, Gomez-Perez e Juristo (1997, p.35)

Embora a atividade de AC, para obtenção de conhecimento de um dado domínio, ocorra simultaneamente às demais, notou-se que tal atividade se encontra descrita de forma superficial. Não foi encontrado, na literatura consultada, detalhamentos a respeito desta etapa, mesmo que referenciada como sendo essencial e relevante para o processo de construção de ontologias (FERNÁNDEZ-LÓPEZ; GÓMEZ-PÉREZ; JURISTO, 1997; FERNÁNDEZ-LÓPEZ, 1999; FERNÁNDEZ-LÓPEZ; GÓMEZ-PÉREZ, 2002; CORCHO *et. al*, 2003). Considera-se, dessa forma, que aprofundar na temática da AC, a fim de buscar melhorias, bem como identificar problemas decorrentes desta atividade, é caminho a seguir para a devida representação do conhecimento, via ontologias, de um dado domínio.

Na busca de tal ação, partiu-se da premissa que considera que ontologias são construídas visando um vocabulário único, baseado em termos definidos a partir do consenso de membros de determinada comunidade (ALMEIDA; BAX, 2003). Assim, é relevante para os propósitos da presente pesquisa, que se identifique o domínio do conhecimento e a comunidade que o constitui – considerando tal comunidade aquela

com respaldo para falar sobre o domínio, de modo a obter desta comunidade, os termos relevantes acerca do domínio. Tal respaldo é conferido pelo caráter científico da comunidade, uma vez que, ontologias bem fundamentadas não são construídas a partir do senso-comum, mas por membros de um campo científico (SMITH, 2003).

2.5.3 Considerações complementares sobre AC para construção de ontologias

Ao considerar que ontologias bem fundamentadas são construídas a partir da participação de membros de uma comunidade, que possuem respaldo para falar a respeito de determinado domínio, é necessário então que se defina tanto o domínio quanto os membros da comunidade que o constitui. Para tal, a subseção 2.5.3.1, aborda a análise de domínio e a subseção 2.5.3.2 apresenta elementos essenciais acerca do especialista e sua expertise.

2.5.3.1 A análise do domínio

A Análise de Domínio (AD) é discutida desde a década de 1980 na Engenharia de Software. De acordo com Prieto-Diaz (1990), esse termo foi empregado pela primeira vez por Neighbors, em 1981, ao se referir à atividade de identificar objetos, operações e relações de uma classe de sistemas similares em um determinado domínio de problema. Nessa atividade, os peritos do domínio tinham a função de informar o que percebiam como importante.

Sob esta perspectiva, Pietro-Diaz (1990) considera a AD como sendo “o processo pelo qual a informação usada no desenvolvimento de software é identificada, capturada e organizada com a finalidade de torná-la reutilizável em novos sistemas”. Ela trata do desenvolvimento e evolução de uma infraestrutura de informação para apoiar a reutilização, cujos componentes incluem modelos de domínio; padrões de desenvolvimento; e repositórios (bibliotecas) de componentes reutilizáveis (PIETRO-DIAZ, 1990).

Em Ciência da Informação, a AD e respectivos desdobramentos foram apresentados por Hjørland e Albrechtsen (1995), no qual apresentam Bates; Amba e Iyer; Taylor e Bliss (1870-1955) como os precursores que contribuíram com

investigações sobre esta temática na área (HJØRLAND; ALBRECHTSEN, 1995). Hjørland e Albrechtsen (1995) apresentam ainda Saracevic, numa referência aos trabalhos sobre relevância e análise de assunto e Ranganathan (1892-1972) ao considerarem a prática da teoria da classificação, como tipos de AD.

Hjørland & Albrechtsen (1995) consideram a análise de domínio na CI uma abordagem teórica que busca entender as áreas através da análise das comunidades de discurso, que são partes da divisão da sociedade do trabalho. Para esses autores, a organização do conhecimento, a estrutura, os padrões de cooperação, a linguagem, bem como as formas de comunicação e os sistemas de informação são reflexos do trabalho dessas comunidades e do seu papel na sociedade.

Entende-se uma comunidade de discurso, a partir das seguintes características (SWALES, 1986), a saber: i) *metas comuns*: um conjunto comum de objetivos que se apresentam em documentos ou em conhecimento tácito; ii) *mecanismos participativos*: formas de intercomunicação entre os membros por meio de encontros, correspondência, conversas, entre outros; iii) *troca de informação*: mecanismos utilizados para prover informação com propósitos definidos; iv) *estilos específicos*: uso de um ou mais estilos de comunicação para atingir seus objetivos, identificados por tópicos de discussão, forma, posição de elementos e mensagens; v) *terminologia especializada*: uma comunidade discursiva tem um vocabulário específico; vi) *alto nível de especialização*: uma comunidade discursiva tem um mínimo de membros com um nível adequado de conhecimento relevante e *expertise* discursiva.

Hjørland (2002) propõe ainda o estudo de domínios do conhecimento, a partir de onze abordagens para AD, a saber:

1. Produção e avaliação de guias de literatura e entradas de assunto;
2. Produção e avaliação de classificações e tesauros;
3. Pesquisa sobre competências na indexação e recuperação de informações especializadas;
4. Estudos de usuários empíricos em áreas de assunto;
5. Produção e interpretação de estudos bibliométricos;
6. Estudos históricos de estruturas e serviços de informação em domínios;
7. Estudos de documentos e gêneros em domínios de conhecimento;

8. Estudos epistemológicos e críticos de diferentes paradigmas, pressupostos e interesses em domínios;
9. Estudos terminológicos, linguagens para fins específicos, e análise do discurso em campos de conhecimento;
10. Estudos das estruturas e instituições na comunicação científica e profissional em um domínio.
11. Métodos e resultados de estudos analíticos de domínios sobre cognição profissional, representação do conhecimento em ciência da computação e inteligência artificial.

Tennis (2003) acrescenta dois dispositivos analíticos complementares para a AD: “Áreas de modulação” e “Graus de especialização”. “[...] numa análise de domínio, estes eixos podem servir de instrumentos de análise para se delinear o que está sendo estudado e que não está sendo estudado”. Para esse autor, a noção de domínio deve ser definida, de modo que seja transferível, ou seja, pode ser usada por mais de um pesquisador, de modo a permitir uma compreensão comum do que seja o objeto da análise de domínio.

De acordo com Tennis (2003), as “*áreas do eixo modulação*” definem os parâmetros para os nomes e a extensão do domínio, sendo essa última seu alcance total. A extensão “afirma o que está ou não está incluído e do que é chamado o domínio”. De acordo com esse autor, as áreas do eixo de modulação fazem isso através da negociação dos termos e suas definições utilizadas por membros do domínio. Esse eixo concilia a questão: *Qual é o domínio? E o que ele cobre?* Ambas necessárias para se definir os parâmetros sobre um domínio.

O detalhe de como um domínio é organizado sob sua extensão e respectivos nomes fazem parte do eixo, “*graus de especialização*”, que qualificam e definem a intensão de um domínio. De acordo com Tennis (2003), pode ser indesejável ou inviável descrever um domínio inteiro. Isso porque um domínio como um todo, mesmo que tenha um nome e uma extensão, torna difícil a sua análise. Esse autor sugere que, ao qualificar um domínio, sua extensão seja diminuída e sua intensão aumentada.

Sob essa perspectiva, Tennis (2003) argumenta que estudar o hinduísmo não significa estudar todas as religiões. Ou seja, o domínio qualificado é o hinduísmo.

Desse modo, o hinduísmo tem uma maior intensão e uma menor extensão em relação à religião. “[...] hinduísmo como parte da religião é uma qualificação da religião. Isso significa mais especificamente um tipo de religião”. O hinduísmo também poderia ser qualificado pela história ou como um grupo político qualificado. No entanto, conforme sugere esse autor, nem todas as qualificações são facilmente aninhadas,

Como forma de se proceder a análise e identificação do domínio do presente trabalho, este se valeu das abordagens apresentadas por Hjørland (2002), especialmente a produção e interpretação de estudos bibliométricos, bem como lança mão dos dois dispositivos analíticos: “*áreas de modulação e graus de especialização*” para a AD, apresentados por Tennis (2003). Vale ressaltar que os demais elementos apresentados por Hjørland (2002), ainda que importantes para as devidas competências de especialistas da informação, não foram empregados.

Como observado, ao considerar a literatura, bem como a comunidade que a produziu, os estudos bibliométricos foram apresentados por Hjørland (2002), como um elemento que possibilita compreender um domínio. Segundo esse autor, a bibliometria contribui, uma vez que ela pode auxiliar no entendimento do modo como, por exemplo, uma área se desenvolveu.

A bibliometria, segundo Guedes e Borschiver (2005), é “um conjunto de leis e princípios empíricos que contribuem para estabelecer os fundamentos teóricos da Ciência da Informação” e de outros campos do conhecimento. De acordo com os autores, a expressão bibliometria é originada do termo “*bibliografia estatística*”,²⁴ empregado por Hulme em 1922, numa conotação de esclarecimento dos processos científicos e tecnológicos, por meio da contagem de documentos. O termo “bibliometria” foi criado por Otlet em 1934 no seu “*Traité de Documentation*”, sendo esse termo utilizado até os dias de hoje (VANTI, 2002; ARAÚJO, 2006).

Segundo Araújo (2006), a bibliometria tem a finalidade de capturar dados em suas respectivas fontes para tratamentos estatísticos, baseados em mapeamento e geração de diferentes indicadores. Partindo de tais princípios, surgiram leis que buscavam subsidiar as análises, a saber: “*Lei da produtividade de periódicos*” (Lei de Bradford ou Lei de Dispersão), “*Lei da produtividade de autores*” (Lei de Lotka ou

²⁴Statistical bibliography (tradução nossa)

Lei do Quadro Inverso) e “*Lei da frequência de ocorrência de palavras*” (Leis de Zipf ou Lei do Mínimo Esforço) (VANTI, 2002; GUEDES; BORSCHIVER, 2005; ARAUJO, 2006). Não foi realizado estudo aprofundado de tais leis, buscou-se apresentá-las de modo a contextualizar seus princípios, com o qual a análise de citações é parte integrante.

