PONTIFÍCIA UNIVERSIDADE CATÓLICA DE CAMPINAS

MESTRADO EM CIÊNCIA DA INFORMAÇÃO

SÉRGIO FURGERI

REPRESENTAÇÃO DE INFORMAÇÃO E CONHECIMENTO: ESTUDO DAS DIFERENTES ABORDAGENS ENTRE A CIÊNCIA DA INFORMAÇÃO E A CIÊNCIA DA COMPUTAÇÃO

CAMPINAS

2006

SÉRGIO FURGERI

REPRESENTAÇÃO DE INFORMAÇÃO E CONHECIMENTO: ESTUDO DAS DIFERENTES ABORDAGENS ENTRE A CIÊNCIA DA INFORMAÇÃO E A CIÊNCIA DA COMPUTAÇÃO

Dissertação apresentada ao Curso de Pós-Graduação, em Ciência da Informação, da Pontifícia Universidade Católica de Campinas, como requisito parcial à obtenção do título de Mestre.

Orientador:
Raimundo Nonato Macedo dos Santos

Campinas
2006

Ficha Catalográfica
Elaborada pelo Sistema de Bibliotecas e
Informação - SBI - PUC-Campinas

t020 **Furgeri, Sérgio.**

F983r Representação de informação e conhecimento: estudo das diferentes abordagens entre a ciência da informação e a ciência da computação / Sérgio Furgeri. - Campinas: PUC-Campinas, 2006.
159p.

Orientador: Raimundo Nonato Macedo dos Santos.
Dissertação (mestrado) - Pontifícia Universidade Católica de Campinas, Centro de Ciências Sociais Aplicadas, Pós-Graduação em Ciência da Informação.

Inclui anexos e bibliografia.

1. Ciência da informação. 2. Redes de informação. 3. Sistemas de recuperação da informação. 4. Linguagem de programação (Computadores) 5. Metadados. 6. Linguagem documentária. I. Santos, Raimundo Nonato Macedo dos. II. Pontifícia Universidade Católica de Campinas. Centro de Ciências Sociais Aplicadas. Pós-Graduação em Ciência da Informação. III. Título.

22.ed.CDD – t020

SÉRGIO FURGERI

REPRESENTAÇÃO DE INFORMAÇÃO E CONHECIMENTO: ESTUDO DAS DIFERENTES ABORDAGENS ENTRE A CIÊNCIA DA INFORMAÇÃO E A CIÊNCIA DA COMPUTAÇÃO

Dissertação apresentada ao Curso de Pós-Graduação, em Ciência da Informação, da Pontifícia Universidade Católica de Campinas, como requisito parcial à obtenção do título de Mestre.

COMISSÃO EXAMINADORA

_____

Raimundo Nonato Macedo dos Santos

_____

José Fernando Modesto da Silva

_____

José Estevão Picarelli

A Deus e a minha família,
companheiros de todas as horas boas e desafiadoras.

AGRADECIMENTOS

A Deus pelo meu existir, pelo cuidado Dele para comigo e pelas oportunidades de aprimoramento e experiências a que me expõe, bênçãos estas que favorecem em um de meus propósitos: de contribuir para uma sociedade organizada e experiente.

Ao meu orientador Dr. Raimundo Nonato Macedo dos Santos, braço amigo de todas as etapas deste trabalho.

Aos meus familiares e pessoas mais próximas, pela confiança, motivação, oração e ajuda em tarefas de meu dia-a-dia.

A querida Ivone, minha esposa virtuosa, pelo apoio e dedicação em nosso precioso relacionamento, buscando ambiente e situações favoráveis também para meus estudos e pesquisas.

Ao Lucas, meu filho prometido, por lembrar de agradecer ao “Papai do Céu” pelos alimentos, inspirando-me assim na busca constante de melhoria da minha carreira profissional.

Aos amigos e colegas, pela força e pela forte vibração em relação a esta jornada.

Aos professores e colegas de Curso, pois juntos trilhamos uma etapa importante de nossas vidas.

Furgeri, Sérgio. Representação de informação e conhecimento: estudo das diferentes abordagens entre a ciência da informação e a ciência da computação. 2006. Dissertação (Mestrado em Ciência da Informação) - Pontifícia Universidade Católica de Campinas.

RESUMO

A Ciência da Informação vem estudando formas de representação da informação e do conhecimento visando à recuperação da informação. Esta pesquisa tem seu foco na representação do conhecimento e da informação, procurando investigar quais são os pontos convergentes e divergentes entre as linguagens documentárias da Ciência da Informação e as linguagens de marcação desenvolvidas e utilizadas na Ciência da Computação, tendo em vista identificar ações, teorias e processos necessários para uma maior integração entre as duas áreas. Para isso, faz-se uma revisão dos elementos fundamentais necessários à representação da informação e do conhecimento no âmbito da Ciência da Informação. Para tornar possível a comparação entre áreas, apresentam-se os modelos mais consagrados de representação do conhecimento e da informação provenientes da Ciência da Informação, tais como metadados, tesauros e ontologias. No âmbito da Internet, apresentam-se as técnicas de representação com o uso das linguagens de marcação mais utilizadas e suas contribuições para o desenvolvimento da Web Semântica. Encerra-se apresentando uma proposta de estrutura de representação para recursos informacionais, especialmente os disponibilizados pela Internet. A proposta foi desenvolvida a partir dos recursos existentes na Ciência da Computação, particularmente, os oferecidos pela linguagem XML. Contempla a definição de uma ontologia e culmina com a criação de uma estrutura em XML para armazenar metadados de artigos eletrônicos.

Palavras-chave: metadados, linguagens de marcação, ciência da informação, ciência da computação, modelos de representação, ontologias.

Furgeri, Sérgio. *Information and knowledge representation: an analysis of different approaches of Information Science and Computer Science. 2006. Dissertation (Master in Information Science)* - Pontifícia Universidade Católica de Campinas.

ABSTRACT

The Information Science studies new forms of information and knowledge representation for improving efficiency of information retrieval. The dissertation focus on information and knowledge representation, investigate the match and differential points between documentary languages, available in Information Science and markup languages, developed and used in Computer Science, with the objective of identify actions, theories and necessary processes for a bigger integration between the two areas. For this, it is established a revision of the necessary basic elements to the representation of the information and knowledge in the scope of the Information Science. To become possible the comparison between areas, consecrated knowledge and information representation models available in Information Science are presented, such as metadates, thesaurus and ontologies. In the scope of the Internet, the techniques of representation with the use of markup languages more used are presented and its contributions for the development of the Web Semantics. It is locked in presenting a proposal of structure of representation for information resources, especially developed for the Internet. The proposal was developed from the existing resources in the Computer Science, particularly, its available in language XML. It contemplates the definition of a ontology and culminates with the creation of a structure in XML to store metadates of electronic articles.

Keywords: *metadates, markup languages, Information Science, Computer Science, representation models, ontologies.*

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE FIGURAS

LISTA DE ABREVIATURAS

API	=	*Application Program Interface*
CC	=	Ciência da Computação
CDD	=	Classificação Decimal de Dewey
CDU	=	Classificação Decimal Universal
CI	=	Ciência da Informação
DCMI	=	*Dublin Core Metadada Initiave*
DTD	=	*Document Type Definition*
FIFO	=	*First In First Out*
HTML	=	*HyperText Markup Language*
KIF	=	*Knowledge Interchange Format*
KQML	=	*Knowledge Query and Manipulation Language*
LIFO	=	*Last In First Out*
LM	=	Linguagem de Marcação
Modelo E-R	=	Modelo de Entidades e Relacionamentos
OO	=	Orientação a objetos
OWL	=	*Web Ontology Language*
PDA	=	*Personal Digital Assistence*
RBU	=	*Repertoire Bibliographique Universal*
RDF	=	*Resource Description Framework*
RDFS	=	*RDF Schema*
RSS	=	*RDF Site Summary*
SGML	=	*Standard Generalized Markup Language*
SKOS	=	*Simple Knowledge Organisation Systems*
UML	=	*Unified Modeling Language*
URI	=	*Uniform Resource Identifier*
W3C	=	*World Wide Web Consortion*
WAP	=	*Wireless Aplication Protocol*
WML	=	*Wireless Markup Language*
XML	=	*eXtensible Markup Language*

SUMÁRIO

1. INTRODUÇÃO

Este trabalho é uma pesquisa interdisciplinar sobre representação do conhecimento a partir das áreas Ciência da Informação (CI) e Ciência da Computação (CC), buscando estudar conceitos e metodologias mais recorrentes de cada área, e culmina com a proposta de uma estrutura para representação de recursos informacionais na Internet, mais especificamente para artigos científicos em forma eletrônica. Para isso, é apresentada a evolução das principais formas de representação criadas pela CI, seus benefícios e limitações. Descreve também os principais modelos de representação do conhecimento disponíveis na CC e que podem contribuir para a CI.

Entende-se por pesquisa interdisciplinar quando ocorre a convergência de duas ou mais áreas do conhecimento, não pertencentes à mesma classe acadêmica, cujo resultado contribua para o avanço das fronteiras das ciências envolvidas, através da transferência de métodos de uma para a outra e gerando novos conhecimentos ou disciplinas (CAPES, 2003, p. 3; 4).

Desde o surgimento da CI no século passado, renomados autores já apontavam algumas relações existentes entre as duas áreas. Foskett (1980, p.64, *apud* Lima, 2003) relatando sobre a interdisciplinaridade da CI e sua inserção nos campos da ciência relata:

> "uma disciplina que surge de uma *'fertilização cruzada' de idéias que incluem a velha arte da biblioteconomia, a nova arte da computação,* as artes dos novos meios de comunicação e aquelas ciências como psicologia e lingüística que, em suas formas modernas, têm a ver diretamente com todos os problemas da comunicação a transferência do conhecimento organizado". (grifo nosso).

De maneira semelhante, Saracevic (1996, p.48, *apud* Lima, 2003) trata das relações interdisciplinares da CI e cita quatro ciências que mantêm uma estreita

relação com ela: a Biblioteconomia, a Ciência da Computação, a Ciência Cognitiva e a Comunicação. Segundo ele, os pontos principais de convergência e divergência da CI com a CC são descritos nos trechos seguintes:

> "A base da relação entre ciência da informação e ciência da computação *se encontra na aplicação de computadores e na recuperação de informação assim como produtos associados, serviços e redes*" (grifo nosso).
>
> "...[a] *ciência da computação trata de algoritmos* que transformam informações, enquanto a CI *trata da natureza desta informação* e sua comunicação para o uso de seres humanos" (grifo nosso).

De forma semelhante, Ferneda (2003) afirma que as duas áreas possuem propósitos diferentes. Enquanto a CC trata a informação como um simples dado, a CI se preocupa com seu significado. Esse fato contribui para o distanciamento entre as duas áreas.

Rodriguez (2002, p. 42), em sua obra sobre o uso de metadados em bibliotecas digitais, afirma ser importante estabelecer maior sinergia entre as áreas da Biblioteconomia e da Computação.

Mais recentemente, Campos (2004) afirma ser necessário estabelecer um maior diálogo entre as áreas da CI, da CC e da Terminologia, visando à soma de esforços no desenvolvimento de ferramentas para a modelização do conhecimento e buscando estabelecer um núcleo comum de conceitos ao ato de modelar. O objetivo de se criar uma metodologia para realizar a modelagem não se restringe ao simples fato de criar métodos de modelar, mas principalmente melhorar os processos de recuperação, independentemente da área do conhecimento.

De forma semelhante, Bohmerwald (2005) buscou encontrar pontos comuns entre CI e CC para melhorar a interação dos usuários com sistemas

usados em bibliotecas digitais. Devido ao surgimento de novos sistemas, principalmente no que se refere à Internet, os usuários têm sentido certa dificuldade quanto à busca de informação. Seu estudo demonstra a importância do conhecimento nos testes de usabilidade, oriundos da CC, pelos profissionais da CI:

> "Este fato reforça a importância de os profissionais da ciência da informação se dedicarem aos estudos de usabilidade, no âmbito da teoria e da prática, para agregar a eles o conhecimento de sua área, seja ele sobre necessidade, uso ou recuperação da informação".

A melhoria na usabilidade dos sistemas tem uma relação direta com a melhoria na eficiência do usuário quanto à recuperação de informações pertinentes e relevantes.

Por causa dessa estreita ligação entre as duas áreas, diversos cientistas da computação têm unido esforços com a CI para aprimorar sistemas, principalmente no que se refere à recuperação de informação. O corpo docente do mestrado em CI da PUC Campinas tem demonstrado interesse em estabelecer relações com pesquisadores da CC, uma maneira de obter maior sinergia entre as áreas.

Aparentemente, a CC não tem se dedicado a conceitos importantes presentes na CI. Conceitos da Terminologia e de modelagem do conhecimento podem ajudar a reduzir conflitos semânticos e melhorar a documentação de sistemas. O termo revocação, praticamente não citado na CC, pode ajudar na elaboração de sistemas que forneçam resultados mais relevantes aos usuários. Essa falta de cooperação entre as áreas contribui para que os sistemas de armazenamento, representação e, conseqüentemente, de recuperação, não sejam criados de maneira eficiente. Pesquisadores e desenvolvedores de

sistemas computacionais compreendem, e aceitam, que existe uma barreira conceitual muito grande entre usuários e analistas, fato que contribui muitas vezes para o insucesso dos sistemas.

Esse insucesso está ligado, principalmente, ao fato de os sistemas de recuperação não possuírem mecanismos adequados para recuperar informações relevantes quando necessário. Muitas vezes ocorre de a informação estar armazenada, mas não existem mecanismos adequados de recuperação ou o usuário não está preparado para realizar a busca. Será que os conceitos presentes na CI podem contribuir, com os profissionais da CC, na criação de sistemas mais eficazes, isto é, que facilitam o acesso à informação pertinente por parte do usuário?

Com tantas tecnologias computacionais existentes, porque existe um verdadeiro caos na recuperação de informações na Internet? Como é possível melhorar o compartilhamento da informação e do conhecimento através do meio digital?

Este trabalho descreve diferentes linguagens de representação presentes na CC com o intuito de estudar a elaboração de um padrão, mesmo que inicial, para armazenamento de artigos científicos acessíveis pela Internet. Ao se definir um padrão de representação, espera-se que a recuperação de informação traga resultados mais satisfatórios, isto é, as informações encontradas sejam mais relevantes.

Apesar de existirem diversos padrões de representação como, por exemplo, *Dublin Core*, será que não seria mais adequado se o próprio recurso de informação (recurso eletrônico), isto é, o próprio artigo eletrônico, já fosse

elaborado com uma estrutura que o representasse? Ao invés de existirem dois "objetos de informação" separados, o recurso e sua representação, será que não seria mais adequado elaborar um recurso e sua descrição no mesmo "objeto de informação"? E também, não seria mais adequado se a própria documentação fosse realizada concomitantemente com a produção do artigo?

1.1. Objetivos

O objetivo principal se concentra em pesquisar estruturas de representação para recursos informacionais que possibilitem a recuperação de informação e conhecimento de maneira mais efetiva, contemplando o melhor das duas áreas: os preceitos da CI, quanto à estrutura de representação da informação, e os recursos tecnológicos disponíveis da CC. Ao se estabelecer uma estrutura de representação espera-se melhorar a recuperação de informações relevantes.

Como objetivos específicos podem ser citados:

- Estudar os elementos bibliográficos fundamentais necessários à representação da informação e do conhecimento do ponto de vista da CI e da CC.
- Pesquisar a existência de pontos convergentes e divergentes da CI e da CC com relação à representação do conhecimento. Espera-se que isso possa contribuir para o desenvolvimento das duas áreas.
- Realizar uma revisão da literatura referente à representação do conhecimento nas duas áreas citadas.

1.2. Justificativas

O conhecimento não pára de se expandir. Segundo Wersig (1993), a invenção das tecnologias impressas despersonalizou o conhecimento. Antes dessa época, o conhecimento era pessoal, armazenado praticamente de forma

interna ao indivíduo. Com a invenção da impressão surge um problema: compartilhar o conhecimento adquirido e armazenado. Esse fato impulsionou a disseminação do conhecimento para tornar possível se conhecer os resultados de pesquisas de outros pesquisadores. A partir de então, as descobertas científicas deixaram de ser atribuídas a uma única pessoa e o conhecimento passou a ser cada vez mais fragmentado.

O acúmulo de conhecimento passou a ser organizado através da Documentação, uma metodologia usada para facilitar o acesso e recuperação de materiais relevantes. Com o surgimento dos meios eletrônicos, o conhecimento tem se ampliado de forma exponencial, dificultando ainda mais a recuperação de informações relevantes e comprometendo os índices de revocação, principalmente no que se refere às questões interdisciplinares. É preciso que as comunidades científicas disponham de ferramentas para compartilhar o conhecimento adquirido, uma vez que as descobertas representam o resultado de pesquisas conjuntas. Diante desse contexto, torna-se necessário pesquisar formas mais eficazes de representar o conhecimento, visando sua posterior recuperação.

Robredo (2004) afirma que apesar de a tecnologia oferecer soluções para organizar grandes volumes de documentos, a organização da informação neles contida ainda é um problema. Caso não sejam criados melhores mecanismos de representação, a perda de informação nos documentos só tende a aumentar. Para Robredo “é preciso aprofundar e aprimorar os processos de análise da informação e representação do conhecimento”. Não basta apenas se ter tecnologia, é preciso estabelecer critérios de seleção, organização e representação da informação para alcançar maior sucesso na recuperação.

As justificativas citadas relevam a importância da recuperação da informação para os dias atuais e, conseqüentemente, apontam para a necessidade de se criar estruturas de representação mais adequadas ao meio eletrônico. É necessário investigar mecanismos de representação que facilitem a recuperação de fragmentos do conteúdo dos documentos, e não apenas do documento em si. Essa característica permite melhorar a interação com os recursos de informação, tanto por parte de usuários como por sistemas automáticos.

1.3. Metodologia

Considerando-se os objetivos propostos, esta pesquisa pode ser classificada como bibliográfica e exploratória. Para fundamentar o estudo, foi realizada uma reflexão sobre fontes bibliográficas adequadas não apenas no campo da CI, como também na CC, área de suporte para atingir os objetivos apresentados. Primeiramente, é apresentada uma revisão bibliográfica em CI com relação à representação do conhecimento. A seguir, o mesmo procedimento é realizado com a CC. Essas revisões fornecem subsídios adequados para estudar as relações existentes entre as duas áreas. Para fundamentar os conceitos, foi elaborada uma ontologia a partir de um *software* consagrado no mercado na área de representação do conhecimento, a ferramenta Protégé, descrita no Anexo A.

Devido à grande quantidade de documentos disponíveis nas duas áreas, não é possível analisar toda a bibliografia relacionada. Para isso, foi necessário estabelecer critérios para a seleção dos documentos. Optou-se por selecionar os textos mais relevantes e que possuíam autores mais consagrados. Para o levantamento do material bibliográfico foram contatados alguns especialistas que, em consultas informais, indicaram os autores mais relevantes.

O trabalho buscou, através da análise e da síntese, investigar pontos de convergência e divergência entre as áreas quanto à representação do conhecimento, estabelecendo comparações entre o objeto da representação (um artigo, por exemplo) e suas formas de representação (metadados, cabeçalhos, etc).

1.4. Resultados Esperados

Conforme citado anteriormente, este trabalho pretende abordar conceitos de duas áreas distintas, CI e CC, com relação à representação do conhecimento. Com isso, espera-se:

- Identificar os conceitos convergentes e divergentes entre as áreas, contribuindo para a melhoria da representação da informação e do conhecimento em ambiente Internet e, conseqüentemente, para a criação de sistemas de recuperação mais eficazes;
- Unir o melhor das duas áreas com relação à representação da informação e do conhecimento, contribuindo na elaboração de uma estrutura alternativa de representação de recursos informacionais. Espera-se que essa estrutura possa ser um ponto de partida no aumento de eficácia dos sistemas de recuperação em ambiente Internet;
- Os resultados deste trabalho possam ser divulgados para a comunidade através da redação de um artigo e/ou congressos;
- Estimular uma maior aproximação entre as áreas citadas, abrindo caminho para que outros pesquisadores possam dar prosseguimento no estudo da representação da informação e do conhecimento.

1.5. Estrutura da Dissertação

Para atingir os objetivos apresentados, este trabalho encontra-se assim definido: o capítulo 2 realiza uma revisão bibliográfica referente aos aspectos conceituais pertinentes à elaboração do trabalho, tais como representação da

informação, a evolução do conhecimento e suas diferentes formas de representação, tais como sistemas de classificação, tesauros e metadados. Descreve ainda o padrão *Dublin Core* e sua importância na representação da informação. O capítulo 3 realiza uma análise das formas de representação do conhecimento disponíveis na CC, em especial as voltadas ao uso na Internet, e suas relações com a CI. O capítulo 4 apresenta diversas relações existentes entre as duas áreas, descrevendo relações interdisciplinares entre as áreas nas questões de representação. O capítulo 5 apresenta a elaboração de uma ontologia e uma possível estrutura de representação de informação para recursos informacionais, analisando seus benefícios e limitações de escopo. O capítulo 6 apresenta as conclusões alcançadas pelo trabalho.

2. REVISÃO BIBLIOGRÁFICA DA CIÊNCIA DA INFORMAÇÃO

Este capítulo apresenta uma revisão bibliográfica da CI pertinente ao tema estudado. Descreve diversos conceitos relevantes a respeito de informação e conhecimento buscando, sempre que possível, estabelecer suas relações com a CC. Encerra, apresentando uma análise comparativa entre as principais formas de representação do conhecimento disponíveis na CI.

2.1. Dado, Informação e Conhecimento

Autores consagrados da CI, uma ciência com desenvolvimento recente, cujo processo evolutivo encontra-se descrito em Pinheiro (2005), têm se desdobrado na difícil tarefa de estabelecer os limites entre os conceitos sobre dado, informação e conhecimento.

Mesmo assim, não há consenso sobre os limites pertinentes aos conceitos citados como pode ser notado em Oliveira (2005), entretanto, Robredo (2003) considera a definição proposta por Boisot como a mais adequada. De acordo com o Boisot, o dado não tem significado próprio, algo atômico que atua como uma matéria-prima a partir da qual a informação é produzida. A informação se refere ao resultado da interação dos dados para produzir ativação e significado; já o conhecimento se refere ao uso prático da informação na formação ou transformação de algo.

De forma semelhante, a CC considera o termo dado como uma estrutura elementar, uma base a partir da qual um problema pode ser resolvido. O termo “processamento de dados” estabelece que o computador recebe dados, faz um tratamento sobre eles e fornece um resultado (uma informação). De forma geral, a

CC considera que uma informação é o resultado do processamento dos dados.

Segundo Robredo (2003) existe uma multiplicidade de definições para o termo informação, um termo “escorregadio” que varia de acordo com a área do conhecimento relacionada, fato que provoca uma definição redutora, isto é, não abrangente. Para melhorar sua comparação apresenta uma série de definições sobre informação retiradas de vários dicionários e conclui que, mesmo assim, não é possível definir seus limites, fato que torna o termo interdisciplinar, isto é, definido em diversas áreas do conhecimento.

Apesar de o grande potencial da informação, as divergências presentes em definições e paradoxos, produzem muito mais discordância do que esclarecimentos, descaracterizando totalmente o termo nas diferentes áreas. Esse é um problema da CI, uma vez que a informação não pode ser considerada, isoladamente, como um objeto de estudo. As descrições presentes aqui procuram apresentar aspectos relevantes ao tema estudado, tentando identificar uma série de características que interessam diretamente a CI, isto é, o aspecto semântico da informação. Esse aspecto é fundamental para a representação do conhecimento e, conseqüentemente, para o desenvolvimento de sistemas mais eficientes de recuperação.

Segundo Le Coadic (1996, p. 4), informação é “um conhecimento inscrito (registrado) em forma escrita (impressa ou digital), oral ou audiovisual, em um suporte” e “comporta um *elemento de sentido*” (grifo nosso). Através dessa definição é fácil identificar a importância do sentido, isto é, do aspecto semântico inerente à informação. Para que algo seja uma informação é preciso que o receptor tenha condições de identificar o sentido daquilo que está se expressando.

Devido à natureza do presente trabalho, preocupado com o processo de representação de informação e conhecimento e sua posterior recuperação, considera-se relevante o conceito de informação como coisa, preconizado por Buckland (1991), uma vez que o objetivo maior da representação é permitir a recuperação de algo, no caso, a própria informação. Buckland considera a informação como "algo usado, atribuído a objetos, tal como dados e documentos que se referem à informação, porque deles se espera que sejam informativos". Nesse contexto, a própria representação do conhecimento pode ser considerada como coisa, uma vez que ela procura substituir aquilo que representa, algo que mantém informações sobre um domínio qualquer e de forma semântica.