Araújo (2006) apresenta a análise de citações como a mais importante área da bibliometria. Forest (1989) *apud* Araújo (2006) explica que citação é um “[...] conjunto de uma ou mais referências bibliográficas que, incluídas em uma publicação, evidenciam elos entre indivíduos, instituições e áreas de pesquisa, visto que mostram o relacionamento de uma publicação com outra”. A análise de citação pode ser definida como “[...] a parte da bibliometria que investiga as relações entre os documentos citantes e os documentos citados considerados como unidades de análise, no todo ou em suas diversas partes: autor, título, origem geográfica, ano e idioma de publicação etc.” (FORESTI, 1989, p. 3 *apud* ARAUJO, 2006).

A análise de citações (ARAUJO, 2006) permite uma série de padrões de identificação e descrição na produção do conhecimento científico.

“Com os dados retirados das citações pode-se descobrir: autores mais citados, autores mais produtivos, elite de pesquisa, frente de pesquisa, fator de impacto dos autores, procedência geográfica e/ou institucional dos autores mais influentes em um determinado campo de pesquisa; tipo de documento mais utilizado, idade média da literatura utilizada, obsolescência da literatura, procedência geográfica e/ou institucional da bibliografia utilizada; periódicos mais citados, “core” de periódicos que compõem um campo” (ARAUJO, 2006, p.20).

Nesse contexto, índices foram criados para “avaliar a produtividade e a qualidade da pesquisa de cientistas, por meio da medição com base nos números de publicações e citações de diversos pesquisadores” (VANTI, 2002, p.155), bem como para a avaliação da importância dos periódicos. Segundo Vanti (2002), a generalização das estatísticas empíricas pela bibliometria acontece por meio do exame das relações entre diferentes variáveis como: recursos humanos-documentos, artigos-periódicos, produção-consumo etc., que apresentam diversas regularidades de distribuição. Desse modo, o número de artigos que originam n citações, o número de instituições produzindo anualmente n doutorados, o número de autores com n artigos, o número de

revistas contendo n artigos constituem exemplos do mesmo tipo de distribuição (VANTI, 2002, p. 155).

Igami (2011) considera as seguintes distinções de indicadores bibliométricos: i) indicadores-produto (ou ainda indicadores de eficácia) para resultados mais imediatos das políticas com a produção de artigos em Ciência e Tecnologia (C&T) ou número de patentes; ii) indicadores de impacto (ou indicadores de efetividade social) quando se referem a desdobramentos a médio prazo ou a efeitos mais abrangentes e perenes do fomento às atividades de C&T, como o Fator de Impacto de Publicações e outras medidas – não bibliométricas – como a Taxa de Inovação Tecnológica, o Balanço de Pagamentos Tecnológico, o grau de apropriação de tecnologia nacional na produção de medicamentos, na Saúde Pública, no desenvolvimentos de novos materiais para construção de moradias, entre outros.

A produção de indicadores bibliométricos se concretizou nas últimas décadas do Século XX, em função da criação, manutenção e informatização de bases de dados para armazenamento e consulta de informação científica (IGAMI, 2011). As bases de dados compiladas pelo *Institute for Scientific Information (ISI): Science Citation Index, Social Science Citation Index e Arts & Humanities Citation Index* encontram-se reunidas na *Web of Science* (VANTI, 2002; IGAMI, 2011). De acordo com Vanti (2002), o ISI foi o pioneiro na disponibilização do acesso a informações referenciais sobre parte significativa da produção científica mundial e um dos principais agentes incentivadores do uso de indicadores bibliométricos para avaliação da produtividade científica.

Outras bases de dados que indexam citações surgiram ao longo do tempo. Entre essas bases, destacam-se: i) *Scientific Electronic Library Online (SciELO)*²⁵: modelo para a publicação eletrônica de periódicos científicos na Internet, desenvolvido para responder às necessidades da comunicação científica nos países em desenvolvimento, especialmente da América Latina e Caribe; ii) *Scopus*²⁶: base de dados desenvolvida e mantida pela Elsevier, que apresenta uma cobertura maior de publicações em termos quantitativos, na qual incluem a produção dos países em desenvolvimento.

²⁵ Disponível em <http://www.scielo.br/>

²⁶ Disponível em <http://www.scopus.com/home.url>

O “*Fator de Impacto*” (FI), desenvolvido por Garfield (1998) ²⁷, foi apresentado em 1963 na publicação da primeira edição do *Science Citation Index (SCI)*. Trata-se de um indicador bibliométrico publicado anualmente desde 1972 no *Journal of Citation Reports (JCR)*, pelo *ISI-Thomson Scientific*. O Fator de impacto identifica a frequência com que determinado artigo de um periódico é citado no decorrer de um ano. “*Este número pode ser utilizado para avaliar ou comparar a importância relativa de um periódico com outros do mesmo campo ou ver com que frequência os artigos são citados, para determinar quais periódicos são melhores para uma determinada coleção*” (SOARES, 2008). Este índice é calculado com base em citações feitas por publicações predominantemente internacionais, indexadas pela *Web of Science*²⁸.

Por sua vez, o “*Índice H*” mais recentemente criado e proposto pelo físico Hirsch, em 2005, na qual sugeria: “Proponho índice h definido como o número de artigos com número de citações maior ou igual a h, como sendo um índice útil para caracterizar a produção científica de um pesquisador” (tradução nossa). Diferente do Fator de Impacto, que avalia frequência de citações de determinado periódico, o índice h foi criado como uma proposta para avaliar a produção de um pesquisador, um departamento ou mesmo um grupo de pesquisa.

Em seu artigo de criação do índice h, Hirsch (2005) fazia a seguinte indagação:

“Para os poucos cientistas que ganham um prêmio Nobel, o impacto e a relevância de suas pesquisas é inquestionável. Entre o resto de nós, como quantificar o impacto cumulativo e a relevância da produção de um indivíduo em sua investigação científica? Em um mundo de recursos limitados, essa quantificação (mesmo que potencialmente desagradável é frequentemente necessária para fins de avaliação e comparação (por exemplo, para recrutamento de professores universitários e avanço, concessão de subsídios, entre outros)).” (tradução nossa). (HIRSCH, 2005, p.16.570)

Conforme observa Hirsch (2005), o registro de publicação de um indivíduo e o registro de sua citação são dados que contêm informações úteis. Informações que incluem o número de artigos publicados (N_p) ao longo dos anos n ; o número de

²⁷ Fundador do *Institute for Scientific Information (ISI)*, hoje pertencente a *Thomson Reuters Corporation*.

²⁸ Base multidisciplinar, consultada a partir do Portal de periódicos Capes, que indexa somente os periódicos mais citados em suas respectivas áreas.

citações (Nc j) para cada artigo (j), as revistas onde os trabalhos foram publicados, o seu parâmetro de impacto. Para esse autor, esta grande quantidade de informações deverá ser avaliada sob diferentes critérios e por diferentes pessoas.

Vanti (2002) chama atenção para certas condições que podem influenciar os índices de citações, como: idioma, histórico do periódico e seu formato, a frequência de publicação e sua área de especialização. Este autor indaga que aqueles periódicos não publicados em inglês ou que não utilizam o alfabeto romano podem ser menos acessíveis aos pesquisadores do mundo, o que pode influenciar seus níveis de citação. Deve-se levar isso em consideração, quando analisar citações de diferentes periódicos.

A discussão aqui apresentada enfocou alguns elementos para a compreensão de domínios. Tal ação buscou meios de subsidiar o uso da literatura e o registro científico produzido por uma comunidade, como parâmetro para o estabelecimento dos limites e definição desse domínio e o reconhecimento pelo contexto, do qual é parte. Somando-se a esses elementos também são necessários parâmetros para subsidiar a escolha dos especialistas, uma vez que selecionar especialistas (CORNELISSEN *et al*, 2003) é uma forma de garantir a elicitación adequada do conhecimento especializado, de modo que seja possível garantir a qualidade do conhecimento necessário para os propósitos da elicitación.

2.5.3.2 Especialista e expertise

Procurou-se caracterizar os especialistas, bem como à *expertise*, de modo a esclarecer melhor a forma como esses foram abordados no presente trabalho. O trabalho seminal de Groot em 1965 é citado por Gruber (2001) como sendo o pioneiro na orientação científica da pesquisa sobre *expertise*. De acordo com este autor, Groot observando mestres do xadrez verificou que a diferença marcante entre os grandes mestres e jogadores novatos havia sido revelada em uma tarefa de memória, na qual as posições do xadrez foram apresentadas por alguns segundos e, em seguida, foi solicitado que fossem imediatamente reconstruídas. Groot então observou que o retorno dos especialistas era superior e, explicado a partir de estruturas específicas de percepções que tinham em memória e que estavam intimamente relacionadas ao conhecimento específico do domínio.

Conforme exposto por Gruber (2001), a interpretação desses achados por Groot guiou pesquisas futuras sobre conhecimento, a partir da perspectiva da teoria de processamento de informação e psicologia cognitiva. O foco na análise especializada de correlação de desempenho cognitivo - percepção, memória, conhecimento, solução de problemas, desde então tem sido mantido e só recentemente complementado por pesquisas sobre os contextos no qual o desempenho de especialistas está situado. Três temas principais de investigação podem ser derivados a partir desta orientação: i) modelagem das estruturas cognitivas e os mecanismos que caracterizam processamento de informações de especialistas, ii) descrição de aquisição da perícia; c) estudo dos meios instrucionais para apoiar a aquisição de conhecimentos (GRUBER, 2001, p.5.146).

Gruber (2001) utiliza o termo "*expertise*" para se referir ao desempenho de um indivíduo em um domínio específico como medicina, física, xadrez e música. "*Especialistas*" são pessoas que, por critérios objetivos e ao longo do tempo, consistentemente mostram um desempenho superior em atividades típicas de um domínio. O desempenho de um especialista muitas vezes é elucidado por meio de comparações com indivíduos de desempenho mais limitado. Tais indivíduos são chamados "*novatos*". A especialização geralmente é considerada como habilidade especial ou conhecimento adquirido de um determinado assunto, através da experiência prática (GRUBER, 2001, p. 5.147).