Diante do exposto, pode-se conceber que o dado gera informação e esta, conhecimento. Essas definições vêm ao encontro deste trabalho, uma vez que a proposta foi desenvolver uma estrutura de representação que armazena dados com um certo significado e que podem ser usados na representação do conhecimento.

2.2. Representação da Informação

De acordo com McGarry (1999, p.11), "a informação deve ser ordenada, estruturada ou contida de alguma forma, senão permanecerá amorfa e inutilizável". Segundo ele, é necessário que a informação receba um tratamento para ser compreensível aos seres humanos, ela *deve ser representada de alguma forma* para que tenha sentido.

Ainda segundo McGarry (1999, p.12), "A informação, portanto, deve ter alguma forma de *veículo*. Este veículo deve possuir um atributo essencial para que possa ser compreendido pelo receptor". O autor distingue três tipos de

veículos: sinais, signos e símbolos.

Os sinais estabelecem relações com as ações a serem desenvolvidas pelo receptor. Um sinal pode sinalizar que um determinado evento está para ocorrer. Uma pessoa, ao se levantar pela manhã e abrir a janela de seu quarto, pode concluir que existe uma boa probabilidade de chuva, ao visualizar o tempo nublado ou ouvir trovões. Esses sinais podem indicar a necessidade de levar o guarda-chuva para ir ao trabalho. De forma simples, pode-se perceber que o receptor recebeu uma informação ao identificar os sinais. Uma pessoa com deficiência visual, ou auditiva, não poderia identificar esse fato, isto é, os sinais não teriam atributos essenciais para serem compreendidos pelo receptor.

Os signos indicam a presença física de algo ou algum evento relacionado a eles. A fumaça indica fogo, o aumento da temperatura do corpo pode indicar a febre, ou um simples olhar entre duas pessoas pode ter um sentido ou significado conhecido.

Os símbolos tendem a possuir significados mais duradouros e constituem-se em representações culturalmente construídas e reconhecidas por uma comunidade específica. Um símbolo pode ter ou não semelhanças com o que representa. Um ícone de uma impressora, por exemplo, pode ter uma boa semelhança com aquilo que pretende representar e pode ter sentido para pessoas do mundo todo, desde que consigam identificar o objeto real. Já a palavra impressora terá sentido apenas para quem conhece, ao mesmo tempo, o objeto real e a língua portuguesa.

Além dos três tipos de veículos descritos, McGarry (1999, p.17) considera ainda um outro veículo de comunicação e transmissão de informação: “a

linguagem é o veículo fundamental da comunicação humana". Segundo ele, é mais fácil identificar o que uma linguagem faz que defini-la. Através do exposto, parece ser mais fácil relacionar a linguagem aos símbolos, uma vez que uma linguagem falada ou escrita está sempre restrita a uma comunidade, seja ela de humanos ou máquinas.

Mesmo utilizando-se dos veículos de comunicação supra citados, é difícil realizar a representação da informação de forma adequada. Alguns autores citados em Cordeiro (1996) consideram a representação da informação como um simulacro: "A idéia de simulacro pode ser trabalhada em nível da analogia com o modelo de representação da informação". Isso se deve ao fato da dificuldade em realizar uma representação de forma correta, já que o próprio processo de representação é "um processo redutor da informação" realizado por um intermediário.

De um modo geral, a representação de uma informação passa pela definição de um conjunto de elementos básicos e de regras para a conjunção desses elementos. Processo semelhante ocorre na CC, em que os computadores digitais representam uma informação qualquer através de um conjunto de elementos (bits) e regras específicas (agrupamento de bits para formar caracteres). Da mesma forma, a criação de *frames* e redes semânticas define uma série de elementos e regras entre esses elementos, conforme descrito no capítulo 3.

Com o desenvolvimento dos capítulos ficará evidente que a linguagem tem sido fundamental não apenas na transmissão de informações entre seres humanos, mas também entre computadores. O uso de linguagens tem propiciado o compartilhamento do conteúdo de documentos entre computadores, fornecendo

um certo sentido as informações armazenadas.

2.3. Evolução do Conhecimento

O Conhecimento já foi concebido como produto da inteligência divina, como uma dádiva de Deus aos homens, como fruto da razão e como produto da experiência. Durante toda a história da humanidade, o conhecimento sempre foi visto como um bem precioso, um mecanismo de dominação e controle, entretanto, parece ser nos dias atuais que sua posse está contribuindo cada vez mais para o avanço de pessoas, organizações e Estados. Em contrapartida, sua falta está produzindo diferenças sociais e econômicas extremamente acentuadas. É fácil verificar que os países mais ricos são os detentores do conhecimento, sobretudo o tecnológico e científico.

Os filósofos da idade moderna Descartes e Locke mudaram a compreensão de conhecimento: do conhecimento no âmbito da divindade para o humano, colocando-o entre o pensamento e o objeto de estudo, como cita Aranha & Martins (1993): “o conhecimento é o pensamento que resulta da relação que se estabelece entre o sujeito que conhece e o objeto a ser conhecido”. Burke (2003) alerta para a importância do conhecimento na atualidade quando cita a era da “sociedade da informação” ou “sociedade do conhecimento”.

A notável importância do conhecimento para a sociedade atual nos remete a necessidade de seu estudo e investigação. A seguir são apresentadas diversas passagens históricas ligadas ao ato de conhecer, cujo resultado dessa relação é o acúmulo de conteúdo produzido pela humanidade.

Segundo Ortega (2004), desde o surgimento da escrita, aproximadamente 3000AC, já se registravam documentos em tábuas de argila em bibliotecas

primitivas de Elba, na Síria. Mais tarde, o mesmo ocorreu na Mesopotâmia, depois na Assíria, Grécia, etc. Observa-se que desde os tempos antigos o homem já começava a se preocupar em armazenar o pensamento, as informações e conhecimentos adquiridos. No entanto, na Grécia antiga, conhecer não significava representar. Esse fato não motivava a representação do conhecimento de maneira eficaz.

Em Moraes e Arcello (2000), observa-se que notórios filósofos do mundo ocidental como Platão e Aristóteles, cerca de 300 AC, propuseram uma mudança de paradigma: pensar e expressar a verdade através da razão, repudiando o mito como uma forma de conhecimento. A partir de então, os filósofos passaram a possuir seguidores (discípulos) transferindo o conhecimento de geração a geração.

Essencial para sua sobrevivência, o homem buscou aperfeiçoar seu conhecimento visando transformar o mundo e dominar tudo o que nele existe. Burke (2003) em seu livro “uma história social do conhecimento”, cita diversas vezes o uso do conhecimento como instrumento de dominação dos povos, como no trecho: “... conquistadores britânicos empregaram seu conhecimento das línguas e do direito indianos para impor seu domínio mais eficazmente”.

Nota-se, principalmente, que a partir da idade média, a prática de coletar informações e armazenar o conhecimento torna-se mais evidente. Com isso os governos poderiam melhor controlar diversos eventos como guerras, pestes, etc. Esse fato produziu o surgimento da burocracia, cuja função era exercer o controle do povo com base no conhecimento. É interessante observar que esse fato, isto é, esta busca por informação como forma de controle, acaba gerando novas necessidades, um tipo de ansiedade por mais informação, para se exercer mais

controle ainda.

Ainda na idade média, Burke descreve diversos fatos onde apresenta a Igreja Católica como um modelo a ser seguido na busca pela informação e conhecimento. Na época era a principal instituição preocupada com a coleta de informações, tais como registros de nascimentos, casamentos e mortes. Segundo Ortega (2004), a ordem religiosa estava também preocupada em manter diversas bibliotecas. Diversas outras instituições, e até mesmo o Governo, buscaram na Igreja sua fonte de conhecimento. De qualquer forma, nota-se claramente o uso do conhecimento como instrumento poderoso nas mãos da classe dominante.

Pensadores da Idade Moderna entre os séculos XV e XVII, como Galileu e Copérnico, estabelecem a consolidação do conhecimento, desafiando o marco teórico mantido pela Santa Inquisição, propondo um conhecimento com base na observação e na experimentação. A partir de Descartes e Locke, o conhecer das sensações e percepções deu lugar ao pensamento, passando a estar relacionado ao representar, ou seja, o conhecer passou a ser “o saber representar”.

A representação do conhecimento firma-se ainda mais quando a constituição de todo pensamento passa a ser realizado através de uma linguagem. Nesse momento, a ciência passa a ser uma descrição do que nela se passa e contribui ainda mais para a necessidade do registro dos conhecimentos. Mais à frente, entre os séculos XVII e XVIII, Comte estabeleceu a primeira forma de hierarquização e os limites de cada ciência (MORAES e ARCELLO, 2002). Nessa época, Francis Bacon em seu livro “O Avanço do Conhecimento”, de 1605 já acreditava que o conhecimento era algo passível de ser acumulado, melhorado e aperfeiçoado. Bacon, Wilkens e Linnaeus foram os pioneiros em tentar classificar o conhecimento (WRIGHT, 2003). A descoberta do infinito e o

surgimento do capitalismo exigiram uma nova concepção do conhecimento.

Desde 1850, com o aumento significativo do número de periódicos, começou-se a estudar maneiras mais adequadas para representar e recuperar a informação, pois manter documentos com conteúdos semelhantes, próximos uns dos outros, como vinha sendo feito até então, já não era suficiente para se recuperar conteúdos relevantes.

A necessidade de se registrar o conhecimento torna-se ainda mais imprescindível quando Marx afirmou que nada do que é conhecido é absoluto: “A ciência constitui portanto uma empresa sem fim, e em certo sentido sem sucesso absoluto, pois o que pode conhecer consiste sempre naquilo que, tendo o estado atual por base, lhe é possível conquistar no futuro” (PINTO, 1979, p. 200).

Muito do que se acredita ser verdade hoje pode não sê-lo amanhã. O momento histórico atual está cheio de exemplos desse tipo, pois os avanços tecnológicos têm atuado como um verdadeiro instrumento revelador e formador de conhecimento. Por exemplo, em novembro de 2005 um programa de TV apresentou uma reportagem que demonstrava já ser possível visualizar claramente, em detalhes, a criança dentro do útero de sua mãe. Isso torna possível diagnosticar e tratar a criança mesmo antes dela nascer. Com certeza muito conhecimento pode ser adquirido com o uso de tecnologias desse tipo. A ciência possui esta característica, está em constante movimento. Se os fatos históricos são importantes, obviamente o registro de todos os conhecimentos históricos se torna imprescindível, mesmo que o tempo venha a demonstrar que esses conhecimentos estavam incorretos.

No final do século XIX, em função da evolução e ampliação constante em

diversos campos do conhecimento, alguns estudiosos, como Otlet e La Fontaine, começaram a questionar a verdadeira função do acervo de uma biblioteca: apenas manter o acervo ou atuar como um serviço de informações e divulgação do conhecimento?

Era necessário analisar mais a fundo o verdadeiro conteúdo dos documentos. Otlet é considerado um dos precursores deste movimento; ele criticava as políticas de seleção das bibliotecas e as resistências oferecidas às inovações técnicas. Provavelmente, Otlet sofreu essa resistência dos bibliotecários da época porque se tratava de uma mudança de paradigma para as bibliotecas, pois, antes disso, elas trabalhavam apenas com o armazenamento e levantamento bibliográfico. Fornecer um serviço de informações seria uma mudança radical (ORTEGA, 2004). Em 1895, Otlet e La Fontaine criaram o *Repertoire Bibliographique Universel* (RBU), uma ambiciosa tentativa de desenvolver uma bibliografia-mestre do conhecimento mundial acumulado.

Uma breve história de Otlet, chamado de “o antepassado esquecido”, é narrada em Wright (2003). Paul Otlet visionou um novo tipo de estação de estudos: uma mesa móvel construída como uma roda, ligada por uma rede de raios de roda dobradiços sob uma série de superfícies móveis. Seu empreendimento foi concebido como um grande “réseau” (teia) de conhecimento humano. As propostas de Otlet foram fundamentais para o surgimento da Documentação e, mais à frente, da CI. De qualquer forma, parece ter sido Otlet um dos mais preocupados, até então, com a representação do conhecimento, ou melhor, com a recuperação de informações relevantes.

Bush, em seu artigo “*As We May Think”*, visto o grande crescimento de periódicos (e do conhecimento acumulado) já alertava sobre o tempo gasto entre

a redação de um trabalho científico, sua publicação e posterior avaliação por outros cientistas da mesma área (BUSH, 1945). Na época, Bush sugeriu um equipamento chamado Memex, cujo princípio de funcionamento era semelhante ao de um computador conectado a Internet. Pode-se dizer que uma das funções do equipamento era o de contribuir para a criação e distribuição do conhecimento adquirido em pesquisas, facilitando o compartilhamento e busca de informações.

Apesar de toda a evolução tecnológica a partir do século XX e de seu uso intensivo até os dias atuais, na perspectiva da CI, constata-se uma turbulência no campo de conhecimento, especialmente quanto à representação, à armazenagem e recuperação de informações, áreas intensamente relacionadas à cognição humana. Torna-se necessário verificar se a representação do conhecimento está sofrendo impacto positivo na era digital, em especial nas questões da socialização da informação.

Segundo Oliveira (2005), essa parece ser uma preocupação do Banco Mundial ao sugerir que “um novo olhar, um olhar do ponto de vista do conhecimento, seja utilizado para pensar a questão do desenvolvimento nesta era da informação”. O relatório aponta para os problemas decorrentes da falta de conhecimento tecnológico e das dificuldades de se interpretar informações pela falta de conhecimento anterior.

Nesse contexto, torna-se necessário se criar melhores mecanismos de representação do conhecimento, em especial dirigidos ao meio digital, pois isto poderá proporcionar uma melhor recuperação e disseminação de informações. Torna-se necessário identificar quais são as formas de representação mais adequadas considerando-se, principalmente, as mídias digitais.

2.4. Tipos de Conhecimento

O que se pretende nessa seção é apresentar algumas concepções de conhecimento. Não existe uma definição única a todas as áreas, nem mesmo na CI. Existem muitas definições sobre conhecimento, dependendo do contexto onde ele está inserido. Mesmo na filosofia, correntes diferentes como o empirismo e o racionalismo, consideram diversos tipos de conhecimento. O mais sensato talvez seja conceber as diferentes formas de interpretar o conhecimento de maneira complementar.

Ricco (2004), da área de Administração, define cinco tipos de conhecimento esperados para um líder: autoconhecimento (conhecer seus próprios pontos fortes e fracos), conhecimento do trabalho (o que deve se conhecer para realizar as rotinas empresariais), conhecimento da organização (conhecer a cultura organizacional de maneira a ser mais eficiente no trabalho), conhecimento do negócio (conhecer os clientes e suas expectativas) e conhecimento do mundo (conhecer a comunidade mundial e suas relações com a empresa).

Segundo Nonaka & Takeuchi (1997), autores com foco na Gestão do Conhecimento nas organizações, existem dois tipos de conhecimento: explícito e tácito. Enquanto o conhecimento explícito pode ser facilmente documentado e reproduzido, o conhecimento tácito é muito difícil de ser documentado, uma vez que faz parte da experiência individual de cada pessoa, algo interno e difícil de ser expresso em palavras. De acordo com os autores, o segredo da captura do conhecimento corresponde à conversão do tácito em explícito: “A pedra fundamental da nossa epistemologia é a distinção entre conhecimento tácito e explícito, o SEGREDO para a criação do conhecimento está na mobilização e

conversão do conhecimento tácito". A partir dessa definição os autores propõem quatro formas de conversão entre os dois tipos de conhecimento.

De forma semelhante, Oliveira (2005) apresenta dois tipos de conhecimento: explícito e tácito. A autora defende a idéia de que Informação e Conhecimento são partes do mesmo processo de construção: "O que sugiro aqui é que a diferença entre informação e conhecimento seja mais como uma diferença de graus em uma escala do que uma distinção de categoria", no entanto destaca que enquanto a informação não tem aplicação e é fácil de ser armazenada, o conhecimento é aplicável e difícil de ser armazenado.

Segundo Piaget (1976 *apud* Rosa, 2005), o conhecimento é o ato de interpretação de um sujeito que, diante de uma nova realidade ou uma situação problema, aciona esquemas de assimilação e os modifica. O processo de construção é assim um processo de reestruturação no qual todo conhecimento novo é gerado a partir de outros prévios. Para ele existem duas formas de se adquirir conhecimento: o conhecimento físico que consiste no sujeito explorando os objetos e o conhecimento lógico-matemático que consiste no sujeito estabelecendo novas relações com os objetos que não têm existência na realidade externa, está presente apenas na mente do indivíduo.

Há ainda outros tipos de conhecimento presentes na literatura: o inato, embutido na razão dos indivíduos e o empírico, adquirido através de experiências sensoriais. De acordo com Lara (2002), o conhecimento empírico evolui dependendo da experiência anterior do leitor, de sua motivação e de seus objetivos de leitura, mais seu estoque prévio de conhecimento. O conhecimento é dinâmico, pois um mesmo texto pode ser interpretado de formas diferentes em função da vivência do leitor, ou em função da motivação que este possui no

momento em que está realizando a leitura, ou ainda de suas pretensões (apenas dar uma olhada ou entender em profundidade). O mesmo leitor pode ainda ter compreensões diferentes do mesmo texto, caso realize sua leitura diversas vezes, ou em momentos diferentes. A cada leitura descobre algo novo em função da experiência adquirida na leitura anterior. O conhecimento está em constante construção. Essa é uma das preocupações da CC, mais especificamente no campo da Inteligência Artificial, em que o *software* busca "acumular conhecimento", a partir de um conhecimento prévio.

Mesmo levando-se em consideração as diversas compreensões sobre o que vem a ser o conhecimento, existe uma concordância unânime: a importância do conhecimento e de sua socialização. Para socializar o conhecimento é necessário representá-lo de forma eficiente para que possa ser recuperado. A seção seguinte aborda a importância da representação do conhecimento.

2.5. A Função Prática da Representação do Conhecimento

Qual motivo de se representar o conhecimento armazenado? Para Moraes e Arcello (2002) "as representações são instrumentos de ordenação e hierarquização da estrutura social e identificam o grupo ou meio que as produziu e que as consome". Com base nessa afirmação, pode-se considerar que o currículo de um candidato representa seu conhecimento profissional, desde que, aquele que o analisa, tenha plenas condições de interpretá-lo.

Da mesma forma, a grade curricular de um determinado curso, deve ser capaz de expressar o conhecimento que os alunos devem tomar posse. Sendo assim, é possível afirmar que cada área científica, ou cada comunidade de trabalho, independentemente da atividade, necessita de um sistema de

representação, de maneira a delimitar, interpretar e recuperar seu conhecimento. Em outras palavras, cada comunidade precisa definir um sistema de conceitos de maneira que todos possam interpretar os mesmos assuntos de forma equivalente.

O estabelecimento de conceitos pode dificultar a transmissão de informação e conhecimento, caso um determinado conceito não seja conhecido. Na construção de representações, Lara (2002) aponta que a analogia é um método muito importante a ser considerado, citando o caso da analogia de Marco Pólo, quando avistou o Rinoceronte e o comparou ao Unicórnio.

Essa “estratégia” de transmissão do conhecimento através de analogias é utilizada freqüentemente por professores ao se ensinar um novo conceito. È muito comum, por exemplo, no ensino da eletrônica e física se remeter a analogias com a hidráulica, mecânica, etc. Quando um assunto é novo, ou pelo menos desconhecido em parte, pode-se tornar difícil segmentá-lo em grupos de “coisas” conhecidas. Essa afirmação reflete o que diz o texto: “Ao tentar corrigir a descrição inicial dos unicórnios, Marco Polo modifica a intenção, deixando a extensão sem juízo", isto é, altera a classe original, modificando as propriedades dos unicórnios. Quando isso ocorre pode surgir uma nova segmentação de conteúdo, no caso de Marco Polo, poderia acrescentar um novo animal na classe de seres vivos.

Segundo Campos (1995), a classificação que estabelece relações entre itens de informação “não pode mais ser vista em seu sentido restrito de estruturas hierárquicas”. A criação de um sistema de conceitos é cada vez mais necessária, cujo objetivo é o de melhorar a representação do conhecimento e, conseqüentemente, melhorar a recuperação de informações e agilizar o processo de transmissão do conhecimento.

De forma semelhante, a CC também estabelece a classificação em diversas áreas. Na modelagem de dados, mais especificamente no Modelo E-R (Modelo de Entidades e Relacionamentos, descrito na seção 3.5.8), busca-se classificar as entidades de acordo com as propriedades que elas contêm. Na herança, um item da Orientação a Objetos (3.5.9), busca-se classificar os objetos através de uma estrutura hierárquica, estabelecendo uma relação de generalização e especialização entre classes semelhantes. O mesmo ocorre com redes semânticas (3.4.2) e *frames* (3.4.3), importantes modelos de representação do conhecimento.

A principal função da representação é criar uma estrutura eficiente com fins de recuperação de informações. No entanto, a transferência do conhecimento (ou da informação) através de sua representação é algo impreciso. Na descrição do conteúdo de uma obra são utilizadas palavras-chave que resumem um assunto, porém essas palavras são representações apenas parciais e “longes” da perfeição. Efeito semelhante ocorre na CC ao buscar representar o mundo real em sistemas de *software*.

Realmente parece algo muito pretensioso que, por exemplo, apenas um pequeno número de palavras-chave tente representar todo o conteúdo de um livro ou artigo, ou que um sistema computadorizado tente criar na máquina um ambiente real e ideal. Muitas informações podem ser desconsideradas e perdidas durante esses processos. O grande problema parece estar da diversidade de representações que podem ser consideradas em função do momento histórico, da cultura, do ambiente inserido ou até mesmo da área de atuação considerada.

Considerando-se esse contexto, não pode ser mais aceitável que uma biblioteca mantenha seu acervo considerando-se apenas a classificação

bibliográfica como critério de pesquisa. Em tempos de buscadores, como Google, isso soa como um contra-senso. Na era digital não é mais admissível considerar uma pesquisa apenas por seus dados bibliográficos. Torna-se necessário criar uma estrutura de representação mais eficiente. Ao invés de ser retornada toda a obra, deve ser possível, por exemplo, retornar apenas a parte relevante do conteúdo de um documento (como o resumo, a conclusão, etc).

A segmentação do conhecimento tem ligação direta com as duas ciências analisadas. Na CI para se organizar o conteúdo de documentos torna-se necessário agrupá-los de alguma forma, seja através de recortes, classificações ou segmentações, isto é, devem existir termos associados que permitam comunicar e compartilhar novas idéias. Em função do uso de analogias, e da segmentação do conhecimento, a representação da informação de conteúdo pode se tornar algo impreciso, ou seja, uma mesma quantidade de documentos pode ser organizada e representada de maneiras diferentes dependendo da experiência de quem realiza tal procedimento. O mesmo problema ocorre na CC ao se projetar um modelo E-R, ao se criar uma estrutura hierárquica de classes na Orientação a Objetos (OO), ou ainda na criação de *frames* ou ontologias.

De maneira oposta ao modelo apresentado por Shannon & Weaver (1949) *apud* Wolf (1999, p. 113), onde existe preocupação apenas com a comunicação do sinal entre transmissor e receptor dentro de um conceito cibernético, buscando a maximização do canal de transmissão, uma das preocupações da CI, já citada anteriormente, se refere ao significado das informações transmitidas. Nesse contexto, a definição de conceitos para a representação do conhecimento torna-se essencial.

2.6. Representação do Conhecimento

Almeida (2005) descreve o processo de representação da seguinte forma: “As representações são conhecimentos construídos socialmente por uma comunidade ou grupo de sujeitos”. Para se representar algo é necessário que o fenômeno observado e suas representações estejam assentadas na consciência do grupo e pressupõe a utilização de categorias, classes ou modelos definidos pelos integrantes.

Alvarenga (2003) parte do princípio que representar o conhecimento significa o “ato de colocar algo no lugar de”. Por causa disso considera essencial o papel dos autores no processo de representação do conhecimento, um processo cognitivo que terá como resultado a expressão dos pensamentos, observações e metodologias aplicadas pelo autor da representação. Para realizar o processo de representação é necessário que o autor utilize uma linguagem apropriada, condizente com o meio social. Alvarenga sustenta que:

> “o processo de representação possui as etapas de percepção, identificação, interpretação, reflexão e codificação, etapas que são envolvidas no ato de se conhecer um novo ser ou coisa, ou aprofundar-se no conhecimento de um ser ou uma coisa já conhecida, utilizando-se dos sentidos, da emoção, da razão e da linguagem”.

Ao descrever o processo de representação no âmbito de um sistema informações, Alvarenga afirma que a representação pode ocorrer em três momentos distintos. Em todos eles o processo cognitivo acha-se presente:

- **Momento 1**: durante a produção dos registros de conhecimento: estágio anterior à entrada dos itens no sistema de informações, contendo diversas etapas de cognição dos envolvidos no processo (produtores de documentos, autores, revisores e editores). Essa etapa corresponde, portanto, a produção daquilo que será representado;

- **Momento 2**: na organização dos sistemas de informações documentais: estágio onde os itens são incluídos no sistema respeitando-se o processo cognitivo do item anterior, isto é, os itens produzidos no estágio anterior são inseridos no sistema de forma organizada;
- **Momento 3**: no acesso às informações pelos usuários: estágio pós-inclusão do item no sistema, quando nova etapa de cognição se processa. Ocorre o contato do usuário com o sistema de informação com fins de recuperação.

Com a evolução do conhecimento foram estabelecidas diversas formas de representação. Desde a época da "árvore de Porfírio" e, mais a frente, a "árvore baniana" de Ranganathan, diversas foram as maneiras de se representar o conhecimento. Davis e Walter (2003) registraram quinze formas diferentes de representação do conhecimento. A grande maioria delas encontra-se descrita neste trabalho.

Pode-se antecipar que existem características comuns entre diversos tipos de representação do conhecimento. Esta seção aborda as características dos sistemas de classificação, unitermo, tesauros e metadados, buscando frisar seus pontos fortes e fracos e suas contribuições para a recuperação da Informação. Vale a pena ressaltar que alguns autores parecem considerar "organização do conhecimento" como similar a "representação do conhecimento", como pode ser visto em Tristão *et al.* (2004).

2.6.1 Sistemas de Classificação

Segundo Tristão *et al.* (2004) classificação significa "a ação e efeito de classificar, e classificar significa ordenar e dispor em classes". Uma classe consiste "de um número de elementos quaisquer (objetos e idéias) que possuem alguma característica comum pela qual podem ser diferenciados de outros

elementos”. Dessa forma, a classificação permite organizar diversas coisas ou objetos, agrupando-as de acordo com as relações existentes. O resultado da classificação é uma estrutura de relações existentes em uma área do conhecimento.

Uma das primeiras formas de organização foi a “classificação do conhecimento”. Estudiosos da antiguidade como Aristóteles, Porfírio e Bacon já buscavam formas de classificar o conhecimento humano. A evolução detalhada das tentativas de classificação é encontrada em (KAULA, 1984). Em 1916, Brown definiu que a classificação era um "processo mental" executado de forma consciente e inconsciente por qualquer ser humano e, portanto, deveria ser considerado como um dos mais importantes campos do conhecimento.

O aumento do número de obras tornou a “classificação do conhecimento” impraticável, fato que contribuiu para a “classificação por assunto”. O primeiro tipo de classificação por assunto, talvez uma das principais formas de representação do conhecimento da época, foi proposto em 1876 por Dewey através da CDD (Classificação Decimal de Dewey*)* que já passou por mais de vinte revisões até hoje. A CDD organiza todo o conhecimento em dez classes principais através de números decimais, um mapa completo das áreas do conhecimento, mostrando seus conceitos e suas relações. O sistema é composto por dez categorias para representar todo o conhecimento humano. Cada classe principal contém dez divisões que podem ser novamente subdivididas, permitindo a inserção de novos temas se necessário.

Alguns anos depois os belgas Otlet e La Fontaine criam a CDU (Classificação Decimal Universal), uma adaptação da CDD. Apesar de existirem algumas diferenças entre elas, a base permanece a mesma. Apesar de serem

usadas até hoje existem basicamente dois problemas relacionados a CDD e CDU: a dificuldade de acesso por parte do usuário comum, pois este fica perdido em relacionar a seqüência numérica e sua localização no espaço físico, e o fato de ambas terem sido criadas com o objetivo de organizar fisicamente as coleções de documentos e não organizar os assuntos presentes nesses documentos, pois a unidade a ser classificada corresponde a todo o documento. Se a menor unidade fosse o conceito seria possível uma infinidade de arranjos e combinações necessárias para representar qualquer assunto. Em tempos de digitalização fica difícil acreditar que esse tipo de representação possa sobreviver por muito tempo.

Outro tipo de representação, proposto por Ranganathan, tinha por objetivo principal garantir uma seqüência útil dos livros nas estantes. A preocupação se referia à localização física dos livros na biblioteca em relação ao tema central abordado na obra. Os livros eram organizados por assunto, isto é, as "idéias centrais" semelhantes ficavam localizadas próximas umas das outras, uma tentativa de facilitar a localização de diversas obras que abordavam assuntos semelhantes (DAHLBERG, 1972). Essa organização não necessariamente tinha relação direta com os assuntos dos livros.

Apesar de sua contribuição e utilidade até os dias atuais, os sistemas de classificação, em especial a CDD e a CDU são inadequados para diversas aplicações. Segundo Shera e Egan, *apud* Tolmasquim *et al.* (1998):

> "os diversos domínios do conhecimento ou da atividade focalizam diferentes pontos na escala ascendente de generalidade. Este é um dos fatores que tornam qualquer esquema de classificação universal inadequado à maioria das finalidades especiais, dando assim ensejo à multiplicidade de sistemas de classificação, cada qual focalizando o nível de generalidade ou particularidade que é fundamental em seu próprio contexto"

A CC também se depara com esses problemas de generalização e especialização na elaboração de sistemas. Diferentes sistemas podem necessitar de especificações diferentes. Para contornar esse problema existem diversas técnicas usadas na tentativa de reaproveitar um código genérico em diferentes aplicações. Exemplos dessas técnicas são: *Procedures*, *Functions*, *Libraries* e APIs (*Application Program Interface*). Essa busca pela reutilização do código se tornou mais marcante com o surgimento da Orientação a Objetos (3.5.9).

2.6.2. Cabeçalhos de Assunto

Segundo Gomes e Marinho (1984) o sistema de cabeçalhos de assunto, desenvolvido na Biblioteca do Congresso em Washington, foi desenvolvido com o objetivo de facilitar o acesso à biblioteca por parte do público comum, uma vez que os sistemas até então disponíveis eram voltados apenas ao público erudito.

No sistema de cabeçalhos de assunto um conjunto de palavras é usado para representar o conteúdo de um documento. Cada vez que uma nova obra não pode ser representada pelos cabeçalhos existentes um novo cabeçalho deve ser criado (*ad-hoc*). O principal aspecto negativo dos cabeçalhos se refere a sua impossibilidade de seu conjunto representar as diferentes áreas de assunto existentes numa base. Apesar de todo cabeçalho representar pelo menos um assunto presente em uma obra, o conjunto deles não consegue representar o todo. Além disso, um documento não poderia ser encontrado quando se usasse como critério de busca uma idéia secundária, não existente no cabeçalho (GOMES, 1996). Outro aspecto negativo dos cabeçalhos de assunto é a necessidade de uma ordem na seqüência dos elementos para se realizar uma busca. Essa ordem parece não ter mais sentido considerando-se mecanismos de busca eletrônicos.

2.6.3. O Sistema Unitermo

Outra tentativa de representação foi a criação do sistema unitermo composto por um conjunto de fichas, onde cada ficha continha uma única palavra e os números dos documentos associados a esta palavra (GOMES, 1996). Com isso uma obra poderia ser representada por um conjunto de termos. No entanto, o uso de termos pode gerar o problema de sinonímia, pois cada termo pode possuir significados diferentes dependendo do contexto, em outras palavras, um mesmo termo pode ter vários sinônimos e correlações ao ser usado durante o processo de busca.

Outro problema se refere à incapacidade de um termo sozinho produzir sentido, muitas vezes é necessário agrupar mais de um termo para se obter significado. Por exemplo, ao se buscar pela palavra “dispositivo” poderiam ser relacionados outros termos que tratam de assuntos similares como “artefato”, “equipamento”, etc. Essa característica permitiria se encontrar mais documentos relevantes, entretanto, não proporcionaria uma busca refinada, pois o resultado de uma busca poderia apresentar muitos documentos, fazendo com que um pesquisador gastasse muito tempo, separando as obras mais relevantes.

Além do exposto, nos sistemas unitermo, o usuário que irá realizar a busca precisa ser especialista na área do conhecimento, isto é, ele precisa conhecer previamente os termos conceitualmente similares antes de iniciar a pesquisa. Pode-se afirmar que nesse tipo de sistema o usuário só faz perguntas cuja resposta ele já tem idéia do que seja. De qualquer forma, trata-se de uma evolução em relação aos cabeçalhos e precursor dos tesauros. Os problemas do unitermo se referem ao fato de que as idéias não podem ser representadas através de uma única palavra e um mesmo termo pode ter significados diferentes.

Com isso, diferentes usuários podem fazer uma pesquisa considerando significados diferentes, o que pode gerar problemas na recuperação.

2.6.4. Tesauros

Segundo Jesus (2002), o termo tesauro teve origem no dicionário analógico de Peter Mark Roget, intitulado "*Thesaurus of English words and phrases*", publicado, pela primeira vez, em Londres, em 1852. Em seu dicionário as palavras não foram agrupadas em ordem alfabética, como ocorre com os dicionários, mas de acordo com as idéias que elas exprimem, isto é, as palavras podiam ser encontradas pelas idéias que elas poderiam expressar.

Segundo Cavalcanti (1978), *apud* Jesus (2002), tesauro é:

> "uma lista estruturada de termos associada empregada por analistas de informação e indexadores, para descrever um documento com a desejada especificidade, em nível de entrada, e para permitir aos pesquisadores a recuperação da informação que procura"

Por essa definição pode-se observar que o objetivo do tesauro é organizar as informações, facilitando sua recuperação.

O tesauro pode ser considerado uma evolução do sistema unitermo, onde cada termo se relaciona com outros que mantém uma proximidade lógica ou semântica, isto é, os termos encontram-se associados entre si. A gênese do processo evolutivo do tesauro encontra-se descrita em Moreira (2003). Essa forma de representação também foi defendida por Bush, a mais de 60 anos atrás. Ele dizia que a mente humana trabalha melhor com associações. O tesauro facilita a busca, pois o usuário não precisa conhecer previamente os termos semanticamente semelhantes, isso já está implícito no sistema.

Gomes (1996) ainda relata: “diferentemente de outros dicionários, o seu (do tesauro) permite que se chegue a uma palavra mais adequada ou que melhor se ajuste as necessidades do escritor, sem que, de início, ele saiba quem é ela”. Além disso, o uso de associações facilita o usuário a adquirir conhecimento sozinho, pois as idéias estão organizadas e devidamente relacionadas.

Para Currás (1995), *apud* Moreira (2003):

> “Tesauro é uma linguagem especializada, normalizada, pós-coordenada, usada com fins documentários, onde os elementos lingüísticos que a compõem – termos, simples ou compostos – encontram-se relacionados entre si sintática e semanticamente.”

É especializada por atuar num certo domínio, normalizada porque seus termos são controlados, pós-coordenada pelo fato de os termos serem combinados no momento de seu uso.

O tesauro se constitui, portanto, num importante instrumento para a representação do conhecimento, uma vez que em sua estrutura encontram-se termos conceituados e associados, agilizando e facilitando o processo de recuperação da informação. Através do tesauro é possível determinar quais termos podem ser usados no sistema e quais termos podem ser usados na busca para alcançar um resultado satisfatório. Além disso, permite a introdução de novos termos em sua estrutura de modo a aproximar a linguagem do usuário à do sistema (MOREIRA, 2003).

Para que a construção do tesauro seja eficiente torna-se necessário escolher os termos que melhor representam um domínio específico. Espera-se que esse procedimento garanta a recuperação de documentos relevantes (revocação) e torne a seleção mais precisa (precisão). Para isso, é necessário fazer um controle da terminologia, fazendo uso de um vocabulário controlado, ou

seja, criar uma forma de delimitar os meios pelos quais as idéias são representadas. Isso pode garantir maior efetividade nas relações entre perguntas e respostas. Para verificar a eficácia da construção do tesauro podem ser elaboradas situações fictícias entre os usuários que utilizaram o sistema.

Em toda a bibliografia consultada, praticamente não se encontraram aspectos negativos referentes ao uso de tesauros na representação do conhecimento. O que existe é uma limitação no escopo, isto é, na abrangência do tesauro em um dado campo de conhecimento e sua constante necessidade de atualização. Os avanços da área representada pelo tesauro tornam sua atualização permanente. Para cada área do conhecimento existe a necessidade da construção de um novo tesauro. Quando a área torna-se complexa, isto é, quando a representação dos assuntos abordados começa a ser ampla, existe a possibilidade da criação de diversos micro-tesauros relacionados (CAMPOS *et al.*, 2006).

2.6.5. Metadados e Dublin Core

O aumento exagerado de material bibliográfico, em especial os disponíveis na Internet, tem dificultado a recuperação de informação. Para alcançar maior eficiência na recuperação de informação torna-se necessário conhecer onde as informações estão localizadas e de que forma elas podem ser manipuladas. Dessa forma, a representação do conhecimento torna-se um elemento fundamental para acesso rápido à informação. Uma forma de descrever informações e representar o conhecimento é realizada através de metadados.

Metadado é definido, na maior parte da literatura, como “dados sobre os dados”. Apesar de essa definição ter sido acolhida tanto na CI como na CC,

parece não existir um consenso a respeito dela, pois “dados sobre dados” não é uma definição clara. O principal objetivo dos metadados se refere a prover uma forma de facilitar a identificação, a localização e a descrição de um recurso de informação, isto é, trata-se de uma estrutura capaz de organizar os dados.

O uso de metadados permite descrever um objeto (um artigo científico, por exemplo) de maneira a melhor estabelecer suas características, auxiliando usuários e sistemas a compreenderem as fontes de informação consultadas. Talvez seja adequado definir metadados como informações sobre os dados, isto é, um conjunto de atributos que permitem identificar um recurso (um objeto) e produzir informação a respeito dele ou, como diz Abreu (2004): “metadado refere-se a alguma estrutura descritiva da informação sobre outra informação ou conhecimento, auxiliando na identificação, descrição, localização e gerenciamento desse recurso”. Seja qual for a definição, é importante observar que, apesar de existirem diversos objetivos na construção de metadados, o maior deles é descrever recursos para que esses possam ser recuperados.

Os metadados são importantes para todos os envolvidos com recursos de informação, tanto para os que o produzem quanto para os que o consomem. Através de sua utilização agrega-se valor a informação em todos os níveis de utilização: o produtor, o utilizador e o administrador dos dados são beneficiados. A adesão de formatos de intercâmbio de representação de informação possibilita que pessoas, empresas ou instituições, trabalhem em conjunto, dividindo tarefas e compartilhando dados e experiências.

Para a criação de metadados são usados termos (ou elementos) que descrevem um determinado recurso, isto é, o objeto de informação. Considerando um livro como um recurso, é possível estabelecer diversos termos para sua

descrição: titulo, autor, editora, etc. O conjunto desses termos forma o vocabulário controlado usado na descrição do metadado, onde o vocabulário controlado se refere à lista dos termos utilizados (descritores) e suas relações com outros termos.

Lourenço (2005) apresenta diversos tipos de classificação para os metadados conforme sua função em um ambiente Internet. De forma geral, os metadados podem ser assim classificados:

- **Descritivos**: usados na descrição do conteúdo de um objeto digital e elaborados a partir de uma linguagem de marcação;
- **Estruturais**: usados para estruturar a apresentação dos objetos e permitir a interação entre eles (*links*, por exemplo);
- **Administrativos**: usados para controle e preservação do recurso informacional

Após estudar as diferentes vertentes sobre a definição de metadados, Lourenço (2005) define que:

> "metadado pode ser entendido como um "identificador" que descreve, contextualiza, administra e recupera um objeto digital, além de relacioná-lo a outros objetos digitais semelhantes ou relacionados a ele dentro de uma biblioteca digital ou no ambiente da Web como um todo. É representado pelas tags das linguagens de marcação, pelos hiperlinks que ligam os objetos digitais entre si e até mesmo pelas URLs que identificam os sites da Web".

Portanto, a elaboração de um padrão de metadado ideal, a ser usado na Internet, deve contemplar as três características citadas de maneira a melhorar a organização, a descrição e a recuperação de recursos de informação. Além disso, o metadado pode ser usado para contextualizar um recurso informacional, identificando o tempo, o espaço e o ambiente político e social a partir do qual o recurso foi elaborado. Beall (2005) cita diversos aspectos essenciais para garantir a qualidade de metadados usados em bibliotecas digitais.

Existem diversos padrões para descrição de metadados. Todos eles fornecem um conjunto de termos para descrever recursos informacionais, tendo como objetivo integrar sistemas e melhorar recuperação desses recursos. Um dos mais usados é *Dublin Core*. Esse trabalho adotou o *Dublin Core* por ser o padrão escolhido pela Web Semântica, item apresentado na seção 4.3.

Trata-se de um padrão para descrição de recursos eletrônicos definido pelo comitê DCMI (*Dublin Core Metadada Initiave*), uma organização dedicada a promover a adoção de metadados interoperáveis e a descrever vocabulários especializados para a descrição de recursos. O objetivo básico do padrão é contribuir para a formação de mecanismos de busca mais inteligentes (DCMI, 2006a), oferecendo suporte à interoperabilidade entre sistemas.

Existem dois tipos de padrão *Dublin Core*: o simples e o qualificado. A diferença básica entre os dois consiste em que o qualificado possui um conjunto maior de termos, conjunto usado para refinar as informações sobre um determinado recurso, isto é, as informações a respeito do recurso tornam-se mais abrangentes. Outra característica importante do padrão qualificado se refere à possibilidade de definir aspectos semânticos a respeito dos termos usados (DCMI, 2006b).

Para descrever um recurso de forma simples, isto é, apenas com seus elementos básicos, são usados os 15 termos apresentados na Tabela 1 (DCMI, 2006c):

Tabela 1. Conjunto de termos usados no padrão Dublin Core.

Item	Termo	Descrição do termo
1	*Contributor*	Uma entidade que contribuiu com o conteúdo do recurso.
2	*Coverage*	Define a abrangência do conteúdo do recurso, tipicamente um período de datas ou uma região.
3	*Creator*	A entidade responsável pela criação do recurso.
4	*Date*	Uma data associada a publicação do recurso.
5	*Description*	A descrição do recurso, tipicamente contém o resumo.
6	*Format*	Define o tipo da mídia ou dimensões do recurso.
7	*Identifier*	Contém um identificador único para o recurso (ex. ISBN).
8	*Language*	A linguagem usada no conteúdo do recurso.
9	*Publisher*	A entidade responsável pela publicação do recurso.
10	*Relation*	Uma referência a um recurso relacionado.
11	*Rights*	A entidade que possui direitos autorais sobre o recurso.
12	*Source*	Uma referência ao local onde o recurso se localiza.
13	*Subject*	O assunto referente ao recurso, tipicamente um conjunto de palavras-chave.
14	*Title*	Título do recurso.
15	*Type*	Define a natureza ou gênero em que o recurso está inserido (imagem, som, simulação, etc.).

Esse conjunto de termos é usado para descrever um recurso qualquer como, por exemplo, um artigo científico disponível na Internet. Cada termo pode ser opcional ou ter mais de uma ocorrência, dependendo das necessidades de descrição do recurso.

Como exemplo nacional de utilização do padrão *Dublin Core*, Souza *et al.* (2000) apresentam detalhes sobre seu uso na descrição do acervo da Embrapa Informática Agropecuária. No artigo, os autores apresentam o conjunto de termos do *Dublin Core* dividido em 3 categorias: conteúdo, propriedade intelectual e instanciação. No mesmo artigo são fornecidas algumas orientações referentes a descrição do conteúdo dos elementos (termos) de metadados.

O

Quadro 1 apresenta um pequeno trecho da descrição de um recurso elaborado a partir da linguagem RDF (*Resource Description Framework*), estudada mais à frente. Os termos especificados pelo *Dublin Core* estão destacados em negrito.

Quadro 1. Exemplo usando o padrão Dublin Core. Fonte (DCMI, 2006c).