Corroborando com essa visão, Cornelissen *et al* (2003) consideram especialista aquele indivíduo cujo conhecimento acerca de um dado domínio é obtido gradualmente, através de um período de tempo de aprendizado e experiência. Difere do não especialista pelo modo que este representa e agrupa seus conhecimentos e pela percepção holística que tem relação aos problemas (BRADLEY *et al*, 2006). Conforme observa Cornelissen *et al*, (2003), um novato embora seja mais focado, possui pouca experiência passada para ajudar na formação de hipóteses. Por outro lado, o especialista considera outros fatores, como fatos anteriores para desenvolver um plano para proceder à investigação.

Relevante considerar a observação de Ford e Sterman (1998), para os quais o conhecimento especializado é influenciado por perspectivas e objetivos individuais. Assim, esses autores consideram relevante para a seleção de especialistas, que se

decida, pela formação de um grupo heterogêneo (por exemplo, cientistas e médicos) ou por um grupo homogêneo (por exemplo, apenas cientistas). [...] “as diferenças nas experiências pessoais afetam menos um grupo homogêneo, mas a vantagem do grupo heterogêneo de especialistas está no fato que este é compensado por diferentes pontos de vista” (FORD; STERMAN, 1998).

Em resumo, os seguintes critérios para identificar especialistas se baseiam em (CORNELISSEN *et al*, 2003): i) período de aprendizagem e experiência de uma pessoa em um domínio do conhecimento, que por sua vez influencia seu comportamento crítico e analítico. Assim, devem ser considerados, o número de publicações científicas de uma pessoa sobre o assunto, o envolvimento em debates públicos sobre o tema, ou o período de aprendizagem e experiência. ii) na escolha entre um grupo heterogêneo ou homogêneo de especialistas, cujos critérios podem ser usados tanto para se escolher o especialista, quanto seus pares.

Hoffman (1995) explica que na prática a escolha de especialistas não tem sido problema, uma vez que estes têm sido selecionados com base nos anos de experiência e em critérios profissionais como nível de escolaridade, treinamento e experiência, registros de publicação, membros de sociedades profissionais, para citar alguns. De acordo com Hoffman (1995), apenas recentemente é que foram desenvolvidos amplamente critérios baseados em métodos empíricos, úteis e sistemáticos para a identificação de candidatos à AC.

Para Hoffman (1995), *expertise* não é uma categoria única, uma vez que a sua definição não deve extrair dessa palavra a sua riqueza conceitual e de dependência contextual, mas sim apontar para operacionalizações. Além disso, propõe esse autor que, questões sobre o modo que indivíduos são selecionados para treinamentos, sobre como a *expertise* é constituída e como é exercitada dependem do domínio. Níveis distintos são definidos de modo a fornecer um contexto com as quais as categorias de *expertise* se aplicam. Conforme sugere Hoffman (1995), tais níveis são um modo de apontar para critérios operacionalizáveis que abranjam fatores gerais, experimentais, sociais, cognitivos e relacionados ao desempenho, conforme apresentado no QUADRO 6 a seguir:

QUADRO 6: Categorias dos níveis da *expertise*

Nível	Descrição
Ingênuo	Aquele totalmente ignorante em relação ao domínio
Novato	Um novato que se encontra numa fase de experiência. Possui exposição mínima no domínio
Iniciado	Alguém que já tenha passado por uma iniciação. Um novato com instrução introdutória.
Aprendiz	Aquele que está aprendendo. Um estudante submetido a um programa de nível introdutório. Tradicionalmente um aprendiz é imerso no domínio pela convivência ou assistindo alguém com alto nível. A duração do aprendizado depende do domínio.
Trabalhador	Uma pessoa que pode desempenhar um dia de trabalho sem supervisão, embora trabalhe sob ordens. Um trabalhador experiente e de confiança, ou aquele que tem adquirido um nível de competência.
Especialista	Trabalhador brilhante ou expressivo, altamente considerado pelos pares, cujos julgamentos são notavelmente acurados e confiáveis, no qual o desempenho mostra habilidade consumada e economia de esforço, e lida efetivamente com casos raros ou difíceis. Também é alguém com habilidade especial ou conhecimento derivado de experiência extensiva em subdomínios.
Mestre	Trabalhador ou especialista que também é qualificado para ensinar aqueles de nível inferior. Tradicionalmente é alguém de um grupo de elite de especialistas, cujos julgamentos definem regulamentos, padrões ou ideais. Também pode ser um especialista considerado por outros especialistas como sendo “o especialista” ou especialista real, especialmente considerado no subdomínio do conhecimento.

Fonte: Adaptado de Hoffman (1995)

Hoffman (1995) lembra que em estudos em que não está claro como a *expertise* foi operacionalizada ou como os especialistas foram identificados, estes devem ser considerados experientes. Em outros casos, lembra o autor, o termo “participante” pode ser utilizado como alternativa, por exemplo: “participantes eram trabalhadores ou aprendizes”. Esse cuidado foi apresentado por Hoffman (1995), uma vez que tarefas ou problemas que desafiariam o especialista “mestre” poderiam ser incompreensíveis pelo novato, e problemas que novatos podem compreender pode ser trivial ao especialista.

Embora não exista uma lista definitiva de critério, se os critérios escolhidos são formulados qualitativamente (CORNELISSEN *et al*, 2003), a contribuição importante é que a base sobre a qual os especialistas são selecionados deve ser transparente e pública.

A revisão de literatura discutida no presente capítulo sugere uma abordagem à AC multidisciplinar, uma vez que métodos e técnicas provenientes de diferentes disciplinas são combinados de modo a se criar um método adequado ao domínio que se pretende explorar. Desse modo, para investigar as atividades de AC no campo da Biomedicina,

propõe-se um roteiro metodológico em que o conhecimento médico é obtido e validado. Para tal, o Capítulo 3 descreve a metodologia proposta para a presente pesquisa, o contexto em que se desenvolvem atividades de AC e o o roteiro para AC em Biomedicina no campo da hematologia. Por sua vez, o Capítulo 4 descreve, a partir do roteiro metodológico proposto, a coleta de dados e respectivos resultados, no qual apresenta as principais dificuldades levantadas no decorrer da atividade de eliciação de conhecimento dos especialistas no domínio em estudo.

3 METODOLOGIA DE PESQUISA

A metodologia não procura soluções, mas contribui na escolha das maneiras de encontrá-las, integrando os conhecimentos adquiridos sobre os métodos em vigor nas diferentes disciplinas científicas ou filosóficas
BARROS & LEHFELD.

A atividade de pesquisa necessita de método a fim de orientar o pesquisador na construção de quadros teóricos do conhecimento. Marconi e Lakatos (2005) definem o método científico como um conjunto de atividades sistemáticas e racionais que, com maior segurança e economia, permitem alcançar o objetivo – conhecimento válido e verdadeiro – traçando o caminho a ser seguido, detectando erros e auxiliando as decisões do cientista. Seguindo orientações de Marconi e Lakatos (2005), propõe-se a presente pesquisa da seguinte maneira:

- Do ponto de vista de sua *natureza*, trata-se de uma pesquisa *aplicada*, uma vez que objetiva gerar conhecimentos para aplicações práticas dirigidas à solução de problemas específicos;
- Quanto à abordagem ao problema, considera-se pesquisa qualitativa, visto que o conhecimento apreendido, no âmbito da atividade de AC para representação em ontologias, não pode ser mensurado quantitativamente;
- Do ponto de vista de seus *objetivos*, caracteriza-se como pesquisa *exploratória*, uma vez que procura maior familiaridade com o problema, com vista a explicitá-lo;
- Do ponto de vista dos procedimentos técnicos, classifica-se como pesquisa bibliográfica, pois foi elaborada a partir de material já publicado, constituído principalmente de artigos de periódicos, livros e material científico disponibilizado na Internet, como teses e dissertações; e como pesquisa *ex-post-fact*, pois o roteiro construído foi aplicado em campo.

O restante do presente capítulo está organizado nas seguintes seções: a seção 3.1 fornece informações gerais sobre a instituição objeto da pesquisa, descrevendo seu papel como instituição pública e no campo científico, apresenta ainda a motivação pelo desenvolvimento da pesquisa na instituição. A seção 3.2 discorre sobre a condução e o

recorte da pesquisa. A seção 3.3 descreve de forma genérica a metodologia de pesquisa adotada, apresentando os instrumentos de pesquisa, os métodos para coleta de dados e instrumentos para coleta de dados e a seção 3.4 apresenta os instrumentos de pesquisa adotados.

3.1 A instituição objeto da pesquisa

O presente trabalho foi desenvolvido na Fundação Centro de Hematologia e Hemoterapia de Minas Gerais – Fundação Hemominas, daqui em diante denominada simplesmente Hemominas. Essa instituição presta serviço na área de hematologia e hemoterapia, desenvolvendo atividades de assistência médica, ensino, desenvolvimento tecnológico, produção, controle de qualidade, educação sanitária, além de ser reconhecida no campo da pesquisa, contribuindo com o conhecimento científico em diferentes áreas da biomedicina.

A instituição possui um corpo diretivo e técnico de diferentes áreas do conhecimento e formado por pesquisadores (mestres, doutores e pós-doutores) voltados para o desenvolvimento no campo científico.

3.1.1 Histórico, missão e visão institucional.

A Hemominas foi criada em 1985, como unidade integrante da Fundação Hospitalar do Estado de Minas Gerais (Fhemig), tendo por objetivo implantar e organizar o sistema hematológico e hemoterápico em Minas Gerais. Em 1989, por meio da Lei nº 10.057, passou a ter personalidade jurídica de direito público, sob a forma de fundação vinculada à Secretaria de Estado da Saúde. A Hemominas tem como “*Missão*” atuar na área de hematologia, hemoterapia, células e tecidos com excelência e responsabilidade social. Sua “*visão*” é ser reconhecida como organização de excelência mundial em serviços de saúde.

De acordo com o art. 2º da lei supracitada, a Hemominas tem como finalidade:

“assegurar unidade de comando e direção às políticas estaduais relativas à hematologia e hemoterapia, garantindo à população a oferta de sangue e hemoderivados de boa qualidade, desenvolvendo, para tanto, atividades nas áreas de prestação de serviço, assistência médica, ensino, pesquisa,

desenvolvimento tecnológico, produção, controle de qualidade e educação sanitária”.

O art. 3º do Decreto Estadual 43.668, de 26 de Novembro de 2003, apresenta algumas das atribuições, que descrevem a identidade da Hemominas, conforme apresentado na sequência:

- garantir à população a oferta de sangue e hemoderivados de boa qualidade em consonância com as diretrizes estabelecidas pela Política Estadual de Saúde;
- garantir à população a oferta de outros tecidos biológicos e células, de boa qualidade;
- desenvolver atividades nas áreas de assistência médica, ensino, pesquisa, tecnologia, produção, controle de qualidade e educação sanitária;
- integrar as funções, serviços e atividades concernentes à hematologia e hemoterapia de Minas Gerais, bem como planejar, executar e desenvolver atividades de serviço social e de assistência médica;
- planejar, coordenar e executar a produção de hemocomponentes, bem como a captação, preparação, preservação e distribuição de tecidos biológicos;
- planejar, coordenar e executar os trabalhos de controle de qualidade relativos à hematologia e hemoterapia;
- realizar pesquisas e implantar novas técnicas e descobertas científicas relacionadas com a coleta de sangue e outros tecidos biológicos (grifo nosso)
- prestar serviços de assessoria em hematologia e hemoterapia aos órgãos e entidades da saúde pública, às entidades privadas e à comunidade em geral;
- coordenar a distribuição dos hemocomponentes, hemoderivados e outros tecidos biológicos à rede pública e exercer outras atividades correlatas.

Ao longo de 25 anos de atuação²⁹, o processo evolutivo da Hemominas tem como resultados a ampliação do atendimento ao doador, registrando, em setembro de 2010, um milhão de cadastros de candidatos à doação de sangue, apenas no Hemocentro de Belo Horizonte, bem como a produção de 640 mil hemocomponentes.

²⁹ Dados obtidos no Plano Diretor Hemominas

Ainda a se destacar: a expansão da cobertura para cerca de 90% dos procedimentos transfusionais, em todo o estado vinculado ao Sistema Único de Saúde (SUS).

Suas unidades atendem a maioria dos estabelecimentos de saúde de Minas, incluindo hospitais públicos, filantrópicos e particulares: são 479 hospitais, em 276 municípios. Reguladas pelo Decreto 43.668, de 26/11/2003, a Fundação integra-se por meio de sistema de rede de unidades hemoterápicas hierarquizadas, descentralizadas nas macrorregiões do estado (Agências Transfusionais, Postos de Coleta, Núcleos Regionais e Hemocentros Regionais), possuindo uma administração central operando o sistema e atuando em linha de comando. Algumas destas unidades prestam atendimento a pacientes portadores de doenças do sangue, especialmente hemofílicos e com hemoglobinopatias congênitas (doença falciforme e outras), através de uma equipe multidisciplinar constituída por profissionais de diversas áreas.

Destaca-se, também, o cadastro de doadores para a realização de transplantes de medula óssea, em parceria com hospitais universitários, já foram cadastrados mais de 25 mil doadores de medula óssea, sendo parceira importante do MG Transplantes. Além disso, a busca pela melhoria contínua revela-se na implantação de programas como o de Hemovigilância, Gerenciamento de Resíduos de Saúde, Cetebio, Doador do Futuro e a Acreditação (busca da Certificação de seus serviços por agências avaliadoras). A FIG.12 apresenta a estrutura orgânica da instituição. Trata-se de uma representação da sua estrutura até o nível gerencial, conforme Decreto Estadual 44.466 de 16/02/2007.

3.1.2 A Hemominas como instituição de pesquisa

A Hemominas está cadastrada, desde 1997, como instituição de pesquisa junto ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Fundação de Amparo à Pesquisa de Minas Gerais (Fapemig). As pesquisas desenvolvidas na instituição têm por objetivo beneficiar o cidadão, melhorando as práticas técnicas voltadas à preservação da saúde do doador e ao atendimento ao paciente. As parcerias com universidades e outras instituições, no país e no exterior, são importantes para a produção científica, a qual conta com o financiamento de agências de fomento. As linhas de pesquisa delineadas na Fundação Hemominas são respectivamente:

- 1) Hemoglobinopatias
- 2) Hemostasia, trombose e coagulação
- 3) Doenças transmissíveis pelo sangue
- 4) Virologia
- 5) Transfusão de sangue e derivados
- 6) Imunohematologia
- 7) Doação de sangue e componentes
- 8) Hemoderivados
- 9) Saúde ocupacional e psicologia
- 10) Gestão em saúde

Para o desenvolvimento de pesquisas, a Hemominas possui alguns grupos relacionados aos temas correlatos à sua área de atuação, conforme breve descrição a seguir:

A. Grupo de Biotecnologia em Hemoderivados (Biohemo)

Desenvolve projetos de pesquisa na Hemominas (Serviço de Pesquisa) e em colaboração com outras instituições semelhantes (Instituto de Ciências Biológicas, UFMG; Centro de Pesquisas René Rachou, Fiocruz/ MG). O Biohemo³⁰ tem por objetivo geral desenvolver e implementar novas metodologias, tecnologias e diagnósticos relacionados a distúrbios das coagulopatias (hemofilia A e B; doença de von Willebrand) e das hemoglobinopatias (anemia falciforme e talassemia alfa). As linhas de pesquisa atualmente são: biotecnologia de peptídeos sintéticos para neutralizar os inibidores do FVIII; imunogenética de anticorpos inibidores do Fator VIII; polimorfismos de citocinas e padrão de expressão das citocinas em hemofílicos que produzem inibidores do FVIII; diagnóstico de alfa-talassemia.

B. Grupo sobre Doença Falciforme

30 Disponível em: <http://dgp.cnpq.br/buscaoperacional/detalhegrupo.jsp?grupo=8779208ORM2XA8>

Os trabalhos desenvolvidos pelo grupo têm contribuído para conhecer a prevalência, a epidemiologia e o comportamento clínico da doença falciforme³¹ na região. Tem como objetivo contribuir para a formação de recursos humanos, especialmente na Iniciação Científica de alunos de graduação, de Mestres e Doutores. Tais pesquisas vêm resultando em dissertações de mestrado e teses de doutorado, artigos científicos e diversas apresentações em eventos científicos.

C. Grupo sobre Segurança Transfusional

Os trabalhos do grupo³² visam aprimorar o conhecimento da atual situação da Doença de Chagas transfusional no Brasil e na América Latina, com foco no significado da sorologia inconclusiva ou indeterminada; promover o diagnóstico situacional das agências transfusionais do Estado de Minas Gerais; pesquisar a prevalência da aloimunização eritrocitária e plaquetária, essa última com foco na refratariedade das transfusões de concentrados de plaquetas; contribuir para a formação de recursos humanos.

D. Grupo Interdisciplinar de Pesquisa em HTLV (GIPH).

A Hemominas desenvolve pesquisas sobre o vírus linfotrópico de células T humanas tipo 1 HTLV (*Human T cell lymphotropic*), desde 1997, com a criação do Grupo Interdisciplinar de Pesquisas em HTLV (GIPH)³³. O GIPH é composto por pesquisadores da Hemominas e de instituições parceiras, como a Faculdade de Medicina e Instituto de Ciências Biológicas da UFMG, Rede Sarah de Hospitais do Aparelho Locomotor, Centro de Pesquisas René Rachou e Universidade Federal dos Vales do Jequitinhonha e Mucuri.

Esses grupos demonstram o caráter da Hemominas sob o ponto de vista científico, que por meio de parcerias com outras instituições e incentivos via agências de fomento, como a Fapemig, contribuem de forma relevante com o desenvolvimento técnico-científico. Para certificação de tal fato, uma simples busca na Base de dados

31 Disponível em: <http://dgp.cnpq.br/buscaoperacional/detalhegrupo.jsp?grupo=0002406Z8GP1AI>

32 Disponível em: <http://dgp.cnpq.br/buscaoperacional/detalhegrupo.jsp?grupo=0002401ZR0ZLY5>

33 Disponível em: <http://dgp.cnpq.br/buscaoperacional/detalhegrupo.jsp?grupo=8779401R8IU3BK>

SCOPUS³⁴, a partir do Portal de Periódicos disponibilizado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), utilizando-se o termo “Fundação Hemominas”, no campo de busca “Afiliação”, foram recuperados 81 trabalhos científicos.

A análise feita nessa base, em abril de 2011, mostra que tais trabalhos correspondem a um índice $H = 13$, o que significa que, dos 81 documentos considerados para o *h-Index*, 13 foram citados pelo menos 13 vezes no período de 1996 a 2011³⁵. Trata-se de um fato relevante, uma vez que a produção e divulgação do conhecimento científico são, sem dúvida, referência para o desenvolvimento de outros trabalhos no campo da ciência.

O desenvolvimento da presente pesquisa na Hemominas foi motivado tanto pelo fato de se tratar de uma instituição que presta importante serviço de saúde pública, quanto pelo desenvolvimento científico, que faz parte de sua identidade, não como uma ação secundária. Somando-se a isso, seus especialistas contribuem com relevante produção de conhecimento científico.

3.1.3 Motivação para a escolha do grupo de pesquisa na instituição

Após levantamento preliminar das pesquisas desenvolvidas na instituição, verificou-se que o GIPH apresentava duas situações em especial que motivaram a sua escolha.

A primeira delas foi detectada em encontros preliminares com a coordenação do GIPH quando havia sido relatada a existência de dificuldades entre comunidades nacionais e internacionais, atuando em pesquisa sobre HLTV e doenças associadas, na organização das terminologias, para representar os sintomas e níveis com os quais eles acontecem, bem como das diferentes manifestações e características.

Como exemplo, a seguinte situação foi apresentada: de um lado encontram-se indivíduos “*sintomáticos*” – aqueles que foram infectados pelo HTLV e desenvolvem

34 Conforme definido pela Capes, trata-se de uma Base de dados referencial que oferece funcionalidades como identificação autores e afiliações, análise de citações, análise de publicações e índice H.