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description rdf:about="http://media.example.com/audio/guide.ra">
    <dc:creator>Rose Bush</dc:creator>
    <dc:title>A Guide to Growing Roses</dc:title>
    <dc:description>Describes process for planting and nurturing different kinds of rose
bushes.</dc:description>
    <dc:date>2001-01-20</dc:date>
  </rdf:Description>
</rdf:RDF>
```

Com os aspectos apresentados fica evidente que os metadados podem ser usados na representação de um recurso de informação, isto é, torna-se possível fazer uma representação da fonte de informação e, conseqüentemente, do conhecimento.

2.6.6. Pontos fortes e fracos de cada tipo de representação

Conforme apresentado nas seções anteriores, cada tipo de representação teve sua importância histórica e foi utilizado dentro de um contexto específico. Todos eles tiveram, e ainda têm, papel importante na representação dentro do sistema em que atuam. Todos os tipos de representação estudados possuem um objetivo em comum: facilitar e recuperar informação relevante para os usuários.

Com base nos enunciados de Gomes (1996), Tolmasquim (1998), Campos (2006), entre outros autores referenciados neste capítulo, a

Tabela 2 apresenta um resumo dos principais aspectos positivos e negativos de cada um dos tipos de representação do conhecimento. O

conhecimento dos benefícios e limitações de cada tipo de representação se faz necessário para subsidiar a elaboração de sistemas de representação mais eficazes.

Tabela 2. Formas de representação usadas na CI.

Formas de Representar	Aspectos positivos	Aspectos negativos
Sistemas de classificação	• Permitem organizar o conhecimento usando um determinado critério, criando classes de assuntos relacionados. • Possibilidade de se estabelecer vários critérios de classificação.	• A menor unidade organizada não corresponde a um conceito e sim um assunto. • Tanto a CDD, quanto a CDU possuem apenas uma forma de classificação. • Indicam um modo de classificação baseado no ponto de vista histórico, cronológico de acontecimentos, fatos etc.; o que não é propriamente história da ciência.
Cabeçalhos de assunto	• Facilidade na interpretação do sistema. •O conteúdo de uma obra pode ser representado por um conjunto de palavras-chave.	• O conjunto de termos usados nos cabeçalhos não consegue representar as diferentes áreas de assunto existentes numa base. • O grupo de palavras-chave nem sempre consegue representar o conteúdo de uma obra, trata-se de algo impreciso. Segundo Dias (2001), para a grande maioria dos livros, as bibliotecas costumam utilizar apenas um cabeçalho. • Quando se usa como critério de busca uma idéia secundária, não existente no cabeçalho, a obra não pode ser recuperada. • Existe a necessidade de uma ordem na seqüência dos elementos para se realizar uma busca.
Sistema Unitermo	• Uma obra pode ser representada por um conjunto de termos • A busca se torna bastante rápida quando o termo representa bem a informação procurada.	• Cada termo pode possuir diversos significados, gerando problemas de sinonímia durante uma busca. • O usuário que realiza a busca precisa ser especialista na área do conhecimento, pois é necessário conhecer os termos conceitualmente similares. • incapacidade de um termo sozinho produzir sentido, torna-se necessário agrupar mais de um termo.
Tesauros	• Possui uma estrutura própria, controlada, padronizada e hierarquizada, além de relações de	• Um número muito grande de conceitos é de difícil sistematização. Por esse motivo o tesauro é delimitado em um campo específico.

	associação. • Essas características o tornam mais eficiente se comparado com sistemas anteriores. • Permite que se chegue a uma palavra mais adequada ou que melhor se ajuste as necessidades do escritor, sem que, de início, ele saiba quem é ela.	• Ao se trabalhar com áreas abrangentes pode ocorrer a homonímia (o mesmo termo com significados diferentes – exemplo: tênis). Apesar disso, esse problema pode ser resolvido pelo uso de um qualificador. • Existe a necessidade de uma equipe de atualização/manutenção em função dos avanços da área representada.
Metadados	• Fornece suporte para detalhar a identificação de um recurso de informação, facilitando sua localização.	• A existência de diversos padrões pode comprometer a interoperabilidade entre sistemas • Problemas tipográficos podem comprometer a confiabilidade do conteúdo dos metadados, impossibilitando a recuperação dos recursos

Este capítulo forneceu uma base conceitual a respeito da representação da informação e do conhecimento na perspectiva da CI. O próximo capítulo descreverá os conceitos referentes à representação do conhecimento na perspectiva da CC.

3. REPRESENTAÇÃO NA PERSPECTIVA DA CIÊNCIA DA COMPUTAÇÃO

Este capítulo apresenta uma revisão bibliográfica da CC pertinente à representação do conhecimento. Descreve diversos conceitos relevantes a respeito de informação e conhecimento buscando, sempre que possível, estabelecer suas relações com a CI.

De maneira breve, pode-se afirmar que a CC é muito “experiente” na arte de representar, visto que tudo o que se apresenta na tela, ou em qualquer outro dispositivo periférico, é gerado a partir de zeros e uns. Com apenas dois valores, presença ou ausência de sinal, tudo pode ser representado, ou seja, na CC a menor unidade de representação é o bit. De maneira similar, a CI busca encontrar

a menor unidade de representação, o conceito, conforme pode ser observado em Campos (1995).

Antes de aprofundar na representação do conhecimento em si, será apresentada uma pequena revisão referente a algumas formas de representação utilizadas na CC. Para resolver os problemas cotidianos, a CC desenvolveu diversas estruturas de dados que visam representar (no computador) a realidade de um ambiente. Dessa forma, pode-se conceber que as informações a serem processadas na máquina representam uma simplificação da realidade e, para reduzir os efeitos negativos dessa simplificação, torna-se necessário escolher as estruturas mais adequadas a serem usadas na representação. A seguir, será fornecida uma visão geral de algumas dessas estruturas.

3.1. Tipos de Dados

Conforme citado anteriormente, os computadores representam todos os tipos de informação utilizando apenas zeros e uns, no entanto, esse não é um tipo de representação adequado a seres humanos, vista a complexidade envolvida na representação de um dado qualquer. Por exemplo, o número decimal 215 é representado pela seqüência binária 11010111. Com o objetivo de reduzir essa complexidade, foram criados alguns tipos de dados. Um tipo de dado pode ser concebido como uma estrutura adequada para armazenar dados de um determinado tipo. Os tipos de dados elementares, conhecidos também como tipos primitivos, são:

- **Inteiro**: um tipo de dado que permite armazenar valores inteiros, incluindo valores positivos e negativos.
- **Real**: um tipo de dado que permite armazenar valores com ponto

flutuante, isto é, com casas decimais.

- **Lógico**: um tipo de dado que pode assumir apenas os valores verdadeiro ou falso.
- **Caractere**: um tipo de dado que possibilita armazenar qualquer caractere (letra, algarismo, sinal etc.).

O tipo de dado define o conjunto de valores que podem ser armazenados em uma variável. De modo simplista, uma variável é um apelido atribuído a um endereço de memória onde o dado será armazenado. Em outras palavras, o tipo de dado estabelece o conjunto de valores que podem ser armazenados em uma dada posição de memória do computador. Por exemplo, com o tipo Inteiro é possível armazenar o número 100, porém, o número 100.42 não (ele não é inteiro). Já uma variável do tipo Real é adequada para armazenar tanto o primeiro quanto o segundo número.

Outras estruturas mais complexas podem ser criadas a partir dos tipos primitivos. Exemplos dessas estruturas são: vetores, matrizes, listas e árvores. As seções seguintes abordam cada uma dessas estruturas.

3.1.1. Vetores e Matrizes

O vetor forma uma estrutura que pode armazenar um grupo de variáveis do mesmo tipo. Numa linguagem mais técnica, o vetor forma um conjunto finito e ordenado de elementos homogêneos. É finito porque existe um número específico de elementos em um vetor; homogêneo porque todos os elementos de um vetor devem ser do mesmo tipo, isto é, todos devem possuir o mesmo tipo de dado.

Todo vetor possui um nome para identificar seus elementos. Para acessar um elemento de um vetor é usado um índice. Por exemplo: um vetor chamado “titulo” pode ser usado para armazenar títulos de obras: titulo[1]=“Representação

do conhecimento", titulo[2]="Informação como coisa", titulo[3]="Uma informação tácita". Todos os elementos de um vetor possuem o mesmo nome (título), no entanto os elementos são diferenciados pelo uso do índice (1, 2, 3). Os elementos de um vetor podem ser manipulados de acordo com as necessidades de informação. Mantendo elementos em um vetor é possível realizar diversos procedimentos sobre eles, tais como a classificação ou a pesquisa.

A matriz forma uma estrutura similar ao vetor, porém multidimensional. Para acessar um elemento da matriz torna-se necessária à utilização de dois ou mais índices, associados ao nome da matriz. Por exemplo: titulo[1,3]="Representação do conhecimento".

Tanto o vetor quanto a matriz são muito usados na CC em inúmeras aplicações. Essas estruturas são muito eficientes quando se necessita representar os mais variados conjuntos de dados dispostos em linhas e colunas, tais como aplicações realizadas através de planilhas eletrônicas.

3.1.2. Listas

Na CC, uma lista é uma estrutura que permite representar um agrupamento de elementos do mesmo tipo, cujo elemento básico é o vetor. Por exemplo, uma relação de seleções da copa do mundo, uma lista de amigos, etc. Cada seleção, ou cada amigo, pode compor um item da lista. Utilizando estruturas desse tipo é possível representar o mundo real no computador, criando um ambiente mais adequado para a manipulação dos elementos que compõem a lista. Os elementos de uma lista podem, por exemplo, ser organizados, consultados, etc.

Tecnicamente, uma lista é formada por nós (ou nodos) que podem conter, em cada um deles, um dado do tipo primitivo (inteiro, real, caractere, etc) ou uma

outra estrutura composta por vários tipos primitivos em conjunto (registro).

As listas podem ser lineares ou não-lineares. Segundo Villas *et al.* (1986), lista linear é “uma estrutura dinâmica caracterizada por uma seqüência ordenada de elementos, no sentido da sua posição relativa”. A lista linear é dinâmica porque seus elementos podem ser reorganizados durante a execução de um programa; é uma seqüência ordenada de elementos pelo fato de seus elementos serem dispostos em ordem seqüencial, um após o outro. Existem dois tipos básicos de estruturas do tipo lista: pilha e fila. Essas estruturas são descritas nas seções seguintes.

3.1.2.1. Pilha

Segundo Villas *et al.* (1993), pilha é “uma lista linear em que as operações de inserção e retirada são efetuadas apenas no final da lista”. Uma operação de inserção é um procedimento que insere um elemento na pilha, enquanto que a retirada faz a operação inversa. Essas operações são realizadas no topo da pilha (final da lista linear). A operação de inserção recebe o nome de *push* e de retirada *pop*, como pode ser observado pela Figura 1. Inicialmente, a pilha encontra-se vazia, ao utilizar a instrução *push* (a), o valor de “a” é inserido no topo da pilha. O mesmo ocorre com os valores de “b”, “c” e “d”. Ao utilizar a instrução *pop* () os valores referentes à “c” e “d” são retirados do topo da pilha.

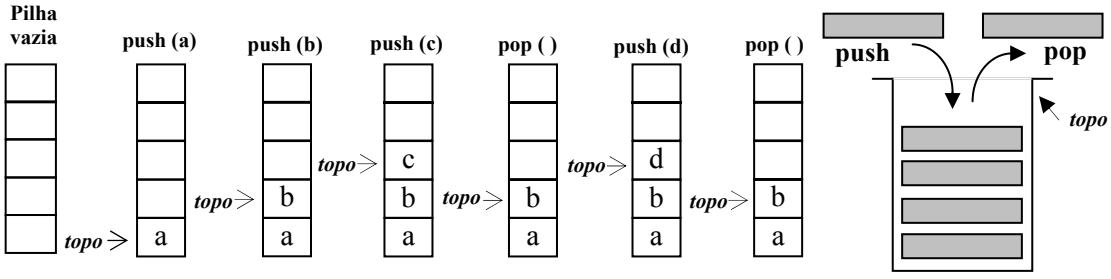

Figura 1. Representação das operações em uma pilha (baseado em Pereira, 1996).

Pelo fato de os elementos serem inseridos e retirados apenas no topo da pilha, o primeiro elemento a ser inserido será o último a ser retirado, enquanto que o último a ser inserido será o primeiro a ser retirado. Essa ordem é conhecida como LIFO – *Last In First Out*. A estrutura criada pela pilha funciona exatamente como no mundo real. Ao se empilhar livros, por exemplo, o primeiro livro a ser empilhado será o último a ser retirado.

A CC utiliza a estrutura de pilhas em muitas aplicações. O botão “Voltar” (*back*), disponível num *software browser*, pode servir de analogia para a compreensão do funcionamento das pilhas. O usuário navega por diversos sites e, através do botão “Voltar”, tem a opção de realizar o caminho inverso, isto é, o último site visitado será o que aparecerá primeiro, e assim sucessivamente até se chegar ao site inicial.

3.1.2.2. Fila

Segundo Villas *et al.* (1993), fila é uma lista linear em que “as operações de inserção são efetuadas apenas no final e as de retirada, apenas no início da lista”. Essas operações estão representadas na Figura 2. A imagem a esquerda mostra uma fila com três elementos (a,b,c) em que “a” é o primeiro elemento da fila e “c” o último. Ao retirar o elemento “a” (imagem ao centro), o primeiro elemento passa a ser o “b”. A imagem a direita mostra o elemento “d” inserido no final da fila. Para melhor compreensão da Figura 2, considere que os elementos são inseridos da direita para a esquerda.

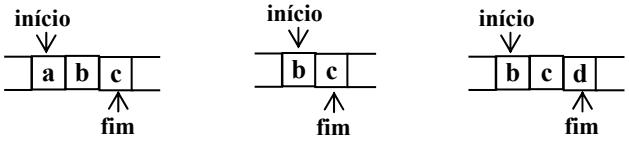

Figura 2. Representação das operações em uma fila.

Pelo fato de os elementos serem inseridos no final da fila e retirados no início dela, o primeiro elemento a ser inserido será o primeiro a ser retirado, enquanto que o último a ser inserido será o último a ser retirado. Essa ordem é conhecida como FIFO – *First In First Out*. A estrutura criada pela fila funciona exatamente como no mundo real. Ao se formar uma fila de pessoas, a primeira a compor a fila será a primeira a ser atendida (retirada da fila), assim como a última que compor a fila será a última a ser atendida.

Assim como as pilhas, existem diversas aplicações para as filas. Elas podem ser usadas, por exemplo, na simulação e análise de filas reais. As filas de atendimento em bibliotecas podem ser analisadas através da simulação de seu comportamento. Outro exemplo prático se refere à lista de *links*, apresentada por um mecanismo de busca como Google. O resultado de uma consulta pode ser colocado em uma lista por ordem de relevância, isto é, os primeiros a serem inseridos na fila serão os primeiros a serem apresentados na relação fornecida ao usuário.

Apesar de a lista ter uma estrutura linear, existem variações em seu "comportamento", dependendo de suas necessidades de utilização. Além de pilhas e filas, Pereira (1996) descreve as listas duplas e as listas circulares.

3.1.3. Árvores

Segundo Villas *et al.* (1993), árvore é "uma estrutura não-linear que representa relações de hierarquia e composição". Relações de hierarquia porque os elementos não são dispostos em forma linear, como em pilhas e filas, mas em forma de uma árvore hierárquica. Relações de composição porque essa estrutura permite representar a composição de um elemento, isto é, permite definir que um

elemento é composto por outros elementos.

A árvore é formada por um elemento principal denominado raiz que possui ligações com outros elementos, chamado galhos (ou filhos). Esses elementos filhos podem conter outros elementos filhos, formando uma estrutura hierárquica, conforme pode ser observado na Figura 3.

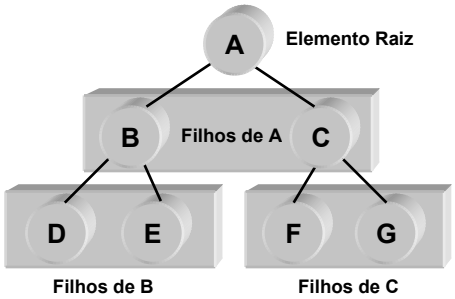

Figura 3. Representação gráfica de uma árvore.

No campo da CC as árvores podem ser usadas em inúmeras aplicações, tais como: em redes de comunicação de dados no envio e recebimento de pacotes entre computadores e, nos sistemas de geoprocessamento, para facilitar a navegação em aplicações visuais. Já o campo da CI utiliza estruturas em forma de árvore nos sistemas de classificação (CDD e CDU) e também em tesauros e na criação de ontologias.

Garcia (2000) apresenta uma forma de extrair conhecimento em bases de dados através de árvores de decisão. O objetivo é extrair informações a partir de dados armazenados em uma base. Para isso, constrói classificadores que separam os dados em pequenos grupos, levando-se em consideração suas características semelhantes. Dessa forma, uma estrutura em forma de árvore pode ser usada para representar o conhecimento e apoiar a decisão.

3.2. Conceitos de Representação do Conhecimento

Existem diversas definições a respeito da representação do conhecimento na Ciência da Computação. Davis *et al.* (1992), *apud* Campos (2004), cita diversas definições ligadas ao papel da representação do conhecimento, ou seja, procura descrever quais os motivos de se buscar a representação.

> 1. Uma representação de conhecimento é um mecanismo usado para se raciocinar sobre o mundo, em vez de agir diretamente sobre ele. Neste sentido, ela é, fundamentalmente, um substituto para aquilo que representa.

Conforme essa afirmação, a CC está em acordo com a CI, pois também procura criar um substituto para aquilo que representa. Da mesma forma, a CC também considera que toda representação é imprecisa e contém simplificações do fato real.

> 2. Uma vez que toda representação é uma aproximação imperfeita da realidade, ao selecionarmos uma representação, estamos tomando um conjunto de decisões sobre como e o que ver no mundo. Portanto, selecionar uma representação significa fazer um conjunto de compromissos ontológicos.

Esse preceito também está em acordo com o descrito em Alvarenga (2003) e Almeida (2005), isto é, para representar o conhecimento é preciso criar um conjunto de termos e conceitos que estejam em harmonia na consciência de um grupo.

> 3. Uma representação de conhecimento é uma teoria fragmentada de raciocínio que especifica que inferências são válidas e quais são recomendadas pelo exame de três componentes: a concepção de inferência inteligente, o conjunto de inferências que a representação sanciona e o conjunto de inferências que ela recomenda.

Pelo fato de a CC utilizar mecanismos de automatização de processos, torna-se importante definir o que significa representar e quais são as etapas inerentes a esse processo. Para isso, a CC tem buscado responder a três questões fundamentais sobre o ato de representar: “(i) O que significa raciocinar de forma inteligente? (ii) O que é possível inferir a partir do que se conhece? (iii)

O que se deve inferir a partir do que se conhece?". A resposta a essas questões tornaria possível às máquinas realizarem inferências sobre uma representação do conhecimento. Procedimento semelhante ocorre na CI, já citado em Alvarenga (2003). A autora procura determinar as etapas seguidas por um ser humano no ato de representar.

Essas perguntas são difíceis de responder, pois se seres humanos fazem representações diferentes de acordo com suas experiências anteriores, quanto mais uma máquina. Para auxiliar esse processo pode ser realizada a modelagem do conhecimento, uma forma de representar o que se conhece através de modelos.

3.3. Modelagem do Conhecimento

Le Moigne, citado em Campos (2004), afirma que conhecer é modelizar (abstrair), ou seja, o processo de conhecer equivale à construção de modelos do mundo/domínio a ser construído que permitem descrever e fornecer explicações sobre os fenômenos observados. Sua preocupação se refere à metodologia usada para se modelar o conhecimento, apesar de acreditar que a razão humana está acima de toda metodologia, não sendo necessário seguir sempre o mesmo algoritmo de modelização.

Campos (2004) estabelece quatro princípios fundamentais e seqüenciais, representados na Figura 4, que podem ser utilizados no ato de modelar domínios de conhecimento.

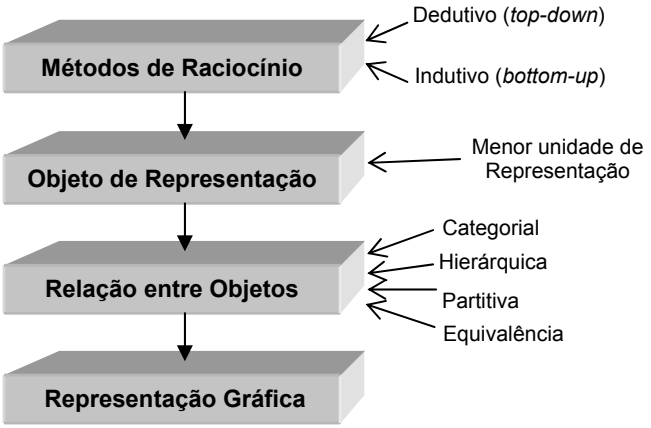

Figura 4. Princípios da modelagem do conhecimento.

Os quatro princípios de modelar o conhecimento, expostos na Figura 4, podem ser assim concebidos:

- **Método de Raciocínio**: princípio de modelar o conhecimento utilizando o método dedutivo (*top-down*) ou o método indutivo (*bottom-up*). No dedutivo, parte-se do geral para o específico, isto é, o processo de modelagem se inicia pelo domínio/contexto e os elementos que irão compor a representação. As relações entre os elementos são consideradas em uma etapa posterior. Já no método indutivo ocorre o inverso. A Tabela 3 apresenta um resumo dos métodos de modelagem usado na representação, considerando-se as áreas da CI, CC e Terminologia, baseado em Campos (2004).

Tabela 3. Métodos de modelagem e formas de representação.

Área	Representação	Método	Característica
CC	Ontologia Formal	Indutivo	de objetos para o estabelecimento do contexto
	Orientação a objetos	Indutivo	de objetos e seus relacionamentos para sua representação
Termi-nologia	Sistema de conceitos	Indutivo	do conceito para um sistema de conceitos
CI	Classificação Facetada de Ranganathan	Dedutivo	De categorias para conceitos
	Teoria do Conceito de Dahlberg	Dedutivo e Indutivo	considerada híbrida, pois agrega dedutivo e indutivo. Apesar de existirem categorias genéricas, chega-se a elas a partir de um processo indutivo de análise do conceito.

- **Objeto da Representação**: princípio de modelar o conhecimento definindo o objeto de representação, isto é, estabelecendo a menor unidade de representação em um dado contexto. Enquanto que na CI a menor unidade de representação é o conceito, exposto por Ranganathan e Dalhberg, para a CC a menor unidade de representação, tanto na OO quanto na Ontologia, é o próprio objeto, ou seja, um objeto é um conceito. Não existe uma diferenciação clara entre o objeto do mundo real e sua representação. O que se busca é simular nos sistemas computacionais o que ocorre no mundo real, sem se preocupar com o entendimento do conteúdo do objeto representado.
- **Relações entre os Objetos**: princípio de modelar o conhecimento objetivando verificar as possibilidades de ligação/separação semânticas entre os conceitos de um dado domínio. As relações entre os conceitos podem ser realizadas através da categorização (agrupamento dos conceitos pela sua natureza), pela relação hierárquica (pela ordem de dependência entre os conceitos), pela relação partitiva (partes que compõem um determinado conceito) e pelas relações de equivalência.
- **Representação Gráfica**: princípio de modelar o conhecimento elaborando diagramas que expressem as relações conceituais. Apesar de a CI estar bem fundamentada em conceitos e relações conceituais, ela carece de representações gráficas. A criação de representações gráficas poderia contribuir para melhorar a comunicação entre o mundo real e sua representação. Na CC, existem diversos modelos de representação gráfica cujo objetivo é aumentar as chances de sucesso na elaboração de sistemas.

Conforme o exposto, a modelagem é um processo importante no ato de representar o conhecimento, uma vez que busca determinar certas “regras” a serem seguidas, buscando criar um ambiente mais adequado à representação. O que se busca com esses princípios é simular o raciocínio humano no ato de representar. Se esse objetivo for atingido, será possível desenvolver sistemas que interpretem uma base de dados e representem o conhecimento armazenado de

forma automática. Para elaborar sistemas desse tipo é necessário utilizar linguagens adequadas para a manipulação do conhecimento.

3.4. Técnicas para representação do conhecimento

A representação do conhecimento tem sido muito estudada na CC, principalmente na subárea de Inteligência Artificial. A aplicabilidade mais comum da representação do conhecimento é servir de base para a construção de sistemas especialistas, em especial os voltados à área médica (Widman, 1998).

A CC representa o conhecimento de forma declarativa, a partir de um conhecimento adquirido anteriormente. Para isso, define formalismos computacionais, baseados em diferentes técnicas, numa linguagem que o computador possa entender e processar, ou seja, o conhecimento precisa ser transformado em códigos.

A escolha da técnica de representação de conhecimento está relacionada ao grau de complexidade do conhecimento que deve ser explicitado, bem como do tipo de aplicação onde será usado. As técnicas de representação do conhecimento mais usadas na CC são: regras de produção, redes semânticas, *frames*, triplas objeto-atributo-valor e lógica de predicados. Em função dos objetivos propostos, serão descritas apenas as três primeiras formas.

3.4.1. Regras de Produção

É uma das técnicas de representação do conhecimento mais simples, dada sua facilidade de compreensão e programação, uma vez que a linguagem usada na determinação das regras é muito próxima à linguagem natural. Cada regra de produção é composta por três partes: o nome da regra, a parte SE (ou IF) e a

parte ENTÃO (ou THEN), onde:

- O nome da regra é um identificador a partir do qual o mecanismo de inferências pode usá-lo (exemplo regra1, regra2, etc.);
- a parte SE contém a premissa para a execução de uma regra. Caso seja verdadeira, as instruções constantes no ENTÃO serão executadas;
- a parte ENTÃO contém as ações a serem executadas caso a condição da parte SE seja verdadeira.

O mecanismo de inferência é responsável por analisar as premissas das regras e executar as ações. Para exemplificar a representação do conhecimento através de regras de produção, foi elaborado um exemplo baseado em Melo (2001).

Considere um mecanismo de busca da Internet procurando por artigos eletrônicos que sejam da área de CI e que tratem de representação do conhecimento. Para cada texto pesquisado, são aplicadas as regras constantes no Quadro 2 na tentativa de classificá-lo como um artigo relevante. A condição para que o texto seja um artigo relevante é: *SE pesoFinal for maior ou igual a 0.8 ENTÃO é um artigo relevante*.

Quadro 2. Regras de produção para representação do conhecimento.

Regra 1:	SE existe o termo “representação do conhecimento” ENTÃO incrementa a chance de ser um artigo válido (peso 0.5)
Regra 2:	SE existe o termo “ciência da informação” ENTÃO incrementa a chance de ser um artigo válido (peso 0.3)
Regra 3:	SE existem os termos “resumo” e “conclusão” ou “considerações finais” ENTÃO incrementa a chance de ser um artigo válido (peso 0.3)
Regra 4:	SE existe o termo “classificação” ou “teoria do conceito” ou “semântica” ou “significado” ou “domínio” ENTÃO incrementa a chance de ser um artigo válido (peso 0.3)
Regra 5:	SE existe o termo “tesauro” ou “ontologia” ENTÃO incrementa a chance de ser um artigo válido (peso 0.3)

Os termos das regras foram definidos considerando-se os termos mais freqüentes que aparecem nos textos de representação do conhecimento na área de CI. Os pesos das regras foram ajustados de acordo com testes realizados. O valor numérico que mede a chance de o texto ser relevante é obtido através a equação: ***PesoFinal = PesoFinal + PesoNovo * (1 – PesoFinal)***. PesoFinal é um índice que estabelece a chance de o texto ser um artigo relevante, inicialmente igual a zero. PesoNovo é o peso da regra atual.

Para que um texto a ser buscado na Internet seja considerado relevante, as regras 1 a 3 devem ser, obrigatoriamente, verdadeiras. Além disso, pelo menos a regra 4 ou a 5 deve ser verdadeira. Caso contrário, o texto não será considerado relevante.

Para Ávila (1991), a possibilidade de agregar novas regras, independente do conhecimento já existente no sistema, permite um certo grau de modularidade as regras de produção, no entanto, a adição incontrolada de regras pode dar lugar a inferências incorretas e resultados imprevistos. Apesar dos benefícios apontados, as regras de produção possuem deficiência quanto à definição de termos e na descrição de objetos de domínio e suas relações. Essas deficiências se tornarão mais claras na leitura das seções seguintes em que será discutida a representação do conhecimento através de redes semânticas e *frames*.

3.4.2. Redes Semânticas

As redes semânticas têm seu foco nas categorias de objetos e nos relacionamentos entre eles, assemelhando-se ao modelo orientado a objetos a ser descrito na seção 3.5.9. O conhecimento é representado através de um conjunto de nós (*nodos*) e um conjunto de arcos (*links*). Os nós representam

conceitos através de substantivos, adjetivos, pronomes ou nomes próprios. O relacionamento entre estes conceitos é representado por arcos através de verbos transitivos ou preposições. Os termos mais comuns usados em arcos são IS_A (é um), HAS_A (tem um), A_KIND_OF ou AKO (é um tipo de) e IS_PART (é parte de).

A rede semântica é representada através de uma notação gráfica, conforme exemplificado no modelo da Figura 5.

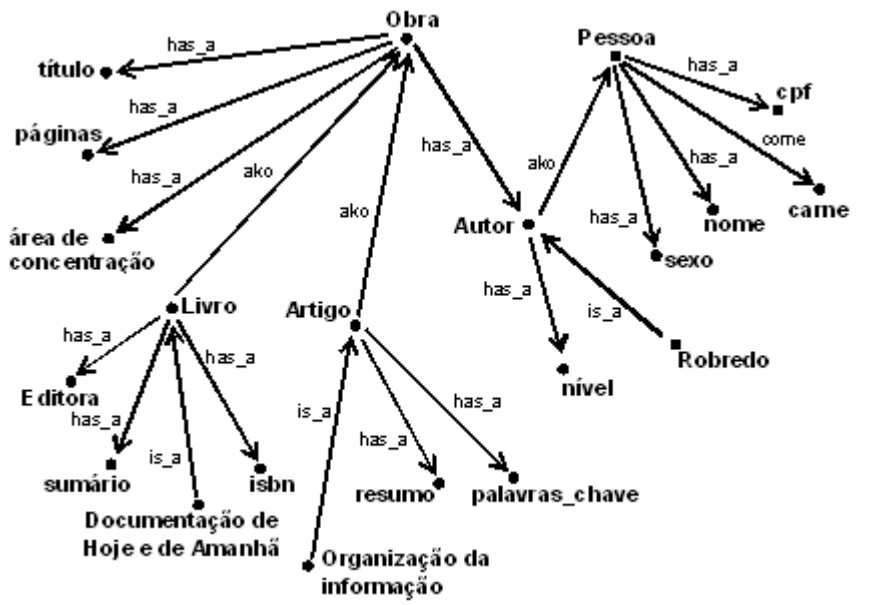

Figura 5. Modelo de uma rede semântica.

Os nomes dos nós representam conceitos e os arcos a ligação entre esses os conceitos. A seta do arco determina o sentido da leitura e, apesar de não estar explícito no modelo, podem ser lidas nos dois sentidos. Algumas inferências podem ser realizadas no modelo: Livro é um tipo de Obra; Artigo é um tipo de Obra; Livro tem um ISBN, um sumário e uma Editora; Robredo é um Autor, um tipo de Pessoa que possui CPF, um nome etc.

Um mecanismo de inferência se encarrega de interpretar o diagrama para

inferir os resultados esperados. A interpretação de uma rede semântica usando a notação gráfica é relativamente simples, no entanto, para se evitar conclusões falsas, pode ser necessário aumentar o número de relações semânticas (arcos), tornando o sistema mais complexo.

Segundo Shastri (1988), *apud* Teive (1997), os nós dos níveis hierárquicos mais baixos denotam indivíduos ou instâncias e são conectados por arcos do tipo IS_A (em nosso exemplo, Robredo e Organização da Informação), enquanto que os nós de níveis hierárquicos mais altos representam classes ou categorias de indivíduos (em nosso exemplo, Obra, Pessoa, etc).

Para Ávila (1991), apesar de as redes semânticas serem usadas para representar o conhecimento, elas possuem algumas limitações. Existe uma baixa precisão de interpretação proporcionada por nós e arcos. Em nossa análise, isso se deve a baixa complexidade empregada no modelo. Isso pode ser observado na Figura 5. Um mecanismo pode inferir corretamente que Robredo é uma Pessoa que possui um CPF, no entanto, pode inferir incorretamente que Robredo come carne. Se no lugar de carne fosse colocado alimento, a inferência seria correta, pois toda pessoa, no caso Robredo, precisa de alimento, mas de carne nem sempre.

Outro problema das redes semânticas se refere ao significado dos termos usados em nós e arcos. Por exemplo, o termo Obra no modelo exposto se refere a documentos de texto, entretanto, o mesmo termo poderia ser usado para representar obras de arte ou construções da área civil. Dessa forma, a interpretação de nós e arcos pode ser ambígua, pondo em risco a confiabilidade nas inferências realizadas.

3.4.3. Frames

No campo da Inteligência Artificial, o termo *frame* (quadro) refere-se a uma maneira especial de representar conceitos e relações, organizado de forma muito similar às redes semânticas. O modelo de *frames* para a representação do conhecimento foi proposto em 1975, por Marvin Minsky. Da mesma forma que nas redes semânticas, na representação do conhecimento através de *frames,* os nós do topo representam conceitos gerais, enquanto que os nós inferiores representam instâncias mais específicas de conceitos. Outra semelhança dos *frames* com as redes semânticas se refere ao fato de também serem definidos como objetos estruturados (BONET, 1985, *apud* TEIVE, 1997).

Assim como nas redes semânticas, uma das características dos *frames* é a possibilidade de se definir novos subtipos de objetos que herdam todas as propriedades da classe original. A diferença básica das redes semânticas em relação aos *frames* é que, enquanto o primeiro organiza o conhecimento na forma de redes, o segundo organiza de forma hierárquica. Apesar disso, uma mesma representação do conhecimento pode ser realizada em ambos (MARTINS, 2005).

A estrutura do *frame* é composta por duas partes básicas: o nome do *frame* e um conjunto de “gavetas” (escaninhos ou *slots*). A Figura 6 apresenta a representação gráfica de um *frame*. O nome é o identificador do *frame* que se deseja definir, no caso, “Artigo”. Cada *slot* possui um nome e é formado por uma ou várias propriedades (facetas) com conteúdos individuais.

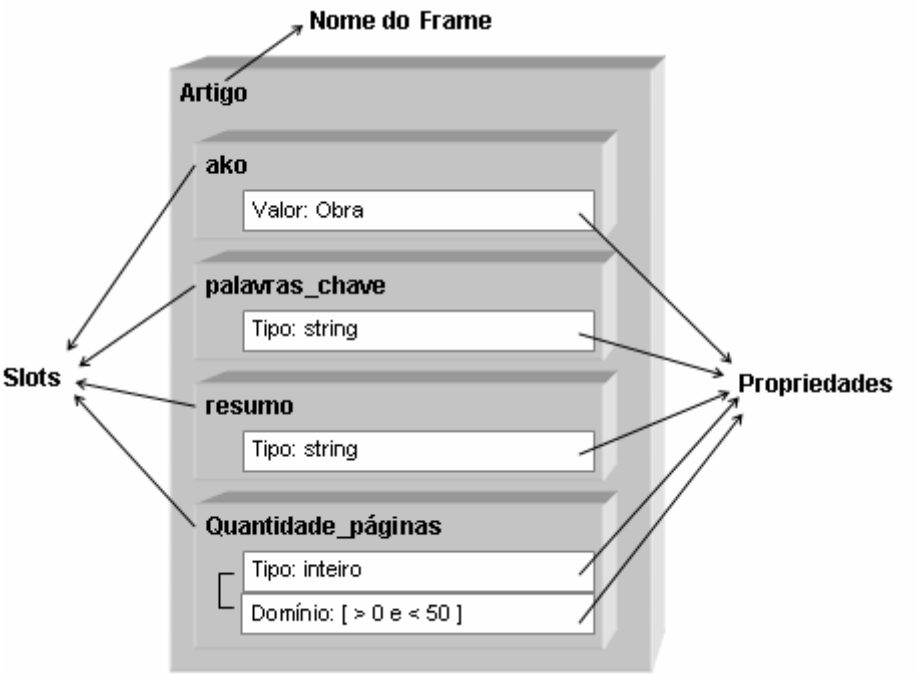

Figura 6. Estrutura básica de um frame.

As facetas descrevem os *slots*, definindo explicitamente o tipo de informação que pode ser armazenada, impondo restrições e domínios. Por isso, as facetas têm um importante papel sobre os *slots*, garantindo a taxonomia dos valores armazenados no *frame*.

De forma comparativa, a Figura 7 traz a mesma representação do conhecimento da Figura 5 (elaborada a partir de redes semânticas), porém realizada a partir de *frames*.

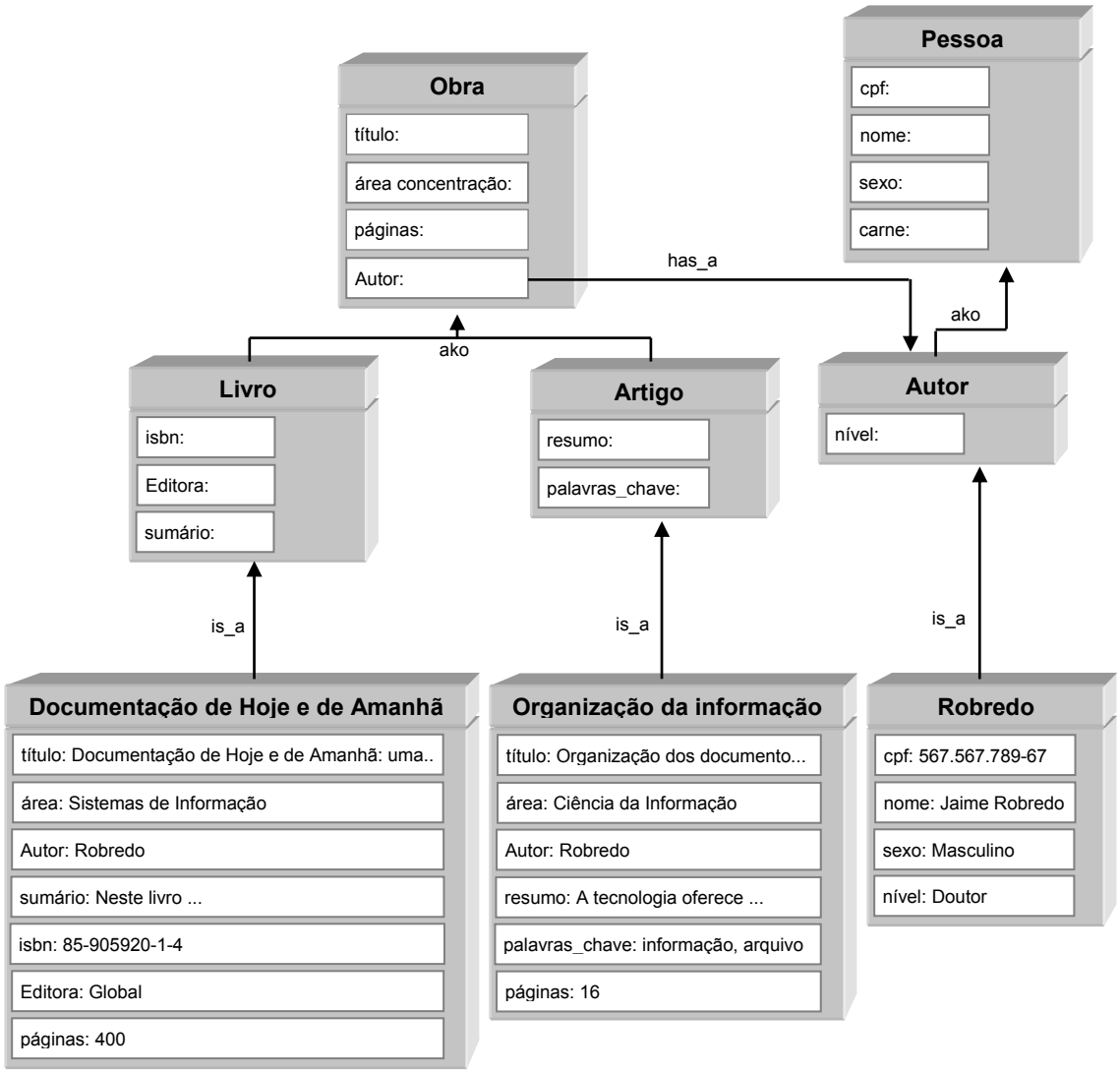

Figura 7. Representação do conhecimento usando *frame*.

A partir de um modelo desse tipo, um mecanismo de inferência pode ser usado para responder perguntas do tipo: “Robredo é uma Pessoa?”, “Organização da Informação é uma Obra?”.

O modelo de *frames* possui uma representação gráfica que mais se aproxima da maneira como os especialistas da CC pensam quando desenvolvem sistemas, isto é, o modelo consegue abstrair o que se conhece a respeito de uma dada realidade. No entanto, as linguagens usadas na manipulação de *frames* não dão suporte à descrição do conhecimento na forma declarativa, como

apresentado nas Regras de Produção.

3.4.4. Pontos fortes e fracos de cada tipo de representação

Em função das pesquisas realizadas, considerando-se principalmente os enunciados de Ávila (1991), Teive (1997) e Melo (2001), não existe uma técnica de representação ideal para qualquer caso. Quando possível, deve-se utilizar várias técnicas de representação do conhecimento de maneira conjunta, aproveitando-se das vantagens de cada uma. Quando as técnicas são usadas em conjunto, aumentam as chances de se obter uma representação mais próxima à linguagem natural. A Tabela 4 procura apresentar os aspectos positivos e negativos de cada forma de representação discutida.

Tabela 4. Formas de representação usadas na CC.

Técnica de representação	Aspectos Positivos	Aspectos Negativos
Regras de produção	• A linguagem usada na determinação das regras é declarativa e muito próxima da linguagem natural. • Possibilidade de agregar novas regras, independente do conhecimento já existente no sistema.	• Deficiência quanto à definição de termos e na descrição de objetos de domínio e suas relações. • O aumento do número de regras pode causar resultados imprevistos.
Redes Semânticas	• Permite representar associações similares ao pensamento humano. • Possibilita a descrição de objetos de domínio e suas relações. • Permite a definição de novos subtipos de objetos que herdam todas as propriedades da classe original.	• Baixa precisão na definição e interpretação de nós e arcos, podendo gerar ambigüidade. • A simplificação do modelo pode gerar inferências incorretas. • Sem suporte a descrição do conhecimento na forma declarativa, apenas procedimental.
frames	• Idem Redes Semânticas • Representação gráfica mais completa. • Uso das facetas contribui para o controle da taxonomia.	• Idem Redes Semânticas. • Suporte apenas a estrutura hierárquica, podendo gerar restrições de representação.

3.5. Estruturas de Representação

Esta seção apresenta uma visão geral sobre algumas linguagens de marcação que possuem ligação direta com a CC e CI. São apresentadas as características principais dessas linguagens, seus pontos fortes e suas limitações. Numa abordagem comparativa, visa determinar o papel de cada linguagem na representação da informação.

3.5.1. Diretórios

Pode-se considerar que a primeira estrutura de representação e organização usada pela CC foram os diretórios. Um diretório corresponde a um local lógico, definido em uma unidade, para armazenar arquivos comuns. Com o advento do sistema operacional Windows o diretório passou a ser conhecido pelo nome de “pasta”, um nome mais sugestivo para usuários comuns. Para organizar as referências dessa dissertação, por exemplo, foram criadas diversas pastas com as letras do alfabeto. Cada autor foi armazenado no diretório de letra correspondente a seu sobrenome. Por exemplo, as referências de Robredo foram armazenadas em um diretório de nome “R”, as referências de Saracevic em um diretório de nome “S”, e assim por diante. Para facilitar ainda mais a localização das referências, todo arquivo foi nomeado seguindo a seguinte nomenclatura: “sobrenome do autor (ano da publicação) - título do artigo”. Por exemplo, uma referência usada de Robredo possui a seguinte nomenclatura: “Robredo (2004) - Organização dos documentos ou organização da informação, uma questão de escolha.pdf”. Essa metodologia garante não apenas uma melhor organização, mas também facilita a recuperação das referências.

A função básica do diretório é organizar os conteúdos armazenados para

facilitar a recuperação. Segundo Cendón (2001) “os diretórios foram a primeira solução proposta para organizar e localizar os recursos da Internet, tendo precedido os motores de busca por palavras-chave” e um dos pioneiros a organizar seus conteúdos de forma hierarquizada em diretórios foi o Yahoo!. Esse tipo de organização era muito usado no início da Internet, época em que a quantidade de documentos armazenados era pequena, comparada com os dias atuais.

As informações disponíveis no site do Yahoo permanecem ainda dispostas em categorias e subcategorias. Para buscar uma informação o usuário dispõe de uma lista, normalmente em ordem alfabética, com *links* divididos nessas categorias. Nesse caso a busca é realizada do item mais geral para o mais específico. Nos diretórios, não existe um ordenamento de sites por relevância da informação desejada: todos os sites de uma dada categoria são exibidos ao usuário (normalmente em ordem alfabética).

Os diretórios têm bases de dados pequenas se comparados a outras formas de indexação como os motores de busca, no entanto, a probabilidade de possuírem conteúdos relevantes é maior. Para Cendón (2001) os diretórios são mais apropriados para buscas por tópicos que sejam de interesse para um grande número de pessoas. Com relação à organização, a maior parte dos diretórios usa listas hierárquicas de assunto, apesar de existirem alguns que utilizam os cabeçalhos de assunto e a classificação de Dewey.

Ainda em termos de Internet, os documentos armazenados nos diretórios (e na enorme maioria das páginas) são elaborados a partir da linguagem HTML (*HyperText Markup Language*) um tipo de linguagem de marcação que determina a formatação com que o conteúdo do documento será apresentado na tela. Um

dos principais problemas do uso dessa linguagem se refere à recuperação da informação contida nos documentos. Isso acontece porque a informação não está estruturada (no documento), ou seja, essa informação está organizada apenas para seres humanos entenderem e não processos automáticos.

3.5.2. Linguagens de Marcação

Almeida (2002) afirma que:

> "A ciência de informação, como campo dedicado às pesquisas científicas voltadas para os problemas da efetiva transmissão do conhecimento, seus registros e recuperação, encontra boas oportunidades para estudo e discussão das linguagens de marcação".

Essa seção busca apresentar a importância da linguagem de marcação (LM) para a CI, mesmo sendo oriunda da CC. São fornecidos os fundamentos essenciais das linguagens de marcação como um pré-requisito para a compreensão da linguagem XML (*eXtensible Markup Language*), tema abordado na seção seguinte.

Diversos autores apresentam várias definições para LM:

> "Uma linguagem de marcação possui códigos para indicar o leiaute e estilo a ser apresentado em um texto" (FKCC, 2004).
>
> "...um conjunto de convenções utilizadas para a codificação de textos" (ALMEIDA, 2002).
>
> "Uma sintaxe e procedimento para embutir em documentos tipo texto, marcadores que controlam formatação, quando são visualizados por aplicações especiais como um browser" (MORGAN, 2004).
>
> "Uma forma de descrever a estrutura lógica ou semântica de um documento e fornecer instruções a computadores sobre como apresentar o conteúdo de um arquivo" (DAVIES, 2004).

As três primeiras definições são apenas parciais e descrevem as LM em seu estado inicial. Já a definição de Davies, apresenta uma visão mais global e

que interessa diretamente aos estudos da CI: a possibilidade de descrever o conteúdo semântico de um texto.

Existem diversos tipos de LM, cada uma contém usos e finalidades específicas. Editores de texto, por exemplo, podem manter marcas internas para controlar diversos atributos de um texto, tais como cor, tamanho, formatação, etc. Aplicativos para navegação na Internet apresentam na tela documentos criados a partir de uma LM, documentos compostos por inúmeras marcações. Um telefone celular também pode utilizar uma LM específica para navegar pela Internet. Enfim, existem muitas aplicações para as LM.

Como o próprio nome sugere, as LM têm como característica principal criar marcas (*tags*) para delimitar um texto. Num primeiro momento, essas marcas eram usadas apenas para definir a forma como um texto seria apresentado. Mais tarde, com a evolução das linguagens, tornou-se possível usar marcas para fornecer um certo significado ao texto.

Uma marca é um tipo de código que envolve uma palavra ou texto, conforme apresenta o Quadro 3.

Quadro 3. Texto com marcadores apontados em negrito.

As **<B>** Linguagens de Marcação **</B>** permitem a criação de muitas **<U>** aplicações **</U>** no campo da **<I>** Ciência da Informação **</I>**.

Neste exemplo, baseado na linguagem HTML, o trecho delimitado pelas marcas **<B>** e **</B>** aparecerá em negrito (*bold*), já o trecho delimitado pelas marcas **<U>** e **</U>** aparecerá sublinhado (*underline*), enquanto que o trecho delimitado pelas marcas **<I>** e **</I>** será apresentado em itálico (*italic*). O restante do texto, que não está envolvido por nenhuma marca, será apresentado em sua forma normal, isto é, sem nenhuma formatação.

Todas as LM possuem um conjunto de convenções utilizadas na definição de suas marcas, isto é, contêm diversas marcas, cada qual com um significado próprio, estipulado previamente. Conforme citado no exemplo anterior, em HTML a *tag* <B> é usada para delimitar um texto que será apresentado em negrito. Cada *tag* possui uma função específica, seja simplesmente para marcar um texto, ou para definir uma estrutura mais complexa como, por exemplo, uma tabela, um formulário, etc. Além disso, ao se delimitar um texto por marcas, é possível estabelecer um certo conteúdo semântico que pode ser tratado e manipulado por programas de computador.

Desde 1994 existe um comitê internacional chamado W3C (*World Wide Web Consortion*), responsável por desenvolver e manter padrões para a Internet. O W3C é dirigido por Tim Berners-Lee, o inventor da *World Wide Web* (W3C, 2005). Dentre esses padrões encontram-se dezenas de linguagens de marcação. É necessária a existência da padronização para que as LM possam ser compartilhadas e utilizadas em nível mundial. Para um fabricante desenvolver um *browser* (software para navegação na Internet), por exemplo, é necessário que ele reconheça que a *tag* <B> indica negrito, e assim por diante.

As LM possuem uma ancestral comum chamada SGML (*Standard Generalized Markup Language*), um padrão internacional independente de sistemas e máquinas, para a definição de métodos de representação de textos em formato eletrônico. Trata-se de uma linguagem complexa a partir da qual surgiram diversas outras, como a HTML e a XML. Todas as LM devem ser capazes de diferenciar as *tags* (que compõem a estrutura do documento) do texto propriamente dito (que constitui o conteúdo do documento). Tendo essa característica, uma ferramenta como um *browser*, por exemplo, tem condições de

identificar as marcas contidas e apresentar ao usuário somente o conteúdo texto do documento, sem as *tags*.

As LM permitem criar documentos com uma estrutura de representação que seja compreendida por diversos sistemas de software, independentemente da máquina onde estão sendo executados. Na prática isso quer dizer que um mesmo documento pode ser manipulado em um computador pessoal ou em um *mainframe* (computador de grande porte), sem que haja dependência de um determinado sistema operacional ou sistema proprietário. Os documentos se tornam independentes do sistema ou *software* onde são visualizados, contribuindo fortemente para a disseminação da informação em ambientes computacionais heterogêneos, isto é, com tecnologias diferentes da fonte onde o recurso de informação foi produzido.

As LM contribuíram para tornar a comunicação livre de formatos proprietários, uma vez que elas representam padrões abertos e podem ser usadas livremente. Apesar de existir o órgão regulamentador W3C, ninguém é proprietário (dono) de uma LM, não é necessário se pagar direitos para elaborar um documento baseado em LM.

As características citadas nos parágrafos anteriores, isto é, a independência de plataforma e o uso livre da linguagem, foram essenciais para o grande sucesso da linguagem HTML, fato que ocasionou o surgimento de bilhões de documentos para a Internet. Essa grande quantidade de documentos espalhados pelo globo contribuiu enormemente para a disseminação de informações. Pesquisadores e estudantes, espalhados em diversas regiões do planeta podem, agora, compartilhar informações e conhecimentos, acelerando o desenvolvimento da sociedade. Apesar disso, é importante destacar que essa

“avalanche” de documentos contribuiu para estabelecer um certo “caos” na Internet, principalmente no que se diz respeito à recuperação de informações relevantes, mesmo utilizando-se de mecanismos de busca como o Google.

As LM têm diversas aplicações no campo da CI. Os tópicos seguintes apresentam algumas contribuições:

- A idéia de Bush (1945), com a proposta do MEMEX, foi finalmente concretizada por Tim Berners-Lee, ao criar a Internet e a linguagem HTML. A proposta de Bush, de criar termos associados que pudessem ser consultados rapidamente, foi concretizada através da criação de *links* em HTML. Documentos pequenos, ou mesmo documentos grandes, podem ser divididos e interligados através de *links* que realizam associação entre termos. Um livro, por exemplo, pode ser decomposto em diversos arquivos, cada um contendo um capítulo, ou mesmo uma seção, de um livro.
- Documentos criados a partir das LM tornaram-se acessíveis através de dispositivos móveis, como celulares e PDAs (*Personal Digital Assistence*) facilitando a disseminação de informação. O protocolo *Wireless Aplication Protocol* (WAP) tem possibilitado que celulares, e outros dispositivos, acessem páginas Internet, criadas a partir da linguagem *Wireless Markup Language* (WML).
- O surgimento do padrão RSS (*RDF Site Summary*), um padrão desenvolvido a partir das linguagens de marcação, cria um ambiente adequado para se compartilhar informações entre *sites* e aplicações, facilitando a integração entre diferentes sistemas. Foi criado a partir da linguagem XML com o objetivo de agilizar o processo de troca e coleta de informações. Muitos *sites* já disponibilizam esse serviço gratuitamente, principalmente *sites* de notícia como jornais, revistas e portais. Rucavina (2005) desenvolveu um serviço, baseado em RSS, para avisar aos usuários de uma biblioteca a data de devolução dos livros emprestados.

Outras contribuições das LM para o campo da CI são descritas nas seções seguintes.

3.5.3. HTML

Apesar de a HTML ser uma “velha conhecida” da CI, torna-se necessário descrever alguns de seus aspectos visando sua comparação com a XML. Em HTML as *tags* determinam o início e o fim do texto marcado como uma unidade ou elemento de informação, isto é, através das *tags* é possível representar uma informação. Por exemplo, o texto delimitado por <TITLE> se refere ao título do documento, assim como o texto delimitado por <P> corresponde a um parágrafo.

Os mecanismos de busca podem analisar o conteúdo delimitado pela *tag* <TITLE> a fim de verificar a ocorrência de uma palavra pesquisada. Por exemplo, ao se pesquisar pelo termo “marcação”, é perfeitamente possível ao mecanismo de busca recuperar o documento do

Quadro 4. É extremamente importante que o título de um documento HTML sempre utilize palavras-chave, isto é, termos que dizem respeito ao conteúdo do documento. Além disso, é importante que o título esteja bem descrito, pois ele será exibido nos resultados de um mecanismo de busca. Empresas especializadas em criação de Web *sites* apontam que “mais da metade dos usuários da Internet no mundo usa, todos os dias, um mecanismo de busca” (Symantec, 2005). Como os mecanismos de busca analisam o conteúdo do título, este possui relação direta com a facilidade de recuperação do documento.

Quadro 4. Exemplo de um documento HTML.