35 Período considerado pela base de dados Scopus. Esta base não possui informações completas para artigos publicados antes de 1996

todos os sintomas da doença, dessa forma preenchem todos os critérios para o diagnóstico da HAM. Do outro, os sujeitos “*assintomáticos*” – aqueles que, embora tenham sido infectados pelo vírus, não desenvolveram nenhum dos sintomas para a doença. Mas, existem ainda indivíduos infectados que desenvolvem níveis diversos dos diferentes sintomas. Esses apresentam poucos sintomas, mas não preenchem os critérios para serem classificados como doentes de HAM. Os indivíduos que desenvolvem um ou outro sintoma, mas que ainda não são doentes de HAM recebem a classificação oligossintomáticos. Esse termo, muitas vezes, não é reconhecido pela comunidade científica para aplicação nesse contexto.

Essa situação gera questionamentos sobre a propriedade das denominações em uso. Tal fato é corroborado por Romanelli *et al* (2010),

“Na prática clínica, os critérios utilizados no diagnóstico da mielopatia estão bem definidos. Porém, os pacientes que apresentam outras manifestações neurológicas ou sistêmicas, como pacientes sintomáticos e com exame neurológico normal, não possuem uma terminologia e critérios clínicos bem definidos quanto à denominação e diagnóstico (ROMANELLI *et al*, 2010, p. 164)”.

Esse autor relata ainda que, do ponto de vista nosológico³⁶, há questionamentos sobre a propriedade da atual denominação para Mielopatia Associada ao HTLV/Paraparesia Espástica Tropical (HAM/TSP). Apresenta como exemplo outras denominações que têm sido propostas pela literatura da área, a saber: Encefalomielopatia pelo HTLV-1, Síndrome do HTLV-1, Doença Neurológica Associada ao HTLV-1, Doença Neuroimunológica Associada ao HTLV-1 e Complexo Neurológico do HTLV-1.

O segundo motivador da escolha se deve ao fato de o conhecimento produzido, ainda que bem difundido entre a comunidade científica, é desconhecido por grande parte de profissionais, principalmente aqueles das áreas da saúde que não estão envolvidos com essa temática. Isso ocorre mesmo considerando que o número de publicação sobre o tema seja relevante, ou seja, em torno de 1800³⁷ documentos disponíveis sobre o assunto. Esse fato ficou evidenciado na participação das reuniões

³⁶Ramo da medicina que estuda e classifica as doenças (Houaiss, 2001)

³⁷ Consulta realizada na base de dados Scopus em 23/09/2011

do grupo, em que era relatada a falta de conhecimento sobre o HTLV e doenças associadas como a HAM/TSP, principalmente por parte da comunidade médica, com relação ao diagnóstico, sintomas, sinais, entre outras características inerentes ao curso da doença. Essa questão era relatada como sendo uma das barreiras existentes para o conhecimento sobre HTLV e doenças associadas, em particular a HAM/TSP.

Tais fatos se apresentaram como um ambiente propício para o desenvolvimento da pesquisa proposta. Trata-se de questões que corroboram com a necessidade de organizar o conhecimento de um domínio, para fins de uso e compartilhamento. O presente trabalho considera as observações de Smith (2003), que sugere que ontologias devem ser concebidas sob a perspectiva de um campo da ciência, em que os especialistas têm o respaldo para falar com propriedade do domínio, o que não ocorre se baseada a partir do senso comum ou não especialista no assunto.

Conforme descrito no Diretório dos Grupos de Pesquisa CNPq, o GIPH tem por objetivo estudar e acompanhar a população com sorologia positiva ou indeterminada para HTLV I/II, buscando avaliar aspectos epidemiológicos, neurológicos, dermatológicos, reumatológicos e laboratoriais desses indivíduos e de seus familiares. A missão do GIPH, expressa em seu termo de criação, é produzir e divulgar para a comunidade conhecimento científico sobre o vírus linfotrópico da célula T humana (HTLV I/II) nas áreas: epidemiológica, clínica e laboratorial, buscando o entendimento da história natural da infecção e fornecendo suporte para Programas de Saúde e formação de recursos humanos.

3.2 Condução e recorte da pesquisa

A presente seção descreve como foi conduzida a pesquisa, bem como apresenta o seu recorte, dentre outras informações complementares necessárias ao entendimento do trabalho. Do ponto de vista da forma de *condução*, cabe explicar que a presente pesquisa é parte do *Blood Project*³⁸, uma iniciativa para organização da informação no domínio da biomedicina na área de hematologia e hemoterapia.

³⁸Disponível na Internet em: < <http://mba.eci.ufmg.br/blood/>>. Acesso em 20 de julho de 2010.

Tal projeto está organizado em cinco eixos principais, a saber: i) organização da informação, que consiste na construção de uma ontologia sobre sangue humano, denominada *Blood Ontology* (BLO) ³⁹; ii) aquisição de conhecimento a partir de especialistas; iii) aquisição de conhecimento a partir de textos da área biomédica; iv) visualização da informação; v) segurança da informação biomédica. A presente pesquisa se insere no item *Aquisição de Conhecimento*, a partir de especialistas (ALMEIDA *et al*, 2011; ALMEIDA; PROIETTI; SMITH, 2011).

A atividade de AC real foi conduzida para, a partir de sua conclusão, resultados pudessem ser avaliados de modo a buscar a consecução do objetivo geral da pesquisa. Usando termos comuns da literatura de AC, o autor da presente pesquisa atuou como o “*engenheiro do conhecimento*” buscando informações específicas do domínio adequadas à estrutura existente da BLO. Ao levar adiante um processo de AC, o engenheiro obtém o conhecimento especializado a ser considerado na ontologia. Conforme demonstrado na revisão de literatura, trata-se de um processo complexo que envolve múltiplas variáveis.

Para a realização de tal atividade, a presente pesquisa visou elicitar conhecimento dos especialistas do domínio do vírus HTLV, especificamente no contexto da Mielopatia Associada ao HTLV (HAM/TSP), doença com a qual se estabeleceu o recorte da presente pesquisa. Durante a atividade AC, dificuldades foram observadas e registradas como forma de buscar melhorias na atividade de AC, uma vez que tal atividade é parte do processo de construção de ontologias.

³⁹ Disponível na Internet em <<http://mba.eci.ufmg.br/BLO/>>. Acesso em 11 de julho 2010.

3.3 Descrição da metodologia de pesquisa

A metodologia de pesquisa buscou, nas considerações apresentadas na revisão de literatura (CAP. 2), o ponto de partida para a sua consecução, bem como para o levantamento e seleção das técnicas empregadas ao método de pesquisa aqui proposto. Conforme apresentado pelos autores pesquisados, a atividade de AC envolve técnicas que podem variar conforme as tarefas a serem executadas em cada etapa, na qual consideram que a AC abrange coleta, extração e validação de conhecimento. No presente trabalho, a atividade da AC, que acontece da interação direta com os especialistas, é referenciada como elicitación para obtenção de conhecimento acerca do domínio. As demais tarefas compõem o todo da atividade de AC.

Para maior clareza, dividiu-se a coleta de dados de acordo com a origem dos dados: *dados primários* e *dados secundários*. Os dados secundários correspondem à Fase de Levantamentos para o planejamento da atividade de AC. Nessa fase, os dados obtidos não exigiram em primeira instância contato com os especialistas para fins de elicitación de conhecimento. Por sua vez, os dados primários, por se tratarem de atividades inerentes à elicitación de conhecimento propriamente dita, exigiram contato direto com especialistas. Os dados primários correspondem às fases denominadas respectivamente Coleta, Validação e Refinamento do conhecimento.

O QUADRO 7 apresenta uma visão do método empregado na presente pesquisa, cujas técnicas e instrumentos são apresentados a seguir na seção 3.4, em que são descritos em detalhe como foi desenvolvido cada um.

QUADRO 7: Roteiro metodológico proposto

Fase	Objetivo da tarefa	Breve descrição da Tarefa
Levantamento	Escopo da Ontologia	Conhecer a audiência; Conhecer a ontologia de alto nível que fundamenta a ontologia em desenvolvimento e; Obter informações básicas sobre o contexto.
	Conhecer os fundamentos	Conhecer o escopo básico do domínio e respectivos conceitos.
	Identificar <i>expertise</i>	Identificar a <i>expertise</i> dos especialistas envolvidos
Contato	Obter conhecimento	Elicitar conhecimento de especialistas
	Esclarecer dúvidas e mal entendidos	Esclarecer dúvidas ou mal entendidos

Validação	Levantar e registrar termos	Registrar termos elicitados, candidatos à ontologia
	Validar conhecimento	Obter aprovação sobre termos adquiridos e suas definições
	Atualizar	Atualizar dados após validação
Refinamento	Fazer integração entre granularidades	Caracterizar genes, proteínas, partes do corpo relacionadas.
	Conectar com ontologias de alto nível	Conectar dados de granularidades diversas com ontologias básicas

Fonte: Dados da pesquisa

3.4 Instrumentos de pesquisa

A presente seção apresenta os instrumentos de pesquisa utilizados no presente trabalho. Considera-se que tais instrumentos são técnicas empregadas nas tarefas que compõem a atividade de AC.

3.4.1 Técnicas adotadas para coleta de dados

A seleção das técnicas para coleta de dados se baseou nas propostas de AC citadas na revisão de literatura, levando em consideração que a busca de melhorias nessas atividades levaria em conta as dificuldades apresentadas. Considerou-se a AC definida em termos de suas etapas, sendo que as tarefas foram combinadas conforme cada etapa do processo. A seleção das técnicas se baseou nos seguintes critérios: i) possibilitar a compreensão do domínio do conhecimento; ii) identificar *expertise* dos especialistas do domínio; iii) permitir a interação com especialistas de modo individual; iv) viabilizar a elicitación do conhecimento do domínio; v) orientar a organização do conhecimento elicitado.

Vale ressaltar que não foram consideradas aquelas técnicas que se propunham elicitar conhecimento, cujos objetivos fossem diferentes da proposta do presente trabalho, como, por exemplo, de verificar como determinado especialista resolveria determinados problemas, em busca de possíveis soluções etc. Tal ação seria necessária se o contexto fosse o desenvolvimento de sistemas, como preconizado pela engenharia do conhecimento, o que não é o propósito desta pesquisa.