```
<HTML>
  <HEAD>
    <TITLE>
    O papel das linguagens de marcação para a ciência da informação.
    </TITLE>
    <META name=“author” content=“Sérgio Furgeri”>
    <META name="keywords" content=“html,xml,informação">
    <META name=“revised" content=“Raimundo Nonato">
  </HEAD>
  <BODY>
  <P> Este artigo procura apresentar uma visão geral sobre as <FONT color=”red”> Linguagens de Marcação </FONT> e suas implicações no campo da Ciência da Informação. </P>
  </BODY>
</HTML>
```

Outro aspecto muito importante, mostrado no Quadro 2, é a utilização da *tag* <META>. Trata-se de uma *tag* que possibilita realizar uma série de tarefas, tais como redirecionar páginas para outro *site*, recarregar uma mesma página de tempos em tempos, produzir efeitos de animação durante a transição de uma página para outra e, o que interessa diretamente a CI, a possibilidade de adicionar informações sobre o documento, isto é, de se criar metadados. A *tag* <META>, como o próprio nome sugere, permite definir metadados através de um par de atributos, chamados “*name*” e “*content*”. O atributo “*name*” armazena o nome de um parâmetro, enquanto que o atributo “*content*” armazena o conteúdo desse parâmetro. Observe a sintaxe seguinte: <META ***name***="nome do Parâmetro" ***content***="conteúdo da informação">

Com isso, torna-se possível adicionar meta-informações para que mecanismos de busca possam melhor identificar o conteúdo do documento, como nos três exemplos descritos no Quadro 2: o autor (*author*) do documento é “Sérgio Furgeri”, as palavras-chave (*keywords*) são “html, xml, informação” e o revisor (*revised*) foi “Raimundo Nonato”.

Apesar da enorme importância da HTML para a disseminação da

informação através da Internet, ela é muito limitada no que diz respeito à semântica dos termos. Somente as *tags* <*TITLE*> e <META> podem ser usadas para fornecer algum significado ao conteúdo do documento. Essas marcações se restringem ao cabeçalho do documento, permitindo a definição de metadados muito parciais, tal qual um sistema de fichas quando apresenta apenas alguns dados sobre uma obra.

Voltando ao Quadro 2, não é possível, por exemplo, fornecer significado a palavra “campo”, presente no corpo do documento. É praticamente impossível para um mecanismo de busca reconhecer que a palavra “campo” se refere à área da CI, coisa tão trivial para seres humanos. Um mecanismo de busca poderia interpretar a palavra “campo” como sendo campo de futebol, campo de plantação de tomates ou campo magnético. Esse tipo de interpretação é muito complexo para um mecanismo de software.

A HTML torna-se limitada para representar o significado das informações presentes no documento, pois ela não foi concebida para esse fim. A HTML foi desenvolvida para definir como a informação deve ser apresentada e não o que ela significa. O conjunto fixo de *tags* e atributos não permitem que representações mais aprimoradas sejam criadas. Esse aspecto, além de dificultar o trabalho humano, praticamente impossibilita que computadores troquem informações entre si, de maneira “inteligente”. Essa falta de flexibilidade da HTML impede que os diversos tipos de comunidades e organizações possam trocar documentos e informações de maneira mais efetiva através da Internet. De maneira geral pode-se dizer que representar uma informação em HTML é algo bem limitado.

3.5.4. XML

Antes de abordar efetivamente os aspectos essenciais da XML que podem ser usados na representação da informação, será fornecida uma visão geral sobre a linguagem, assim como seus principais pontos comuns e divergentes em relação a HTML.

A linguagem XML foi criada por Jon Bosac da empresa de *software* Sun, uma das principais empresas da área de computação. Assim como HTML, XML foi definida como um padrão de marcação para ser utilizado na Internet, constituindo-se em uma versão simplificada da SGML, cujo objetivo principal foi fornecer aos desenvolvedores da Internet uma maneira de definir e criar seus próprios marcadores e atributos, em vez de estarem restritos ao esquema de marcação da HTML. O próprio significado da XML sugere essa característica, pois é uma Linguagem de Marcação Extensível.

Da mesma forma que HTML, XML é um padrão aberto e independente de plataforma, entretanto, enquanto a HTML apenas trata de especificar a formatação de uma palavra ou um trecho de texto, a XML trata de criar estruturas para representar seu significado. Enquanto a HTML indica como algo deve ser exibido, a XML procura indicar o que a informação significa. Pode-se considerar que a XML é uma evolução da HTML. Apesar disso, tem ocorrido uma "convivência pacífica" entre HTML e XML. No princípio do aparecimento da XML muito se questionava se ela seria a substituta da HTML, fato que ainda não ocorreu. Conforme citado anteriormente, o que o W3C tem incentivado é que desenvolvedores em HTML criem seus documentos de acordo com a especificação da XHTML, atualmente em sua versão 2.0.

A XML pode ser usada em conjunto com a HTML na elaboração de

documentos para a Internet. O mais freqüente é inserir trechos de código XML em uma página HTML, processo conhecido pelo nome de *Data Islands* (SILVA, 2001). Isso permite que um mesmo documento, escrito em XML, possa ser reutilizado em diversos outros documentos.

A seguir será apresentado um exemplo que ilustra algumas diferenças da estrutura criada pela XML em relação a HTML. Tanto o Quadro 5, quanto o Quadro 6, contêm as mesmas informações publicadas em HTML e XML.

Quadro 5. Documento HTML para descrição de livros.