Dentre as técnicas pesquisadas avaliaram-se aquelas consideradas adequadas ao contexto da presente pesquisa. Por exemplo, verificou-se que o grupo escolhido para as sessões de elicitación de conhecimento possuía grande número de publicações científicas. Desse modo, optou-se por métodos que viabilizassem o mapeamento do domínio a partir da produção científica. Nesse caso, a análise de domínio foi considerada como um método adequado, por possibilitar ao pesquisador identificar o domínio por meio dos textos científicos produzidos pelos especialistas. Os métodos selecionados para a presente pesquisa são apresentados à seguir, de acordo com as etapas definidas, descritas em detalhes no capítulo 4 (Coleta de dados e Análise dos resultados).

- *Entrevista*, escolhida por sua popularidade e eficiência, e por ser uma técnica que possibilita acesso de modo formal ou informal ao especialista. Possibilita, ainda, uma visão geral do domínio, elicitando o conhecimento do especialista de modo verbalizado e detalhamento do conhecimento sobre termos, conceitos, suas propriedades e relações;
- *Imersão na Literatura*, por ser um método que possibilita obter os conceitos básicos do domínio em questão, para conhecimento teórico mínimo requerido;
- *Análise de Domínio*, pela visão acerca do domínio, em que é possível elencar os principais elementos que o constituem. Em especial pela possibilidade de aplicar métodos como a *bibliometria e análise de citação*, métodos válidos e amplamente utilizados na Ciência da Informação e Biblioteconomia, para mapear campos do conhecimento a partir do domínio;
- *Análise de protocolo*, por ser um método que auxilia na transcrição de entrevistas, que objetiva analisar os registros gerados pelos especialistas;
- *Teachback*, por ser uma técnica que visa obter um *feedback* dos especialistas, a partir da descrição de uma parte do conhecimento adquirido em sessões anteriores. Além disso, por auxiliar na correção de equívocos conceituais e ajuda a esclarecer mal entendidos.

Alguns problemas na avaliação das técnicas para a atividade de elicitación de conhecimento (PAYNE *et al.*, 2007), a saber: dificuldades de envolver uma amostra

suficientemente grande de especialistas para dar significância estatística; dificuldade de se avaliar e validar o conhecimento adquirido, dada a falta de padrões para comparação; necessidade de se descobrir uma variedade de tarefas e domínios para que as técnicas sejam úteis; dificuldade de isolar qual valor agregado uma técnica oferece, quando cada técnica pode precisar trabalhar ao lado de outras.

3.4.2 Instrumentos para coleta de dados

Os instrumentos para a coleta de dados foram desenvolvidos ou adaptados conforme a necessidade das etapas. Foram empregados:

- Entrevistas não estruturadas, entrevistas estruturadas e entrevistas de *feedback* fizeram uso *de* três formatos: i) entrevistas com perguntas abertas; ii) entrevistas para elicitare o conhecimento acerca da doença; iii) entrevista de *feedback* não estruturada.
- Formulários aplicados: formulário para levantamento dos especialistas, roteiros de entrevista, agenda para sessão de aquisição de conhecimento, síntese de entrevista;
- *Template do Protégé Frames* com roteiro para entrevista estruturada, a partir do raciocínio médico criado por Scheuermann (2009).
- Gravação: para registros dos contatos e transcrição

Para entendimentos acerca do roteiro para fins de elicitare de conhecimento, cabe, nesse momento, esclarecer conceitos utilizados como referência. Cada um dos termos descritos no QUADRO 8 a seguir, compõe o “quadro terminológico”, que abrange as doenças, suas causas, manifestações, diagnóstico e outras entidades relacionadas ao modo como as doenças são reconhecidas e interpretadas pela medicina (SCHEUERMANN *et al.*, 2009).

Para se chegar a tais definições, Scheuermann *et al* (2009), revisaram e criticaram as atuais terminologias padrão: i) *SNOMED Clinical Terms* (SNOMED CT), que se propõe a fornecer uma linguagem comum de indexação, armazenamento, recuperação e agregação de dados clínicos através das especialidades e locais de atendimento; ii) do *National Cancer Institute Term* (NCIT), que hospeda e

disponibiliza um conjunto de vocabulários organizados por categorias das fontes de padrões e terminologias.

Após a definição dos termos adotados na AC, apresenta-se uma explicação do próprio raciocínio subjacente de acordo com a proposta do quadro clínico sobre a perspectiva da doença (FIG. 12).

QUADRO 8: Termos de referência para o roteiro da entrevista

Termo	Definição	Exemplo
(1) Processo etiológico	Um processo no organismo que leva a uma desordem subsequente.	a) exposição a substâncias químicas tóxicas resulta numa mutação genômica no DNA de uma célula. b) herança de um defeito no receptor da membrana das plaquetas que afeta a homeostase;
(2) Desordem	Base física da doença. Uma combinação de causas relativamente isoladas de componentes físicos: (a) clinicamente anormal e (b) máxima, no sentido de que não é uma parte de alguma combinação maior.	a) glóbulos vermelhos com um saldo negativo de ferro; b) plaquetas com uma anomalia do complexo glicoproteína Ib (GP Ib) ou (ausência de GP Ib) quantitativa ou qualitativa (mutação de genes, ou seja, os genes GPIBA, GPIBB ou GP9); c) cromossomo 4 com mHTT anormal; d) mutação genômica do DNA; e) porções de endotoxina no sangue; f) sangue com níveis sanguíneos reduzidos de cortisol causando crise adrenal.
(3) Manifestação pré clínica da doença	Uma manifestação da doença existente antes desta ser detectada clinicamente ou por exame físico, ou relatado pelo paciente ao médico. Uma manifestação da doença que ocorre antes desta se tornar detectável clinicamente ou por exame físico.	Disfunção diastólica do ventrículo esquerdo (VE) representa a manifestação pré-clínica mais precoce da cardiomiopatia diabética, precedendo a disfunção sistólica e podendo progredir para insuficiência cardíaca sintomática.
(4) Doença	Uma disposição que se submete a: (i) processos patológicos que (ii) existe em um organismo por causa de um ou mais desordens no organismo.	a) Anemia microcítica associados ao transtorno de ferro. b) Epilepsia (doença) dispõe à ocorrência de convulsões (processos patológicos), devido a uma anomalia subjacente ao circuito neuronal do cérebro (base física)
(5) Manifestação clínica da doença	Uma manifestação da doença detectável por meio da história clínica do paciente ou por exame físico.	Metástase orbital como primeira manifestação clínica de hepatocarcinoma.
(6) Fenótipo clínico	Fenótipo é uma (combinação de características) qualidade (s) de um organismo, determinado pela sua constituição genética e do ambiente, que se diferencia (em instâncias específicas) de outras instâncias da mesma espécie. Um fenótipo clínico anormal é um fenótipo anormal de acordo com padrões clínicos.	Esta técnica não invasiva pode ser empregada para detectar regiões muito pequenas de angiogênese associada a tumores de melanoma nascente e para o fenótipo de melanoma em estágio inicial em um ambiente clínico.
(7) Achado clínico	Uma representação de um “achado” no resultado da história clínica;	de um exame físico ou de imagem; ou de alguma combinação entre ambos.

(8) Processo patológico	Manifestação da desordem. Alguns proc. patológicos podem ser alguma mudança no modo que uma função fisiológica normal é realizada (hiperventilação) ou não tem nenhuma contrapartida fisiológica normal (inflamação aguda)	a) acumulação de fragmentos de proteínas mHTT, regulação da transcrição anormal, morte celular neuronal no estriado; b) síntese de hemoglobina deficiente; c) adesividade plaquetária anormal da matriz extracelular durante a fase inicial de formação de plug
(9) Característica anormal do corpo	Um processo corporal clinicamente anormal.	“ <i>síntese de hemoglobina deficiente</i> ” produz uma característica corporal anormal, reconhecida como um sintoma ou sinal, como uma <i>fadiga</i> , <i>lesões nas unhas</i> etc.
(10) Sintomas	Uma qualidade (característica física) observada pelo paciente ou hipotetizada pelo paciente para ser a realização da doença. Geralmente se refere a dores e outros sentimentos e sensações que podem ser observadas apenas pelo paciente	a) tontura e vertigem; b) fraqueza; c) dor
(11) Sinais	Uma qualidade (característica física) de um paciente observada por um exame físico e considerada pelo médico para ser de significância clínica	a) icterícia (coloração amarelada da pele); b) alopecia (perda de pelos e cabelos); c) febre d) perda de peso; e) presença de infecções recorrentes na pele, na urina, nos pulmões e em outros locais
(12) Processo interpretativo	Processo interpretativo que produz uma hipótese para o resultado	
(13) Diagnóstico	A representação da conclusão de um processo interpretativo que tem como entrada o quadro clínico de um determinado paciente e como saída uma afirmação (declaração do diagnóstico) no sentido de que o paciente tem uma doença de tal tipo.	
(14) Hipótese	Prognóstico sobre o curso de uma doença.	a) descartar anemia microcítica associada ao transtorno de ferro; b) descartar Síndrome de Bernard-Soulier c) descartar a doença de Huntington.
(15) Testes laboratoriais	Um ensaio de medição que tem como entrada uma amostra obtida do paciente e como saída um resultado que representa uma qualidade desta amostra.	
(16) Resultado de teste	a) nível de hemoglobina reduzida no sangue; granulocitopenia leve e leucocitose neutrofílica, trombocytosis; nível ferro reduzido de macrófagos na medula óssea, baço e fígado, concentração sérica de ferro reduzido; b) detecção molecular do gene HTT com > 39 repetições CAG c) plaquetas com morfologia anormal no esfregaço de sangue, a adesão ao colágeno defeituoso in vitro, aumentou o tempo de sangramento, a aglutinação das plaquetas defeituosas em resposta à ristocetina.	

(17) Resultado	a) diagnóstico que paciente X tem uma doença que leva à anemia microcítica associada ao transtorno de ferro b) diagnóstico que paciente X tem uma doença que leva à doença de Huntington c) diagnóstico que paciente X tem uma doença que leva à Síndrome de Bernard-Soulier
(18) Tratamento	a) Conjunto de meios: higiênicos, farmacológicos, cirúrgicos ou físicos que tem finalidade de cura ou alívio de enfermidades ou sintomas, após a elaboração de um diagnóstico.

Fonte: Adaptado de Scheuermann et al (2009)

A partir destas definições, o quadro clínico com respectivo curso da doença (FIG. 12), desenvolvido por Smith *et al* (2008) e Scheuermann *et al* (2009), se encontra apresentado na íntegra. A partir deste, explica-se o roteiro desenvolvido.

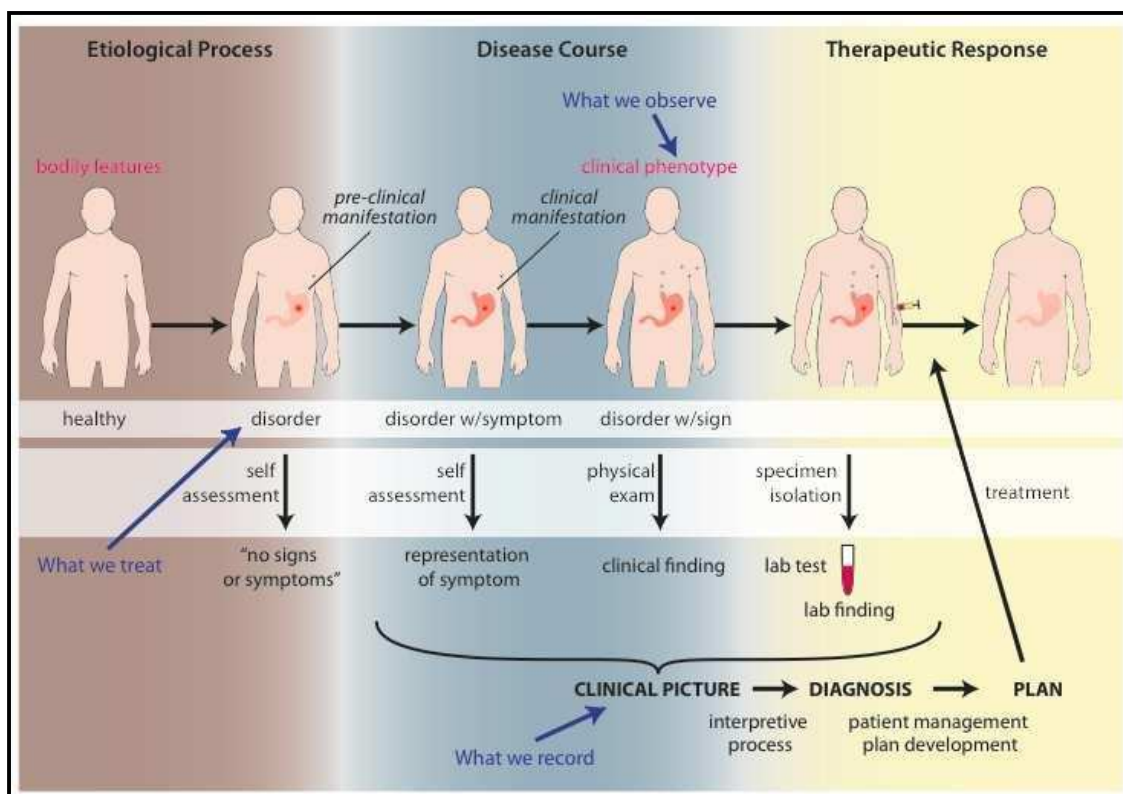


FIGURA 12: Quadro clínico sob a perspectiva do curso da doença
Fonte: Adaptado de Smith *et al* (2008) e Scheuermann *et al* (2009)

Na fase chamada *processo etiológico* (FIG. 13), considera-se que há um corpo humano saudável, com características normais de acordo com parâmetros médicos. Como resultado desse processo, ocorre uma *mudança física* no indivíduo, dando origem à doença. Na *manifestação pré-clínica* da doença, o corpo desenvolve *desordens*, que são portadoras de disposições. Tais disposições são naturalmente associadas à existência de entidades, por exemplo, a disposição do corpo humano para ficar doente e da disposição da fruta amadurecer (SMITH, 2008). O paciente então, em alguns casos, já percebe mudanças no organismo, embora ainda não existam sinais ou sintomas.



FIGURA 13: Processo etiológico, desordem e doença (disposição)
Fonte: Desenvolvido pela autora

A fase *curso da doença* começa com a *manifestação da doença* (FIG 14). Nesse momento, a desordem se manifesta por meio de *sintomas e sinais* (*manifestações clínicas*), sendo o primeiro aqueles que os pacientes conseguem sentir ou identificar e o segundo, que é determinado pelo médico através de *exames físicos e testes laboratoriais*. Nesta fase, é possível determinar os *fenótipos clínicos* – totalidade de sinais e sintomas-, que é a principal característica observável da doença.



FIGURA 14: Doença, processo patológico e condições anormais

Fonte: Desenvolvido pela autora

Na fase *resposta terapêutica*, uma amostra do paciente é obtida de alguma parte do corpo, a fim de submetê-la a *testes laboratoriais*. Esses *resultados* são registrados no prontuário médico como um quadro clínico. O quadro clínico é *interpretado* pelo médico na busca pelo diagnóstico. Nesse ponto, é possível que se estabeleça um plano de *tratamento*, baseado neste *diagnóstico*, de modo que o corpo possa retornar à normalidade. O plano é o *resultado do diagnóstico* fundamentado no processo de interpretação do quadro clínico. O quadro clínico é composto por registros da representação de sintomas e sinais, bem como os resultados do exame físico e laboratorial, como apresentado na FIG 15.



FIGURA 15: Sinais, sintomas e processo interpretativo

Fonte: Desenvolvido pela autora

A adoção de tal perspectiva para o roteiro da entrevista decorre-se do fato de se tratar de uma abordagem cujas definições são apresentadas como referência para melhores práticas no desenvolvimento de ontologias, promulgadas no *OBO Foundry*, conforme defendido por Smith *et al* (2008) e Scheuermann *et al* (2009).

O desenvolvimento na *template* do Protegé-Frames ocorreu no período de fevereiro a março de 2011. Ao final, o roteiro foi apresentado a um especialista que não faria parte das sessões de elicitação, para considerações da organização e entendimentos acerca dos conceitos utilizados, conforme proposto pelos autores de referência. Após as

considerações apresentadas pelo especialista, ajustes foram feitos e se chegou à versão final. O exemplo apresentado na FIG. 16 mostra um fragmento do roteiro final.

FIGURA 16: Roteiro no *Protegé-Frames*

The screenshot displays the Protegé-Frames software interface. At the top, there is a menu bar with icons for file operations (open, save, print, etc.) and a toolbar. Below the menu bar, there are tabs for 'Classes', 'Slots', 'Forms', 'Instances', and 'Queries'. The 'Instances' tab is active, showing two panes: 'INSTANCE BROWSER' on the left and 'INSTANCE EDITOR' on the right.

The 'INSTANCE BROWSER' pane shows a list of instances for the class 'About disease'. The instances listed are 'DAB', 'VAU', and 'JIY'. The 'INSTANCE EDITOR' pane shows the details for the instance 'DAB' (instance of About disease, internal name is interview_v1_Class23). The editor contains several text boxes for data entry:

- Name Of Respondent:** A text box containing 'DAB'.
- About Disease (4):** A text box containing 'Bernard-Soulier Syndrome'.
- About Etiological Process (1):** A text box containing 'inheritance of a defect in the platelet membrane receptor that affects the homeostasis'.
- About Clinical Manifestation Of A Disease (5):** An empty text box.
- About Disorder (2):** A text box containing 'platelets with an abnormality of glycoprotein Ib complex (GP Ib), either quantitative (absence of GP Ib) or qualitative (mutation of genes, namely, the genes GP1BA, GP1BB, or GP9)'.
- About Clinical Phenotype (6):** An empty text box.

Fonte: Desenvolvido pela autora

O modelo ontológico formal de doença cabe ainda lembrar, foi utilizado somente para as entrevistas e para notas, não em uma iniciativa formal para formalizar conhecimento diretamente.

Antes de se proceder à entrevista, uma agenda era preparada pelo pesquisador de modo que fosse possível registrar a finalidade do encontro, etapas, tempo utilizado, data do encontro, para citar alguns, conforme formulário apresentado no QUADRO 9.

QUADRO 9: Formulário de agenda para elicitação de conhecimento

Agenda para elicitação de conhecimento		Tabela nº:	FORM. nº
Pesquisador:	Unidade:	Data:	
Assunto da sessão:			
Desmembrado em mais de uma sessão? S - N		nº dessa sessão:	
Etapas da sessão: (1) (2)		Tempo estimado: (1) (2)	
Descrição da etapa (1)			
Finalidade (1):			
Documentação (1)			
Descrição da etapa (2)			
Finalidade (2):			
Material utilizado (2)			

Fonte: Adaptado de Almeida (2006)

Para proceder ao levantamento dos especialistas foi desenvolvido um formulário de modo que fosse possível levantar dados que pudessem identificar cada um dos especialistas, áreas de atuação, pesquisas envolvidas, fontes de literatura da área, entre outros. O QUADRO 10 apresenta o modelo do formulário desenvolvido para a presente pesquisa.

QUADRO 10: Formulário utilizado para levantamento dos especialistas

Pesquisador ⇒		FORM . Nº
Formação / Titulação		
Especialidade		
Instituição (ões) de Vínculo		
Assunto (pesquisa na área)		
Projeto(s) de pesquisa(s) correlato(s)		
Linhas de pesquisa		
Fontes sugeridas		

Fonte: Desenvolvido pelo autor

Por sua vez, o formulário de síntese das entrevistas (QUADRO 11) foi utilizado para registro das informações relevantes que foram levantadas durante a entrevista com o especialista.

QUADRO 11: Formulário para síntese de entrevista

Síntese de entrevista			FORM. nº
Membro consultado:	Setor:	Data:	nº tabela:
Objetivo da entrevista			
Síntese:			

Fonte: Adaptado de Almeida (2006)

5 CONSIDERAÇÕES FINAIS

Embora árduo o trabalho da escrita, da síntese e do método, amo tudo isso.
Lamentável chegar ao fim, mas é belo recomeçar e traçar novos caminhos;
melhor ainda é saber que se quer voltar à escrita, à síntese e ao método e
tomar de início aquilo que chega ao fim...