```
<HTML>
  <HEAD><TITLE>Livros</TITLE></HEAD>
  <BODY>
    ISBN 85-719-4797-X, título Ensino Didático da Linguagem XML, autor Sérgio Furgeri, editora Érica, publicado em 2001, preço atual R$ 20,00, disponível em estoque.
  </BODY>
</HTML>
```

Quadro 6. Trecho de um documento XML para descrição de livros.

```
<livros>
  <livro isbn="85-719-4797-X">
  <titulo>Ensino Didático da Linguagem XML</titulo>
  <autor>Sérgio Furgeri</autor>
  <editora>Érica</editora>
  <ano>2001</ano>
  <preco>20,00</preco>
  <disponivel>sim</disponivel>
  </livro>
</livros>
```

A estrutura criada pela XML apresentada no Quadro 6 é facilmente legível por seres humanos e máquinas. É fácil identificar que se trata de um catálogo com informações sobre livros. Cada livro possui um número de ISBN, um título, autor, editora, ano, preço e disponibilidade (poderiam existir muitos outros dados). É perfeitamente possível para um *software* de busca reconhecer que o número 2001 corresponde ao ano de lançamento do livro, o que não acontece com o documento em HTML do Quadro 3. Fica difícil para um *software* conseguir

identificar cada unidade de informação através da HTML, entretanto na XML cada unidade de informação está devidamente delimitada por uma *tag* que fornece um certo significado.

Segundo Almeida (2002), “A definição de *tags* próprias confere à linguagem XML habilidades semânticas, que possibilitam melhorias significativas em processos de recuperação e disseminação da informação”. Esses são os aspectos mais importantes da XML que interessam diretamente a CI. Eles possibilitam a criação de metadados, isto é, estruturas de representação da informação. Através da elaboração de estruturas padronizadas, torna-se perfeitamente possível que diversas comunidades consigam fazer intercâmbio e comparação de dados, além de otimizar a recuperação de informações.

Observando os marcadores usados no Quadro 6 é possível detalhar as informações sobre os livros, criando dados sobre eles, isto é, metadados. Os livros passam a possuir diversas propriedades identificáveis e distintas. O nível de detalhamento dos dados pode ser ampliado na medida da necessidade. Por exemplo, a *tag* editora poderia conter outros elementos filhos, como nome, endereço, estado e assim, detalhando ainda mais a estrutura e ampliando a representação da informação. Com esse pequeno exemplo é possível observar que a XML constitui um importante recurso na criação de metadados, que por sua vez, constituem-se num recurso de vital importância para a representação e recuperação de informações (RODRIGUEUZ, 2002, p. 29 a 33).

A XML permite representar o documento como uma estrutura padronizada, em forma de árvore, conforme apresenta a Figura 8.

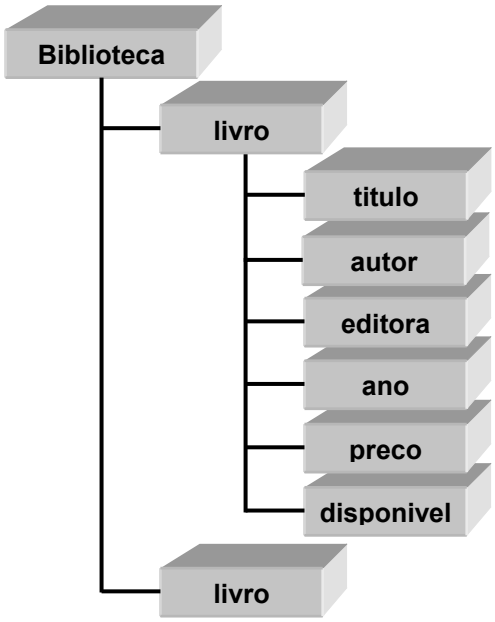

Figura 8. Estrutura em forma de árvore criada pela XML.

Dessa forma, torna-se perfeitamente possível para um *software* realizar uma varredura no catálogo com a finalidade de recuperar informações, que podem ser do tipo: quais os livros do catálogo que foram publicados pela editora Érica, ou quais os títulos dos livros que possuem a palavra XML em seu conteúdo, e assim por diante. A estrutura criada em forma de árvore, além de proporcionar procedimentos de pesquisa, torna possível a manipulação dos elementos do catálogo, permitindo inserção ou remoção de elementos.

Pelo fato de XML ser uma linguagem que permite ao criador elaborar suas próprias *tags*, torna-se possível se desenvolver outras linguagens de marcação a partir da XML. Diversos padrões para marcação de documentos podem ser construídos, baseados na estrutura que a XML oferece. Já existem dezenas, talvez centenas, de outras linguagens que foram desenvolvidas a partir da XML, nas mais diversas áreas do conhecimento (DESIGNERZ, 2005). Isso possibilita que comunidades formem vocabulários próprios, fornecendo significado a termos específicos e utilizados em conjunto. A criação de novas linguagens é possível

através da utilização do DTD (*Document Type Definition*) ou XML *Schema.* Este representa uma evolução daquele. Apesar de o DTD ser ainda bastante utilizado, será focado apenas o XML *Schema*, mais especificamente com relação à representação da informação.

3.5.4.1. Limitações da linguagem XML

Apesar do exposto até aqui, e de todas as vantagens da XML, há de se considerar suas limitações. Os parágrafos seguintes descrevem não apenas suas limitações, mas também sugerem o que pode ser realizado para superá-las.

Por exemplo, analisar dois documentos diferentes (em XML), que seguem padrões diferentes de construção, e utilizam a mesma *tag* é um problema em potencial. Considerando-se que diferentes comunidades utilizam estruturas diferentes para representar uma mesma informação, é improvável que um computador possa reconhecer o verdadeiro significado dos termos. *Tags* com nomes iguais podem possuir significados e conteúdos diferentes. Se, por exemplo, a *tag* <TITULO> é usada em dois documentos diferentes, representando informações diferentes, como tratar isso? Uma alternativa a esse problema se refere ao uso de *namespaces*. Segundo Marchal (2000), *namespace* pode ser definido como “um mecanismo para identificar os elementos da XML”. Trata-se de um prefixo associado a *tag* que permite torná-la exclusiva. Por exemplo, um documento elaborado para a área CI poderia ter a *tag* <CI:TITULO>, já para a área de computação poderia ser <CO:TITULO>, uma maneira de diferenciar, e tornar exclusiva, uma *tag* qualquer. O uso de *namespaces* torna possível a criação de um vocabulário controlado, para descrever algum domínio do conhecimento.

A estrutura definida pela XML permite representar a mesma informação de maneiras diferentes. Por exemplo, suponha que seja necessário representar a seguinte informação: o autor Bush escreveu um artigo de título "*As We May Think*". Qual é a estrutura correta para representar essa informação? O Quadro 7 apresenta quatro possibilidades diferentes para realizar isso. Existem ainda diversas outras.

Quadro 7. Estruturas diferentes para a mesma informação.

<artigo>
  <titulo> As May We Think **</titulo>**
  <autor>
  <nome> Bush **</nome>**
  </autor>
</artigo>
	<autor>
	  <nome> Bush **</nome>**
	  <artigo> As May We Think **</artigo>**
	</autor>

<autor>
  <nome> Bush **</nome>**
  <publicação tipo="artigo**">** As May We Think **</publicação>**
</autor>
	<artigo autor="Bush**" titulo="**As We May Think**"/>**

Esse problema pode inviabilizar a utilização da XML na descrição de metadados. Para contornar esse problema é criado um outro tipo de documento que contém regras a partir das quais um documento XML é criado. Essa é a função do *Schema* XML, descrito na seção seguinte.

Outras limitações, talvez menos importantes, são: a baixa escalabilidade e a estrutura em forma de árvore.

A escalabilidade se refere à ampliação da estrutura do documento. Isso pode representar um problema quando o número de registros for muito grande. Como representar um catálogo de uma biblioteca contendo milhões de registros? Além disso, antes de adicionar novos metadados em um documento (em nosso exemplo considere a inserção da data de publicação), o *Schema* deve ser alterado.

Em XML a estrutura gerada sempre se apresenta em forma de árvore, porém nem sempre as informações podem ser assim representadas. Uma outra linguagem chamada RDF (*Resource Description Framework*), descrita na seção 3.5.6, foi proposta pelo W3C com o propósito de reduzir as limitações da XML.

3.5.5. XML Schema

Apesar de a XML permitir que se criem marcações na medida da necessidade, torna-se necessário considerar que as coisas não são bem assim "totalmente livres". Para que um documento XML seja criado, e possa ser compreendido em diferentes contextos, é necessário que esteja de acordo com certas regras, definidas em outro documento chamado de XML *Schema*. Por meio dele se definem padrões que o autor deve seguir para que o documento seja considerado válido. Para melhor compreensão dos objetivos do XML *Schema* será apresentado um possível documento em XML para a confecção de um catálogo de livros. O Quadro 8 apresenta o documento XML completo e o Quadro 9 um possível XML *Schema* para realizar sua validação.

Quadro 8. Catálogo de livros em XML.

```
<?xml version="1.0" encoding="iso-8859-1"?>
  <biblioteca xmlns="http://www.w3schools.com"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.w3schools.com biblioteca.xsd"
    universidade="PUC">
      <livro isbn="85-7194-797-X">
         <titulo>Ensino didático da Linguagem XML</titulo>
         <autor>Sérgio Furgeri</autor>
         <keywords>
            <key>XML</key>
            <key>HTML</key>
            <key>DTD</key>
            <key>CSS</key>
         </keywords>
         <editora>Érica</editora>
         <ano>2001</ano>
         <preco moeda="R$">20</preco>
         <disponivel>sim</disponivel>
      </livro>
  </biblioteca>
```

O catálogo da biblioteca apresentado no Quadro 8 foi elaborado de acordo com as regras definidas em “biblioteca.xsd” apresentado no Quadro 9.

Quadro 9. Schema para o catálogo de Livros (desconsiderar numeração de linhas)

```
1.  <?xml version="1.0"?>
2.   <xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
3.    targetNamespace="http://www.w3schools.com"
4.    xmlns="http://www.w3schools.com"   elementFormDefault="qualified">
5.     <xs:element name="biblioteca">
6.       <xs:complexType>
7.         <xs:sequence>
8.           <xs:element name="livro" minOccurs="1" maxOccurs="1000">
9.             <xs:complexType>
10.              <xs:sequence>
11.                <xs:element name="titulo" type="xs:string"/>
12.                <xs:element name="autor" type="xs:string" minOccurs="1" maxOccurs="10"/>
13.                <xs:element name="keywords">
14.                  <xs:complexType>
15.                    <xs:sequence>
16.                      <xs:element name="key" type="xs:string" minOccurs="3"
    maxOccurs="10"/>
17.                    </xs:sequence>
18.                  </xs:complexType>
19.                </xs:element>
20.                <xs:element name="editora" type="xs:string"/>
21.                <xs:element name="ano">
22.                  <xs:simpleType>
23.                    <xs:restriction base="xs:integer">
24.                      <xs:minInclusive value="1700"/>
25.                      <xs:maxInclusive value="2010"/>
26.                    </xs:restriction>
27.                  </xs:simpleType>
28.                </xs:element>
29.                <xs:element name="preco" minOccurs="0">
30.                  <xs:complexType>
31.                    <xs:simpleContent>
32.                      <xs:extension base="xs:decimal">
33.                       <xs:attribute name="moeda" type="xs:string" use="required"/>
34.                      </xs:extension>
35.                    </xs:simpleContent>
36.                  </xs:complexType>
37.                </xs:element>
38.                <xs:element name="disponivel" type="xs:string"/>
39.              </xs:sequence>
40.             <xs:attribute name="isbn" type="xs:string" use="required"/>
41.           </xs:complexType>
42.         </xs:element>
43.       </xs:sequence>
44.     <xs:attribute name="universidade" type="xs:string" use="required"/>
45.   </xs:complexType>
46.  </xs:element>
47. </xs:schema>
```

O *Schema* do Quadro 9 apresenta uma série de regras que devem ser seguidas pelo documento do Quadro 8 (o catálogo da biblioteca). Cada elemento presente no *Schema* define uma regra a ser seguida (GRAVES, 2003). As regras estabelecidas podem ser assim definidas:

- As *tags* do arquivo XML são declaradas no XML *Schema* através da palavra *element*;

- A linha 5 inicia as declarações do elemento de nome "biblioteca". Ele deve possuir de um (minOccurs[1]="1" na linha 8) a mil (maxOccurs="1000" na linha 8) elementos filhos de nome "livro"; as declarações correspondentes ao elemento "biblioteca" se encerram na linha 46. A linha 44 define que o elemento "biblioteca" deve conter um atributo obrigatório (*required*) de nome "universidade" que, por sua vez, deve conter um conteúdo do tipo texto (*string*);
  - As linhas 8 a 42 declaram um elemento de nome "livro" que deve conter a seguinte seqüência de elementos filhos:
    - Linha 11: um elemento com o nome "título" de conteúdo tipo texto (*string*);
    - Linha 12: de 1 a 10 elementos com o nome "autor" de conteúdo tipo texto;
    - Linha 13: um elemento com o nome "*keywords*" que, por sua vez, pode possuir de 3 a 10 elementos com o nome "*key*" de conteúdo texto (linha 16);
    - Linha 20: um elemento com o nome "editora" de conteúdo tipo texto (*string*);
    - Linha 21: um elemento com o nome "ano" de conteúdo tipo inteiro (*integer*), com valores entre 1700 (linha 24) e 2010 (linha 25);
    - Linha 29: um elemento com o nome "preco" de conteúdo numérico decimal (linha 32) que contenha um atributo obrigatório (required), com o nome "moeda" (linha 33) e de conteúdo tipo texto (*string*);
    - Linha 38: um elemento com o nome "*disponivel*" de conteúdo tipo texto (*string*);

[1] minOccurs e maxOccurs são propriedades definidas em XML *Schema*.

- Linha 40: o elemento de nome "livro" deve conter um atributo obrigatório de nome "isbn" de conteúdo tipo texto (*string*);
- Todos os elementos devem, obrigatoriamente, ser usados no documento XML na mesma ordem em que foram declarados no *Schema*;

Pelos comentários citados, observa-se que o XML *Schema* contém regras rígidas que determinam a estrutura que o documento XML deverá assumir. Existem apenas alguns pontos flexíveis, mas que também estão sujeitos a determinadas regras. Por exemplo, o elemento “preco” da linha 29 é opcional (minOccours=”0”), entretanto, quando for usado, ela deve estar imediatamente depois do elemento “ano” e imediatamente antes do elemento “disponível”. Podem existir de 3 a 10 elementos “key”, mas não fora dessa faixa.

Conforme exposto, pode-se observar que o uso de *Schemas* permite que sejam criados documentos XML estruturados, seguindo determinados padrões, porém, com alguma flexibilidade. De forma resumida, as principais declarações do XML *Schema* são apresentados na Tabela 5.

Tabela 5. Principais definições do XML Schema.

Declaração	Exemplos
Declarar os elementos permitidos e seu conteúdo (filhos, texto, vazio);	O elemento <biblioteca> contém o filho <livro>; O elemento <editora> contém um texto.
Declarar os elementos filhos, sua quantidade, e a ordem em que aparecem;	O elemento <livro> contém o filho <titulo>, seguido de um ou mais filhos <autor>, seguido dos filhos <keywords>, <editora>, <ano>, seguido talvez pelo filho <preco> e seguido pelo filho <disponível>.
Declarar os atributos de um elemento;	O elemento <livro> contém o atributo “isbn”.
Declarar os tipos de dados para elementos e atributos, além da faixa de valores permitidos;	O elemento <ano> contém um número inteiro entre 1700 a 2010; O atributo “isbn” contém um texto.
Declarar valores padrão e fixos para elementos e atributos.	O atributo “moeda” do elemento <preco> contém o texto “R$” como padrão, apesar de outros tipos de moeda serem permitidos.

Para eliminar dúvidas serão apresentadas algumas analogias. O *Schema*

pode ser visto como um molde a partir do qual objetos podem ser arranjados ou armazenados. Por exemplo, os livros se encaixam no *Schema* prateleira, o bolo se encaixa no *Schema* forma. Para um usuário familiarizado com algum tipo de *software* para banco de dados (Microsoft Access, por exemplo) é fácil notar que é impossível inserir um dado do tipo texto como, por exemplo, o nome de uma pessoa, em um campo definido como tipo numérico (o *Schema*). As coisas “não se encaixam”. De forma semelhante, só é possível usar elementos XML quando declarados e definidos em *Schemas*. O uso de *Schemas* permite se definir padrões para descrição de informações. Uma última analogia: se uma pessoa precisa desenhar uma figura qualquer e só dispõe das formas geométricas quadrado, triângulo e círculo, não será possível usar uma elipse, pois o esquema não permite isso. Só é possível utilizar aquilo que é permitido.

Pelo exposto, pode-se notar que o uso de *Schemas* permite que um documento escrito em XML defina o conhecimento de sua própria estrutura e significado, o que faz com que sua utilização seja mais adaptada às necessidades de automatização entre comunidades quaisquer: bibliotecas (ou qualquer outra instituição) podem compartilhar ou trocar dados entre si, uma vez que a estrutura de representação da informação é a mesma. Baseado no XML *Schema* do Quadro 9, diferentes bibliotecas podem elaborar seus catálogos, de forma padronizada, para permitir que uma ferramenta de *softwar*e, desenvolvida para a Internet, possa realizar buscas no acervo de todas elas e trazer um resultado ao usuário. Muitas são as aplicações que podem ser desenvolvidas nessa área baseadas em XML.

3.5.6. RDF

A RDF (*Resource Description Framework*) é uma linguagem desenvolvida pelo W3C para codificar, trocar e reutilizar metadados na *Web* (W3C, 2004). Para a CI, a RDF pode ser concebida como uma linguagem que possibilita a elaboração de uma "estrutura para a descrição de recursos", ou seja, permite a descrição de metadados, ou ainda, a representação do conhecimento. Conforme Holzner (2001, p. 678), "a RDF é armazenada em documentos separados dos recursos que ela descreve". Recursos são, normalmente, documentos eletrônicos disponíveis na *Web*, ou qualquer outro tipo de obra (livro, CD, etc).

Elaborada a partir da XML, tem como objetivo prover intercâmbio de informações entre aplicações, garantindo o controle do significado. Para isso, define um vocabulário padrão para descrever coisas ou objetos e permitir a troca de metadados entre diferentes aplicações. A RDF permite criar declarações sobre objetos através de propriedades que representam um relacionamento entre recursos (Meissner, 2004). Uma declaração é realizada através de triplas do tipo "recurso-propriedade-valor", em que:

- **Recurso** – é o sujeito de uma declaração. Pode ser um *website* ou parte dele, ou ainda um objeto não acessível via *Web*. Um recurso pode ser acessado e reconhecido de forma única através de um URIs (*Uniform Resource Identifier*). Um artigo científico é um exemplo de recurso.
- **Propriedade** – é o predicado de uma declaração. Trata-se de um atributo usado para descrever um recurso. Um artigo científico pode conter diversas propriedades: nome do autor, título do artigo, data de publicação, etc.
- **Valor** – é o objeto de uma declaração. Representa o conteúdo das propriedades. (o conteúdo do nome do autor, o conteúdo do título do artigo, etc).

Exemplos de declarações podem ser:

- **www.noticias.com.br/saude.html** tem um **autor** chamado **Junior**
- **www.noticias.com.br/saude.html** tem um **titulo** chamado **o celular faz mal a saúde**.

As declarações RDF contêm o formato: "<sujeito> tem <predicado> <objeto>". Essa tripla é representada através de um grafo em que o sujeito (recurso) é um nó (node) em forma de elipse, o predicado (propriedade) é um arco com uma seta apontando para o objeto (valor), cuja representação é um retângulo (se for um literal) ou uma nova elipse (se representar um outro recurso). A Figura 9 apresenta um exemplo contendo declarações na forma de grafo.

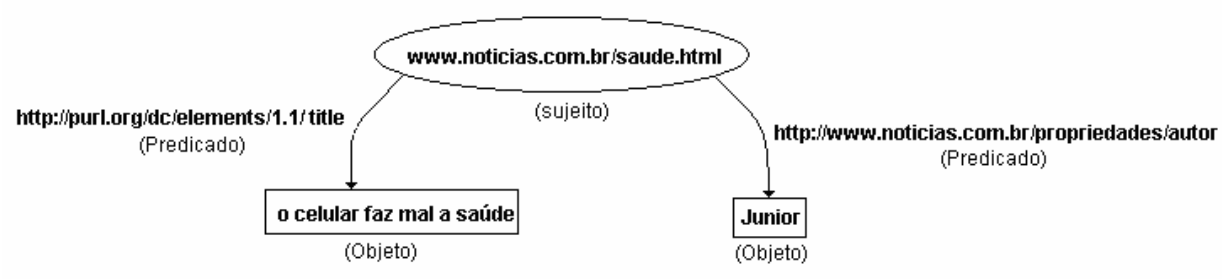

Figura 9. Representação gráfica de declarações em RDF.

Algumas considerações sobre o grafo da Figura 9:

- O recurso "**www.noticias.com.br/saude.html**", uma página imaginária da Internet, tem uma propriedade chamada "**title**"**,** cujo conteúdo é "**o celular faz mal a saúde**". A propriedade "**title**" possui referência a um *namespace*, uma espécie de vocabulário, usado para definir o significado de uma propriedade, no caso, da propriedade "title".
- O trecho de código RDF apresentado no Quadro 10 mostra como é possível se definir metadados sobre uma propriedade. A linha 1 define a URI onde o namespace se localiza. A linha 2 define o nome da propriedade (title), a linha 3 define um breve comentário sobre ela, a linha 4 faz uma descrição sobre ela. As linhas 5 em diante, definem outros metadados sobre a propriedade: o endereço onde o padrão se encontra, quando foi criado e modificado, etc. Essa estrutura permite a criação de vocabulários sobre termos.

Quadro 10. Trecho de um namespace para descrever uma propriedade (traduzido).

```
1. <rdf:Property rdf:about="http://purl.org/dc/elements/1.1/title">
2.    <rdfs:label xml:lang="en-PT">title</rdfs:label>
3.    <rdfs:comment xml:lang="en-PT">Um nome dado a um recurso.</rdfs:comment>
4.    <dc:description xml:lang="en-PT">Tipicamente, um titulo será um nome pelo qual um
   recurso é conhecido formalmente</dc:description>
5.    <rdfs:isDefinedBy rdf:resource="http://purl.org/dc/elements/1.1/" />
6.    <dcterms:issued>1999-07-02</dcterms:issued>
7.    <dcterms:modified>2002-10-04</dcterms:modified>
8.    <dc:type rdf:resource="http://dublincore.org/usage/documents/principles/#element" />
9.    <dcterms:hasVersion rdf:resource="http://dublincore.org/usage/terms/history/#title-004"
   />
10. </rdf:Property>
```

- O recurso “www.noticias.com.br/saude.html” possui também uma propriedade chamada “autor”, cujo conteúdo é “Junior”. Da mesma forma, a propriedade “autor” possui um significado próprio em uma URI. O objeto “Junior” poderia também possuir diversas propriedades como “email”, “cargo”, etc., ampliando o grafo e a cadeia de ligações entre recursos e objetos.
- Pelo fato de o projeto do RDF ter sido fortemente influenciado pelo padrão Dublin Core, as propriedades mais usadas seguem esse padrão (title, creator, etc). Apesar disso, é possível definir padrões próprios, como a propriedade “autor”, definida em “www.noticias.com.br/propriedades/autor”.

O Quadro 11 apresenta o código RDF/XML referente ao grafo da Figura 9.

Quadro 11. Arquivo RDF referente a Figura 9.