Para a realização da presente pesquisa introduziu-se o assunto, discorrendo sobre a busca sistemática para a organização da informação, onde as ontologias são introduzidas como alternativa para representar o conhecimento e a informação, para uso por pessoas e legível por sistemas de informação automatizados. Para tal, foram mostradas as atividades de sua construção, onde a AC foi apresentada como integrante ao processo e tema do trabalho. Por sua vez, no Capítulo 2 a revisão de literatura dissertou sobre a temática da AC, explorando-a, desde o seu surgimento no contexto da engenharia do conhecimento, perpassando pela ciência da informação e biomedicina, apresentando em seguida as técnicas discutidas pela literatura da área. No capítulo 3, a metodologia de pesquisa e os instrumentos desenvolvidos para a coleta de dados foram propostos e apresentou-se a instituição. Por fim, dificuldades foram apontadas de modo a chamar a atenção para os principais fatores. No capítulo 4, foram apresentados o modo como os dados foram coletados e respectiva análise. Por sua vez, o Capítulo 5 apresenta as considerações finais, que se inicia neste momento.

Conforme observado, embora a AC seja indispensável para esse propósito, a literatura da área aponta-a como um “gargalo” (HAYES-ROTH; WATERMAN; LENAT, 1983; ZAF; MCNEESE; SNYDER, 1993; MASTELA, 2005), em função, tanto da necessidade de metodologias confiáveis, quanto das dificuldades enfrentadas pelo engenheiro do conhecimento em obter o conhecimento dos especialistas. Nesse contexto, surgiu a motivação para a presente pesquisa que busca responder à indagação suscitada pelo problema que se propôs a investigar: identificar fatores que geram dificuldades na AC em domínio especializado como a biomedicina. Espera-se contribuir com a discussão científica a esse respeito, tendo a Ciência da Informação como o meio científico.

Pela proposta apresentada, o objetivo geral do trabalho foi alcançado à medida que se apontam problemas e propõem-se práticas alternativas, através de técnicas combinadas, para reduzir a distância entre o que o especialista tem a oferecer e aquilo que é de fato registrado como o que ele sabe. Uma contribuição evidente nesse sentido é o roteiro metodológico descrito na seção 3.3 e sua aplicação prática em campo, conforme apresentado no capítulo 4. Ainda que o roteiro metodológico necessite ser testado e melhorado ao longo de outros estudos, verificou-se a viabilidade de sua adoção para eliciação de conhecimento em domínios além da biomedicina, estendendo-o para o domínio das ciências da vida, cujos diferentes aspectos das doenças podem ser vistos de forma multidisciplinar.

Vale ressaltar que, com relação à técnica de entrevista estruturada proposta para eliciação de conhecimento, a qual é baseada em Scheuermann, Ceusters e Smith (2009), e desenvolvido no *Protegé-Frames*, sua aplicabilidade foi essencial para estabelecer diálogo com os especialistas e com isso levantar os principais termos candidatos ao domínio da doença HAM. Essa estrutura se mostrou necessária e importante para conduzir o processo, inclusive por levar em consideração o tempo disponível para a realização tarefa - fator preponderante para os especialistas

Deve-se levar em consideração que esse modelo ontológico de doença não foi aplicado como uma iniciativa formal para formalizar conhecimento diretamente na ontologia. O conhecimento não registrado por esse motivo foi manipulado pelo uso de técnicas complementares.

Por sua vez, as conclusões da pesquisa revelam aspectos problemáticos que devem ser resolvidos para alcançar eficiência no processo de AC. A atenção a tais aspectos detectados aliada ao uso de técnicas, conforme apresentado ao longo do trabalho, possibilita a almejada redução da distância entre especialista e engenheiro. As principais conclusões obtidas com a pesquisa bibliográfica levam ainda considerações de autores estudados na presente pesquisa (HAYES-ROOT *et al*, 1983; SHADBOLT; SWALLOW, 1993; TEIVE, 1997; Hanes, 2003; MILTON, 2007), que abordaram a temática, nas quais consideram fatores que geram dificuldades na AC.

- Os indivíduos nem sempre estão dispostos a compartilhar conhecimento com colegas, e se o fazem é apenas pela oferta de algum tipo de benefício;

- Os indivíduos participam de atividades de AC por determinação da gerência e nem sempre espontaneamente;
- Os indivíduos vêem o conhecimento que detêm como um tipo de “propriedade intelectual” que deve ser resguardado para ser usado em consultoria ou outro emprego;
- Existe o medo de demissões no caso do compartilhamento de conhecimento; a percepção é de que conhecimento singular fornece proteção do emprego, e torná-lo disponível para outros aumenta a vulnerabilidade;
- Existe incerteza sobre a validade do conhecimento que o indivíduo detém;
- Existe a expectativa de que o conhecimento elicitado seja “engavetado”, o que dá a sensação de tempo perdido;
- Existe o medo da perda de status ao revelar conhecimento, pois dessa forma não haverá apenas um especialista na organização;
- Especialistas são profissionais altamente requisitados e suas atribuições de trabalho não deixam tempo disponível para elicitación do conhecimento;
- Existe dificuldade na representação do conhecimento pelo especialista decorrente da inabilidade de expressar claramente o conhecimento que detém;
- Especialistas armazenam grandes quantidades de conhecimento, mas grande parte deste conhecimento é tácito, o que o torna difícil sua descrição;
- Conhecimento especializado se torna obsoleto rapidamente, além de na maioria dos casos ser incompleto e nem sempre essencial;
- Existe dificuldade de expressão relacionadas a dificuldade de introspecção, dificuldade de verbalizar sob pressão, dificuldade com relação ao vocabulário;
- Existem dificuldades relacionadas a forma de apresentação da questão (geral ou específica demais) e às reações do especialista (falta de conhecimento, má interpretação, nível básico demais, dentre outras).

As conclusões aliadas a observações feitas ao longo dos estágios permitiram identificar grupos de fatores genéricos que geram resultados não desejados no processo. Tais grupos de fatores são:

- fatores relacionados ao perfil do especialista, tais como: treinamento e participação prévia em projetos similares;
- fatores contextuais, tais como: questões culturais, geográficas, políticas e financeiras; falta de acesso a fontes de informação relevantes e deficiências organizacionais;
- fatores relacionados a interação entre especialista e engenheiro, tais como: visão de curto prazo, complexidade do domínio, etc;
- fatores que tornam difícil registrar resultados, tais como: não aprovação pelo especialista dos resultados da atividade, constante avanço do campo científico, para citar alguns, e a partir daí especificar terminologias ou direções para esta especificação em determinado domínio.

Cabe citar que os desdobramentos técnicos e sociais desses resultados são evidentes, ao possibilitar melhorias na forma de registro do conhecimento especializado e ao auxiliar no esclarecimento dos sintomas, viabilizando o diagnóstico e compartilhamento de conhecimento sobre a doença. Do ponto de vista de contribuições para a pesquisa do campo da informação, a pesquisa proposta poderá ser útil para detectar e reduzir problemas relativos à Representação do Conhecimento.

No contexto explorado pelo presente trabalho, a atividade de AC a partir de especialistas, como parte do processo para construção de ontologias, também pode ser entendida como uma atividade preliminar à extração automática de termos, pois é necessária para que se tenha conhecimento especializado para julgar se os termos extraídos fazem de fato sentido para o domínio. Os vocabulários formais biomédicos têm como importante função definir consensualmente o significado de termos utilizados na prática e pesquisa médica (SMITH, 2006). Isso se torna possível justamente quando se pensa na aquisição de conhecimento a partir de especialistas.

Observou-se, por exemplo, no repositório BioPortal⁴⁴ (*NCBO's ontology repository*) que contêm terminologias da SNOMED, MeDRA⁴⁵ e NCI⁴⁶ *Thesaurus* os

⁴⁴ Disponível em: <http://bioportal.bioontology.org/>

⁴⁵ *Medical Dictionary for Regulatory Activities Terminology*

⁴⁶ *A vocabulary for clinical care, translational and basic research, and public information and administrative activities*

termos relacionados ao HTLV, à Mielopatia Associada ao HTLV, à Paraparesia Espástica Tropical, para citar alguns. Tais termos compõem uma taxonomia de classes e subclasses existentes. No entanto, não ficou claro se havia nos critérios para a escolha dos termos candidatos à ontologia algum tipo de validação de especialistas da área. Trata-se de fontes públicas que fornecem dados para anotação automática dos termos associados.

Como resultado da presente pesquisa, espera-se que as discussões aqui apresentadas contribuam para a Ciência da Informação, ao enriquecer a discussão sobre atividade de AC. Do ponto de vista de contribuições para a pesquisa do campo da informação, a pesquisa proposta poderá ser útil para detectar e reduzir problemas relativos à Representação do Conhecimento. Cabe destacar os evidentes desdobramentos técnicos e sociais dos resultados que, mesmo parciais, já possibilitam melhorias na forma de registro do conhecimento especializado favorecendo o diagnóstico e o compartilhamento de conhecimento sobre doenças negligenciadas.

Espera-se que as discussões apresentadas na presente pesquisa contribuam, enriquecendo a discussão sobre modos de ajudar, tanto o engenheiro do conhecimento quanto os pesquisadores envolvidos com a atividade de AC, tornando seu trabalho mais efetivo e, no longo prazo, auxiliando o próprio especialista do domínio, que, no caso do presente trabalho, são médicos, biólogos, bioquímicos, para citar alguns. Entretanto, algumas questões ainda permanecem em aberto, nos quais cabe citar: o uso de uma ferramenta colaborativa, como a wiki, possibilita a efetiva participação de todos especialistas no domínio? O uso de uma *wiki* pode ser considerado um facilitador para as anotações diretamente? O especialista deveria fazer anotações sozinho ou em conjunto com o cientista da informação? Além disso, identificar como as validações acontecem na prática e se existe alguma dificuldade por parte do especialista em fazer anotações, ou do cientista da informação em anotar o que de fato é obtido do especialista – o conhecimento registrado pelo engenheiro do conhecimento é o conhecimento do especialista?

A tarefa de aquisição de conhecimento, conforme apresentado no presente trabalho, é apenas parte do processo de construção de ontologias de determinado domínio. Como parte, é também responsável pela tarefa de apreensão, organização, formalização e compartilhamento do conhecimento, via ontologias, entre especialistas de um domínio do conhecimento.