```
1.  <?xml version="1.0" encoding=”ISO-8859-1”?>
2.  <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
3.           xmlns:dc="http://purl.org/dc/elements/1.1/"
4.           xmlns:property="http://www.noticias.com.br/propriedades/">
5.       <rdf:Description rdf:about="www.noticias.com.br/saude.html">
6.            <property:autor>Junior</property:autor>
7.       </rdf:Description>
8.       <rdf:Description rdf:about="www.noticias.com.br/saude.html">
9.            <dc:title>o celular faz mal a saúde</dc:title>
10.      </rdf:Description>
11. </rdf:RDF>
```

Em resumo, as linhas de código significam o seguinte:

- Linha 1: o documento foi elaborado a partir da versão “1.0” da XML e possui o padrão “ISO-8859-1” para a codificação dos caracteres;

- Linha 2: aponta para um endereço da Internet que contém um *namespace* chamado “rdf“ que define o vocabulário para a descrição de elementos em RDF;
- Linha 3: aponta para um endereço da Internet que contém um *namespace* chamado “dc” (Dublin Core) que define um vocabulário padrão de termos em metadados;
- Linha 4: aponta para um endereço fictício da Internet que contém um *namespace* chamado “*property*” que define o vocabulário próprio para a descrição de propriedades. Isso foi realizado para demonstrar que vocabulários próprios podem ser criados de acordo com as necessidades;
- O restante das linhas descreve as triplas declaradas no grafo.

O modelo criado a partir da RDF possui aspectos semânticos que lhe conferem diversos aspectos positivos na representação de informação e conhecimento:

- A RDF fornece uma forma de representação da informação única, criando ligações exclusivas entre recursos e estabelecendo vocabulários através de *namespaces*.
- A RDF especifica os predicados que podem ser usados nas declarações e limita a faixa de valores que esses predicados podem assumir, reduzindo a probabilidade de declarações sem significado.
- A RDF é bastante indicada para a criação de metadados, pois um recurso é referenciado a um objeto através de um predicado com significado próprio. A ordem em que as declarações são realizadas é irrelevante.
- A RDF reduz os problemas da representação da informação em forma de árvore, criando uma estrutura mais flexível, possibilitando a formação de uma cadeia de informações e estabelecendo uma rede de conhecimento.
- Com RDF é possível criar um vocabulário controlado para descrever um domínio do conhecimento, fato que torna possível para qualquer organização publicar informações de maneira semântica. Com isso, agentes de software podem agir de maneira automática e inteligente sobre recursos da Internet, inferindo sobre o significado dos elementos.

A RDF opera em conjunto com a recomendação RDF *Schema* (RDFS) de

Fevereiro de 2004. Trata-se de uma linguagem para descrição de vocabulário RDF (W3C, 2004), responsável pela definição do modelo de dados a ser seguido pelos documentos RDF e que provê mecanismos para descrição dos recursos e propriedades. Um modelo de dados é semelhante ao grafo apresentado na Figura 5 que contém relações entre os recursos, definidos no documento RDF. Com a RDFS é possível definir os termos (as triplas) que serão usados nas declarações dos documentos RDF, uma maneira de se criar um vocabulário controlado.

3.5.7. RDF Schema

Conforme citado, a RDF *Schema* define um vocabulário controlado. Esse vocabulário permite estabelecer relações e restrições entre os recursos. Relações do tipo: o autor pertence a uma classe chamada pessoa, um site pertence a uma classe chamada recurso, um artigo que pertence a uma classe publicações deve possuir um título, cujo conteúdo deve ser um texto, e assim por diante. São criadas regras e restrições que devem ser seguidas pelos documentos RDF.

De acordo com os conceitos expostos, observa-se que a RDF, juntamente com a RDF *Schema*, possui atribuições que lhe conferem diversas possibilidades de uso na área da Ciência da Informação, uma vez que permite criar relações entre documentos e estabelecer significado aos termos por meio de namespaces. RDF provê uma estrutura mais flexível que XML, aproximando-se da forma como os seres humanos relacionam informações, isto é, através de associações. Um computador, que normalmente mantém dados em estruturas rígidas, passa a poder associar pedaços de informações com estruturas diferentes e de forma automática. Esse é o objetivo principal da WebSemantica proposta por Berners-Lee, isto é, dotar os computadores da possibilidade de fazer associações entre informações relacionadas (BERNERS-LEE, 2001).

Existem ainda outras possibilidades com o uso da RDF. Agüera propõe a criação de um tesauro baseado em RDF (AGÜERA, 2004). Outros autores propõem um padrão baseado em RDF chamado SKOS (*Simple Knowledge Organisation Systems*) para representação de tesauros e outros tipos similares de sistemas de organização de conhecimento (MILES, 2005).

Apesar dos inúmeros aspectos positivos citados do RDF, ainda existem algumas restrições impostas pelo modelo. Conforme Ahmed *et.al.* (2001), *apud* Garcia (2002), a RDF *Schema*

> "coloca limites para valores possíveis que podem ser usados com um predicado, sendo também possível especificar quais predicados podem ser usados com quais tipos de recursos. Mas para algumas aplicações isso não é suficiente, pois pode ser necessário representar restrições como por exemplo: se uma pessoa possui informação de cartão de crédito em seu metadado, então ela também deve ter um endereço. Restrições desse tipo não podem ser especificadas com o conjunto de restrições (constraints) usadas no RDF Schema corrente".

Para diminuir essas restrições, estendendo as possibilidades do modelo RDF, podem ser usadas outras linguagens como, por exemplo, OWL (*Web Ontology Language*), não tratadas aqui.

3.5.8. Modelo de Entidades e Relacionamentos

A CC desenvolve sistemas e bancos de dados a partir de modelos computacionais, cujo objetivo principal é armazenar e recuperar dados de maneira adequada. Um modelo de dados pode ser considerado como uma estrutura que descreve como os dados serão armazenados e como ocorrerão as ligações entre eles. Um modelo muito usado em CC para a descrição dos dados é conhecido como modelo ER (Entidade Relacionamento), ou ainda como Diagrama ER, desenvolvido originalmente por Peter Chen em 1976 (MACHADO e ABREU, 1995, p.30). A partir de 1998, a CI também desenvolveu um modelo

conceitual semelhante ao modelo E-R que apresenta conceitos e definições de entidades, relacionamentos e atributos, denominado FRBR (*Functional Requirements for Bibliographic Records*), uma tentativa de se estabelecer requisitos funcionais para registros bibliográficos centrados no usuário e em suas ações (MORENO, 2005) .

De forma geral, a elaboração do modelo E-R tem como objetivo principal permitir que os envolvidos no desenvolvimento do sistema (analistas, programadores, usuários, etc.) tenham uma boa noção do conteúdo do banco de dados a ser manipulado, buscando especificar a semântica dos dados. Um modelo ER simples pode ser visualizado na Figura 10. Trata-se de uma pequena parte de um possível banco de dados usado em bibliotecas.

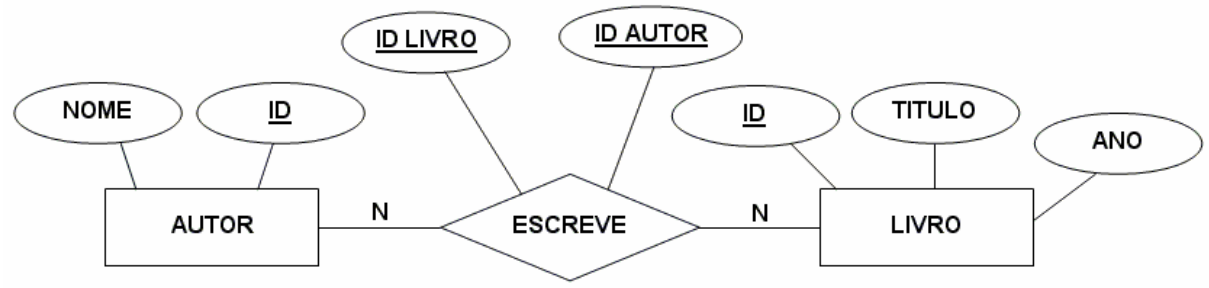

Figura 10. Parte de um modelo ER para Bibliotecas.

Como pode ser observado, o modelo contém símbolos padronizados (retângulos, losangos, elipses) com descrições internas em linguagem natural. Os retângulos representam conjuntos de entidades, ocorrências reais de objetos semelhantes como, por exemplo, autores. Cada retângulo contém um nome interno (normalmente um substantivo) que identifica o conjunto de entidades de forma exclusiva.

O modelo da Figura 10 está representando que o banco de dados armazenará diversas ocorrências de autores. Os losangos representam as

relações existentes entre os dados. Por exemplo, um Autor *escreve* um Livro. O verbo *escreve* indica que no mundo real (e no banco de dados) existirá uma ligação entre pelo menos um Autor e um Livro. As elipses representam os atributos que serão usados em cada entidade. Por exemplo, cada Autor terá uma identificação (id) e um nome (nome).

Após a criação do modelo é possível verificar se o banco de dados contempla todas as necessidades de armazenamento especificadas nos requisitos dos usuários. Uma vez definido pelo analista, em conjunto com os usuários, o modelo é entregue a equipe desenvolvimento para implementar o sistema. Como esse processo envolve diversos tipos de pessoas, é essencial se utilizar termos adequados e padronizados de maneira que todos possam compreender da mesma forma. Um erro de interpretação pode ser trágico e comprometer todo o desenvolvimento do sistema.

De certa forma, pode-se conceber que o modelo ER é capaz de representar o conhecimento a respeito dos dados envolvidos em um ambiente e suas relações existentes. Antes de desenvolver um sistema é necessário que o conhecimento dos usuários seja capturado. O modelo ER é uma das formas possíveis de capturar esse conhecimento e torná-lo conhecido pela equipe de desenvolvimento.

O capítulo 4 define alguns pontos importantes que podem auxiliar a CC no que diz respeito à padronização dos termos usada no processo de modelagem, item essencial para o sucesso de desenvolvimento de um sistema.

3.5.9. Orientação a objetos e UML

Segundo Ricarte (2001), o termo orientação a objetos “pressupõe uma

organização de *software* em termos de coleção de objetos discretos incorporando estrutura e comportamento próprios". O *software* elaborado através dessa metodologia será composto por um conjunto de objetos interligados.

Na programação orientada a objetos, um objeto é uma abstração dos objetos reais existentes. Em uma sala de aula, por exemplo, existem diversos objetos: alunos, cadeiras, mesas, lousa etc. Se for necessário manter controle sobre uma sala de aula, pode ser elaborado um *software* que manipula objetos desse tipo.

Um objeto do tipo mesa, por exemplo, possui uma estrutura, cor, dimensões, peso, enfim diversas propriedades (ou características). Diferentes objetos possuirão diferentes propriedades. Associado às propriedades do objeto, existe outro fator: as ações que podem ser realizadas com ele. Um objeto do tipo motor, por exemplo, pode estar desligado ou em funcionamento, pode existir uma ação para aumentar (acelerar) e uma para diminuir (frear) sua velocidade.

A programação orientada a objetos procura modelar, internamente no computador, a realidade dos objetos, permitindo a elaboração de um sistema que representa, ou pelo menos procura representar, a realidade de um ambiente. Conforme cita Ricarte (2001), os objetos têm dois propósitos: "promover o entendimento do mundo real e suportar uma base prática para uma implementação computacional". De forma semelhante, Matos (2002) afirma: "a proposta da Orientação a Objetos é permitir que os programadores organizem os programas da mesma forma que nossas mentes enxergam os problemas".

Para modelar uma realidade são definidas classes, estruturas iniciais a partir das quais os objetos são criados. Uma forma adequada para realizar a

modelagem das classes é através da linguagem UML (*Unified Modeling Language*). A UML é uma linguagem gráfica para definição de dados (não é uma linguagem de programação) que possui uma série de símbolos com significados próprios. Esses símbolos permitem criar um diagrama para representar as funcionalidades de um *software*. Sendo assim, a UML pode ser usada para representar o conhecimento. Através de seus símbolos gráficos, tanto o funcionamento de um sistema, quanto à realidade de um ambiente, podem ser representados.

A UML proporciona um ambiente adequado desde para o desenvolvimento de sistemas, desde o planejamento até a implementação. Matos (2002), descreve a UML como “uma ferramenta ideal para conceber, compreender, testar, validar, arquitetar e ainda identificar todos os possíveis comportamentos do sistema”.

Para representar graficamente todas as etapas de desenvolvimento de um sistema, a UML dispõe de uma série de modelos de diagramas, cada um com um objetivo específico. Um deles é o diagrama de classes, mostrado na Figura 11.

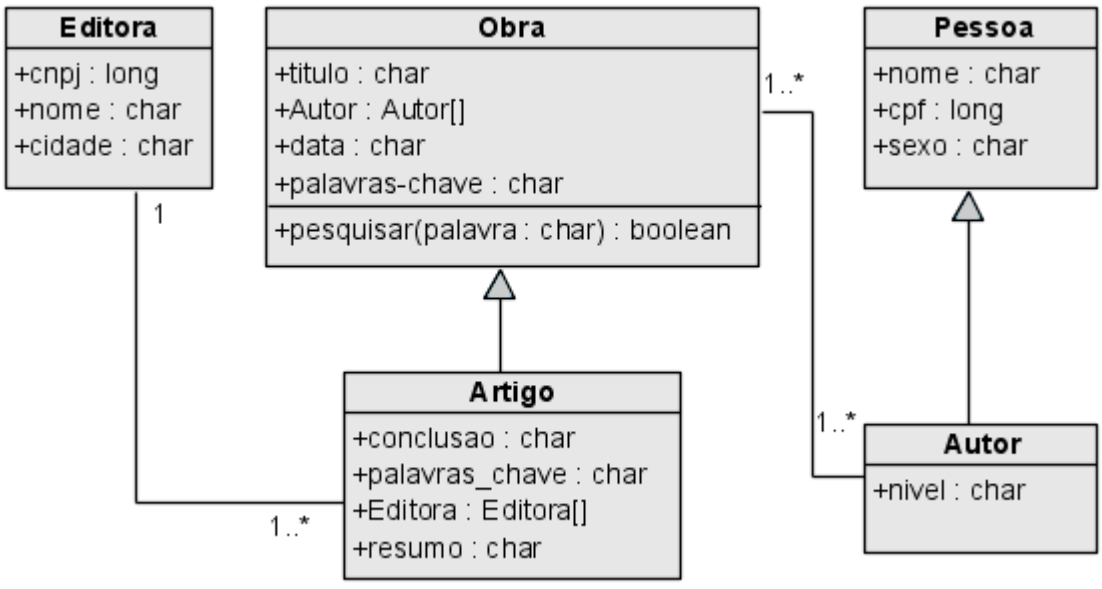

Figura 11. Diagrama de Classes conforme a UML.

O diagrama de classes da Figura 11, desenvolvido a partir da ferramenta *Visual Paradigm*, versão 5.2 (VPUG, 2006), em conformidade com os padrões definidos pela linguagem UML. O diagrama representa a realidade (simplificada) de um controle de publicações e define as classes a partir das quais os objetos serão criados. Cada classe é representada por um retângulo. Na parte superior de cada retângulo é inserido o nome da classe. No interior do retângulo são inseridas propriedades e operações. Enquanto as propriedades se referem às características que podem ser atribuídas ao objeto, os métodos se referem às ações que podem ser realizadas com ele. Por exemplo, a classe Obra é composta pelas propriedades titulo, data, palavras-chave, autores e pelo método pesquisar. O método pesquisar pode ser uma ação que recebe uma palavra e verifica se ela é uma palavra-chave da publicação.

Pelo diagrama apresentado, observa-se que apesar de a UML e a OO realizarem a representação do ambiente através de objetos, não é possível definir o significado do conteúdo desses objetos. Apesar de evidenciar a composição de um objeto e suas relações com outros objetos, fica difícil compreender o significado dos termos que compõem o objeto. Em outras palavras, apesar de ser possível representar a realidade de um ambiente, fica difícil representar o conhecimento inerente a ele. Por exemplo, é evidente que a propriedade título pertence ao objeto Obra, mas qual é o significado do termo título? O capítulo 4 procura discutir esse aspecto.

Apesar de a UML ser uma linguagem padronizada, a criação e decomposição do modelo de classes dependerá do julgamento do projetista e da natureza do problema. Algo semelhante ocorre na CI no processo de documentação e representação do conhecimento.

Este capítulo apresentou as principais iniciativas da CC com relação à representação da informação e do conhecimento. O próximo capítulo descreverá algumas iniciativas em conjunto entre a CI e a CC, cujo objetivo é o de demonstrar que as áreas se beneficiam mutuamente quando atuam de maneira interdisciplinar. Destacam-se a criação de ontologias e o projeto da Web Semântica.

4. RELAÇÕES INTERDISCIPLINARES

Este capítulo apresenta diversos aspectos convergentes entre as áreas de CI e CC no que tange a representação do conhecimento. O objetivo é demonstrar que essas áreas “agregam valor” quando atuam em conjunto.

4.1. Uso de uma terminologia

De acordo com Moreira e Oliveira (2005), diversos autores demonstraram a importância da Terminologia para os sistemas de classificação usados pela CI. O aspecto fundamental está na busca pela compreensão do significado de um termo, dentro de um domínio específico. Para Lara (2001) o termo é a unidade básica da terminologia: “A unidade básica da Terminologia é o termo, designação verbal de um conceito dentro de um domínio específico”. Isso demonstra que uma vez estabelecido o significado de um termo de maneira única, torna-se mais precisa sua representação e recuperação.

Da mesma forma, o uso de uma terminologia pode ser um agente facilitador no desenvolvimento de sistemas computacionais, isto é, ela pode contribuir também com a CC. Moreira e Oliveira (2005) relatam que “a ciência da computação, mais particularmente a representação do conhecimento, pode se beneficiar dos resultados e métodos oriundos da terminologia“.

A Terminologia pode contribuir para a melhoria de diversos aspectos no campo da CC como, por exemplo, permitir a padronização dos termos mais adequados para cada conceito da área, facilitando a discussão pelos pares. Caso contrário um sistema pode ser compreendido apenas por uma comunidade específica que o desenvolveu. A Terminologia pode contribuir também com a subárea de Engenharia de Software na escolha de termos mais semanticamente

adequados para a criação de modelos (E-R, OO, etc). A Terminologia pode ajudar ainda diversas outras subáreas, melhorando o aspecto semântico das definições.

Outro aspecto que começa ser aproveitado pela CC se refere à adoção de nomenclaturas e convenções. O uso de uma nomenclatura semântica começa a ser adotada no *design* de sites[2]. O uso de nomenclaturas e convenções pode contribuir para tornar mais compreensível o conteúdo de linguagens e sistemas.

Conforme citado, todos os termos usados na elaboração de um modelo devem ser padronizados e devem denotar um significado conhecido e único. Isso é importante não apenas para manter a equipe de desenvolvimento em "harmonia semântica", mas também pode ser importante também na integração com outros sistemas.

Ao analisar o modelo da Figura 10, apresentado no item 3.5.8, parece estar claro o significado dos termos Autor e Livro, pois denota claramente um Autor como sendo uma pessoa (um ser humano) dotada de plenas capacidades intelectuais, e Livro como um objeto (físico ou digital) com conteúdo escrito. O termo *escreve,* que representa o relacionamento entre as duas entidades, parece também deixar claro o significado da relação entre Autor e Livro.

O mesmo modelo poderia ter sido elaborado com outros termos. A Figura 12 apresenta o mesmo modelo com outro termo (*possui*) usado para descrever o relacionamento entre Autor e Livro.

[2] The Beauty of CSS Design – www.csszengarden.com

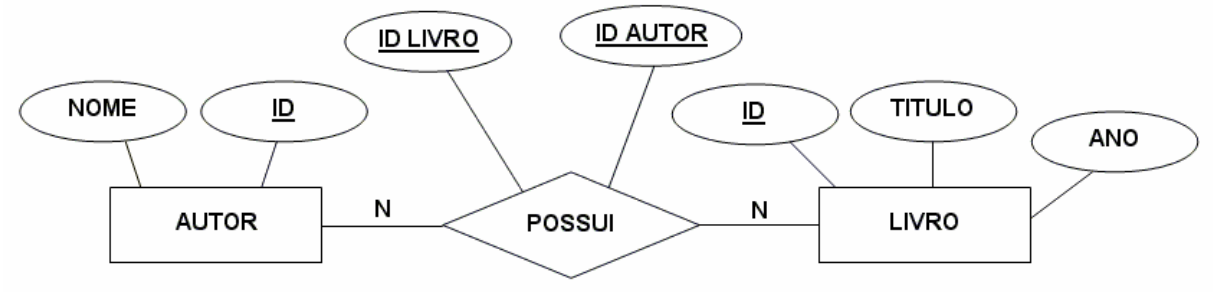

Figura 12. Modelo ER com o termo do relacionamento trocado.

Como pode ser observado, o novo modelo pode causar um erro de interpretação apenas pela substituição do termo *escreve* por *possui*. O termo *possui* pode fazer com que uma pessoa o interprete como possessão e não como autoria, conforme o modelo original.

Os termos usados para descrever os atributos das entidades também devem ser interpretados de forma correta para não gerar dúvidas. Por exemplo, o termo *título*, usado como atributo da entidade Livro, expressa corretamente seu significado. Se fosse usado o termo *nome* poderia gerar dúvidas de interpretação. No exemplo original foi usado o termo *ano* como atributo de Livro; esse talvez não seja o mais indicado. Provavelmente o mais correto seria usar *ano de publicação* ou *data de publicação*.

Os mesmos problemas citados ocorrem em modelos OO. Para finalizar essa seção, vale a pena ressaltar a importância da criação de uma terminologia adequada na modelagem de todo o sistema. Essa ação trará maiores chances de se utilizar termos mais adequados e que possam ser compreendidos por toda a comunidade, facilitando a utilização e manutenção dos sistemas. Além disso, ferramentas de *software* poderiam analisar o conteúdo dos dados de maneira mais eficiente.

4.2. Ontologias

O termo ontologia possui diferentes significados, dependendo da área estudada. Em uma mesma área, como ocorre na CC, pode ocorrer variações em sua definição. Guarino (1997), *apud* Garcia (2002), define ontologia como "uma caracterização axiomática do significado do vocabulário lógico", já para Sowa (2000), *apud* Garcia (2002), a ontologia "define os tipos de coisas que existem em uma aplicação". A primeira definição parece levar em consideração o significado dos termos. Já a segunda parece levar em conta a composição de um domínio, e não o significado.

Uma definição mais apropriada à nossos propósitos é apresentada por Swatout e Tate (1999), *apud* Garcia (2002): "a ontologia é definida como um conjunto de preceitos e termos que podem ser usados para descrever alguma área de conhecimento, ou construir uma representação para o conhecimento".

Navega (2005) afirma que a CC tem buscado diversas técnicas de representação, compartilhamento e manipulação do conhecimento, tais como *Knowledge Interchange Format* (KIF) e *Knowledge Query and Manipulation Language* (KQML), no entanto, o maior impacto atualmente se refere ao uso de ontologias.

Em função do enorme avanço da Internet e, principalmente, pelo acumulo de conhecimento disponível, a criação de ontologias pode trazer maiores benefícios considerando-se esse ambiente, uma vez que agentes de *software* podem compartilhar informações. Lima e Carvalho (2005) apresentam alguns aspectos essenciais a serem considerados na criação de uma ontologia a ser usada em ambiente da Internet:

- A identificação do contexto de um termo: quando dois agentes trocam informação a respeito do termo “braço” é preciso diferenciar “braço” de um ser humano de um “braço” de sofá.
- O compartilhamento de definições: é preciso que um agente consiga estabelecer relações de equivalência como, por exemplo, entre os termos “veículo” e “carro”.

Segundo Souza e Alvarenga (2004), o objetivo maior de uma ontologia é “criar um vocabulário controlado para se trocarem informações entre os membros de uma comunidade, sejam eles humanos ou agentes inteligentes”. Em outras palavras, uma ontologia é um processo que permite organizar a informação e auxiliar sua recuperação. Para isso, define um vocabulário de termos para descrever uma determinada realidade. Esses termos são categorizados para representar um determinado domínio. Dessa forma, uma ontologia pode ser concebida como uma linguagem, pois estabelece um conjunto de termos que, posteriormente, poderão ser usados para formular consultas sobre uma base de conhecimento (ALMEIDA e BAX, 2003). Os usuários formulam perguntas baseadas em conceitos anteriormente especificados, isto é, definidos no momento da criação da ontologia.

A CC possui uma série de linguagens adequadas para a construção de ontologias. Almeida e Bax (2003) citam quinze dessas linguagens, descrevem os diversos tipos de ontologias existentes e também apresentam diversas ferramentas de *software* usadas na criação de ontologias.

Uma das principais linguagens usadas na construção de ontologias é OWL (*Web Ontology Language*). Através dela é possível definir um vocabulário de termos e as relações existentes entre esses termos, ou seja, criar uma ontologia. A OWL possui “capacidade semântica” para dar suporte a linguagens como XML,

RDF, e RDF *Schema* (W3C, 2004)

A ontologia representa uma alternativa para a organização da informação e representação de um domínio de conhecimento. Navega (2005) acredita que uma ontologia só faça sentido se estiver sendo proposta para uma comunidade específica, funcionando como um "acordo de cavalheiros". Por outro lado, é importante que existam mecanismos capazes de permitir o intercâmbio de dados entre comunidades diferentes.

Como a ontologia é armazenada em dispositivos computacionais, mecanismos de *software* podem exercer inferência sobre o conhecimento registrado, extraindo o máximo de informação e gerando novos conhecimentos. Em função disso, a ontologia pode contribuir para melhorar a recuperação da informação. Uma ontologia implementada em um sistema computacional pode processar o conteúdo armazenado e fazer deduções lógicas de acordo com as perguntas realizadas, facilitando não só a comunicação entre mecanismos de *software*, como também possibilitando a comunicação homem-máquina. Conforme Barquín *et al.* (2006), a livraria virtual Amazon utiliza uma ontologia para melhorar a venda de produtos.

Moreira (2003) diz que o termo ontologia é mais comum na CC (principalmente na subárea de representação do conhecimento) e é praticamente ignorado pela CI. Como uma das preocupações da CI se refere à representação do conhecimento, torná-se essencial estudar esse termo. Já para Barquín *et al.* (2006), tanto a CI como a CC têm buscado o desenvolvimento de ontologias de maneira interdisciplinar, uma vez que seu uso pode melhorar a recuperação de informação e facilitar a representação do conhecimento armazenado:

"Estamos assistindo a transição da Web tradicional para a Web

> Semântica, onde as ontologias contribuem soluções a alguns dos problemas causados pelo volume elevado e desordenado de informação, já que são capazes de interagir com o usuário respondendo a suas perguntas, ao mesmo tempo em que podem processar os documentos de forma inteligente, e são capazes de representar o conhecimento abarcando um conteúdo informativo muito grande".

Em função das fontes bibliográficas consultadas, principalmente pelos trabalhos de (Moreira, 2003), Freitas (2003) e Barquin *et al.* (2006), as ontologias representam o "estado da arte" em termos de representação do conhecimento: para a CI ela pode ser considerada como uma evolução dos tesauros, uma vez que permite melhor especificar as relações entre os termos; já para a CC, a ontologia atua de forma complementar a outras técnicas de representação do conhecimento, tais como redes semânticas e *frames*. Enquanto essas técnicas de representação atuam nos níveis lógico e epistemológico, as ontologias atuam em um nível superior, o ontológico. No entanto, nem todos os autores consideram esse fato, já que Souza *et al.* (2004) consideram que a ontologia também atua no nível epistemológico e não no ontológico.

Um importante setor onde as ontologias têm sido muito estudadas e utilizadas se refere à Web Semântica, item tratado a seguir.

4.3. Web Semântica

Antes de iniciar as descrições a respeito da Web Semântica, convém ressaltar que seu objetivo principal é organizar as informações na Internet com fins de recuperação, ou em outros termos, representar o conhecimento armazenado na Internet e torná-lo recuperável. A recuperação pode ser realizada tanto por seres humanos como por mecanismos de *software*, denominados agentes inteligentes (programas de computador que simulam algum

comportamento humano na resolução de tarefas).

O principal responsável pela criação da Web Semântica, o físico inglês Tim Berners-Lee, a considera como uma extensão da Internet atual. Na Web Semântica a informação passa a ter um significado bem definido. A pretensão é que a informação torne-se "*machine-readable*" (lida automaticamente), e também "*machine-understandable*" (entendida automaticamente). A estrutura da Web Semântica permite que as informações sejam armazenadas de maneira organizada e padronizada e, por conseguinte, recuperada (BERNERS-LEE, 2001).

Para cumprir o exposto, a Web Semântica define padrões para a descrição dos elementos que compõem os documentos a serem disponibilizados na Internet. Os documentos são elaborados a partir de determinadas regras, definidas através das linguagens XML e RDF, os alicerces da Web Semântica. Faria e Rosário (2002) descrevem a Web Semântica em uma arquitetura tecnológica de três camadas:

- **Camada de Schema**: contém as regras a serem seguidas na elaboração dos documentos (XML e RDF);
- **Camada de Ontologia**: responsável pela definição formal dos termos e os relacionamentos existentes entre eles – desenvolve uma taxonomia de conceitos para definição de classes e subclasses de objetos;
- **Camada Lógica**: define um mecanismo para realizar inferências sobre os dados. É composta por conjunto de regras de inferências que os agentes de *software* utilizarão para relacionar e processar informações. A manipulação dos objetos através do raciocínio lógico, a definição do significado do objeto e as relações entre eles, permitem a um agente computacional extrair conclusões, simulando o comportamento humano.

Souza e Alvarenga (2004) acreditam que a Web Semântica pode contribuir

para a melhoria dos índices de revocação e precisão:

> "Podemos esperar que a Web tenha grande melhoria dos índices de revocação e precisão no atendimento das necessidades de informação, porque a semântica embutida nos documentos permitirá aos dispositivos de recuperação evitar os problemas comuns de polissemia e sinonímia, além de considerar as informações em seus contextos de significado".

Outro ponto importante, citado em Souza e Alvarenga (2004), diz respeito às mudanças que a Web Semântica pode trazer para atividades pertinentes aos profissionais da CI:

- A melhoria dos motores de busca da Internet;
- O surgimento de interfaces do usuário mais intuitivas (gráficas e cognitivas);
- A construção automática de tesauros e vocabulários controlados;
- A indexação automática ou semi-automática de documentos.

Para que a Web Semântica atinja seus objetivos, isto é, crie um ambiente adequado para representação e recuperação de informação e conhecimento, torna-se necessário a presença de diversas "entidades", desde a criação dos documentos, sua disponibilização na Internet e posterior recuperação. A Figura 13 apresenta as principais "entidades" presentes na Web Semântica.

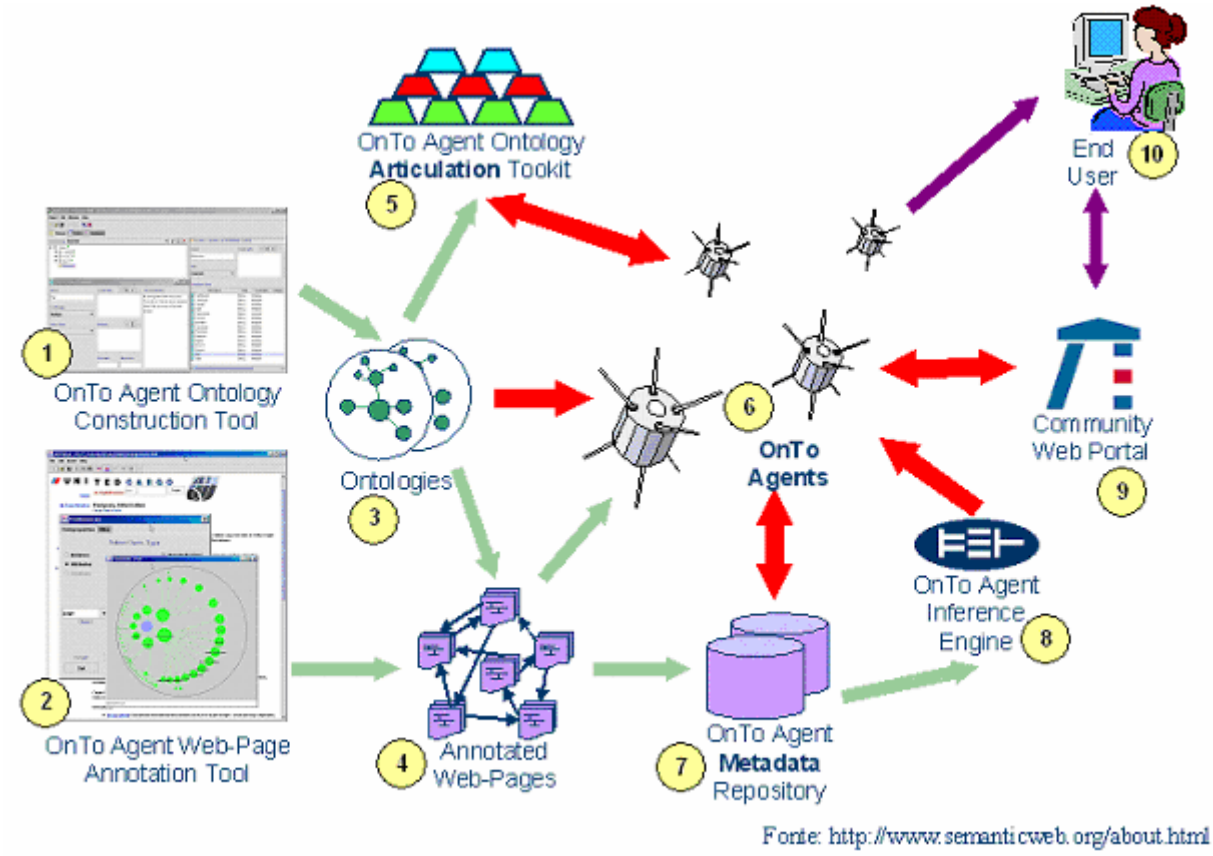

Figura 13. Principais entidades da Web Semântica.

As principais entidades envolvidas com a Web Semântica, apresentadas na Figura 13, já foram abordadas aqui. Através da definição da função que cada entidade exerce, é possível compreender como a Web Semântica se tornará um importante sistema de recuperação de informações.

As entidades citadas na Figura 13 são:

1. Ferramentas de *software* para a criação de ontologias.
2. Ferramentas de *software* para a criação de páginas marcadas semanticamente, isto é, suportam a criação de páginas em diversas linguagens de marcação, como HTML, XML e RDF.
3. As ontologias compartilhadas no sistema, produzidas por (1).
4. Páginas da Internet marcadas semanticamente, produzidas por (2).
5. Kit de Ferramentas de *software* para articulação de ontologias. Elas auxiliam os agentes (6) na interpretação de ontologias desconhecidas. Atuam como um mediador entre os agentes (6) e as ontologias (3).

6. Agentes inteligentes de *software* que interpretam as ontologias (3) através de (5) e atuam em conjunto com os mecanismos de inferência (8) para interpretarem o conteúdo das páginas (4). Com isso podem criar novos conhecimentos (novas páginas) e os armazenar em (7), além de fornecer informações relevantes a portais (9) e usuários (10).
7. Repositórios de metadados formados por (4), manipulados pelos agentes (6) e interpretados pelos mecanismos de inferência (8).
8. Mecanismos de inferência que auxiliam os agentes (6) na interpretação dos conteúdos das páginas marcadas semanticamente (4).
9. Portais comunitários de acesso a Web Semântica.
10. Usuário final que pode obter informações mais relevantes durante as pesquisas.

Conforme apresentado, a Web Semântica utiliza praticamente todos os itens tecnológicos e conceituais citados neste trabalho pertinentes à representação do conhecimento. Espera-se que esses mecanismos contribuam para tornar a Internet mais organizada e mais "inteligente".

Este capítulo descreveu algumas iniciativas em que se percebe a atuação das duas áreas estudadas de forma interdisciplinar. O próximo capítulo apresenta alguns pontos importantes a serem considerados na elaboração de uma representação, tratando sobre conflitos semânticos entre domínios diferentes e propondo uma estrutura de representação a partir de uma ontologia criada, denominada OntoArt.

5. PROPOSTAS DE REPRESENTAÇÃO

Conforme apresentado no decorrer deste trabalho, percebe-se que a representação do conhecimento possui diversas frentes, ou seja, existem diversas maneiras de representar um domínio de conhecimento.

Este capítulo apresenta um possível modelo para realizar a representação do conhecimento de artigos eletrônicos. Trata também de alguns problemas inerentes a integração entre domínios diferentes (domínios que representam o conhecimento de maneira diferente). Espera-se que o modelo sugerido sirva de base para estudos futuros.

A necessidade de prover interoperabilidade entre sistemas computacionais é cada vez mais evidente. Bibliotecas e provedores de informação podem agregar valor a seus serviços se dispuserem de sistemas que troquem dados de maneira automática.

5.1. Conflitos semânticos

Conforme citado anteriormente, a XML oferece um modelo eficiente para descrever metadados, permitindo a descrição não apenas da estrutura de um documento, como também dos próprios dados que transporta. Isso faz com que seja possível estabelecer aspectos semânticos aos dados transportados.

Para um computador, o texto “o livro custou 50,00” não quer dizer muita coisa. Fica muito difícil para um *software* reconhecer que 50,00 é o preço do livro, algo tão trivial para seres humanos. Conforme descrito anteriormente, através da XML é possível criar um marcador para delimitar o preço do livro, tal como: <preco_unitario> 50,00 </preco_unitario>. Além disso, torna-se possível também descrever se o valor está representado em reais, dólares, euros ou outra moeda qualquer. Esse aspecto poderia provocar um conflito semântico entre dois domínios: caso fosse necessário se pesquisar livros em diversos domínios para se descobrir os livros com custo inferior a 50,00, qual moeda deveria ser usada? Para que mecanismos possam interoperar é necessário se estudar e

compreender a semântica dos dados.

Conflitos semânticos são considerados aqui como conflitos decorrentes da discordância na descrição de um termo ou assunto, em domínios diferentes. Seja a descrição de um estado do Brasil, onde um artigo tenha sido publicado. Um determinado domínio pode representá-lo pelo nome completo (São Paulo, por exemplo), porém um outro domínio pode representá-lo apenas pela sigla (SP). Mesmo tendo representações diferentes para estado, um sistema de recuperação pode ter condições de recuperar a informação, independentemente do domínio analisado.

Conforme Abdalla (2003), os conflitos semânticos podem ser classificados em três tipos: expressões diferentes, unidades diferentes e faixas diferentes.

- **Expressões diferentes**: ocorre quando domínios diferentes possuem números equivalentes de elementos para representar um conceito (ou termo), mas com expressões diferentes. Conforme já citado, o estado de um país pode ser representado pelo nome completo ou pela sigla. De forma semelhante, o sexo de uma pessoa pode ser representado pela descrição completa (Masculino, Feminino) ou por apenas um caractere (M, F).
- **Diferentes unidades**: ocorre quando domínios diferentes possuem unidades diferentes para representar o mesmo conceito. Medidas de temperatura (celsius ou fahrenheit), de unidades monetárias (reais, dólar, euro, etc.), de peso (grama, quilo, arroba, etc.) e de velocidade (Km/h, Milhas/h, etc) são exemplos típicos. Com o mundo globalizado, se torna cada vez mais necessário se desenvolver mecanismos de conversão automática.
- **Diferentes faixas de valores**: ocorre quando domínios diferentes possuem uma faixa de valores diferente para representar o mesmo conceito. Por exemplo, um determinado domínio A pode usar os valores A, B, C, D e E para mensurar as notas de alunos. Outro domínio B pode

utilizar os valores Insuficiente, Bom e Muito Bom. Não apenas a descrição é diferente, mas também a quantidade de elementos usados para descrever um conceito. Enquanto o domínio A contém cinco variações possíveis, o domínio B possui apenas três.

Para reduzir os efeitos dos conflitos semânticos pode ser criada uma estrutura em XML que represente, para cada domínio específico, o significado de seus elementos. Um determinado domínio A pode representar os elementos sexo e peso usando a estrutura de representação apresentada no Quadro 12.

Quadro 12. Representação no domínio A.

```
<elemento>
  <nome>sexo</nome>
  <valor>Masculino</valor>
  <valor>Feminino</valor>
</elemento>
<elemento>
  <nome
unidade="Real">preço</nome>
  <valor>10.00</valor>
</elemento>
```

Essa estrutura está representando que o elemento “sexo” poderá assumir os valores: “Masculino” ou “Feminino”. Da mesma forma, o elemento “preço” está sendo representado na moeda Real (“R$”). Um outro domínio B pode representar os mesmos elementos sexo e preço de maneira diferente, conforme mostra o Quadro 13.

Quadro 13. Representação no domínio B.

```
<elemento>
    <nome>sexo</nome>
    <valor>M</valor>
    <valor>F</valor>
</elemento>
<elemento>
    <nome
unidade="Dolar">preço</nome>
    <valor>4.35</valor>
</elemento>
```

Apesar de os conteúdos dos elementos serem diferentes nos domínios A e B, eles são semanticamente iguais, apenas a forma de representação é diferente.

Sendo assim, é possível realizar o mapeamento entre domínios diferentes usando documentos criados a partir da XML. Isso pode ajudar a reduzir os efeitos dos conflitos semânticos.

5.2. Enriquecimento do conteúdo de artigos

Marcondes *et al.* (2005), consideram importante que dados factuais de uma publicação eletrônica, tais como: formulação do problema, metodologia utilizada, hipóteses e conclusões, elementos essenciais em publicações científicas, possam ser inseridos em publicações eletrônicas e, por conseguinte, capturados e validados de forma automática. Além de melhorar o processo de comunicação científica, isso poderia validar o conteúdo de um artigo, comparando-o com ontologias já existentes e consolidadas em uma determinada área do conhecimento.

Além disso, uma publicação poderia fazer relações a outras publicações existentes em uma base de dados pública, incluindo elementos do tipo “se baseia em”, “contesta”, “referenda”. Ainda segundo Marcondes *et al.* (2005) “as publicações científicas eletrônicas, os periódicos eletrônicos, são ainda hoje fortemente calcadas no modelo periódico em papel”. Esse fato restringe a potencialidade do meio digital na Internet. A proposta deste trabalho visa complementar a estrutura proposta por Marcondes *et al.* (2005), adicionando elementos contidos no padrão *Dublin Core*, além de outras recomendações.

5.3. Ontologia para representação de artigos

A criação de uma ontologia requer esforço conjunto de profissionais das diversas áreas envolvidas em um determinado domínio. Uma iniciativa importante no que diz respeito a publicações eletrônicas se refere à ontologia *Portal Ontology* (LAUSEN e STOLLBERG, 2004). Essa ontologia contempla a definição de 16 subconceitos referentes a publicações eletrônicas, dentre eles o conceito de artigo (Article), conforme apresenta a Figura 14. Todos os conceitos se referem a um tipo de publicação.

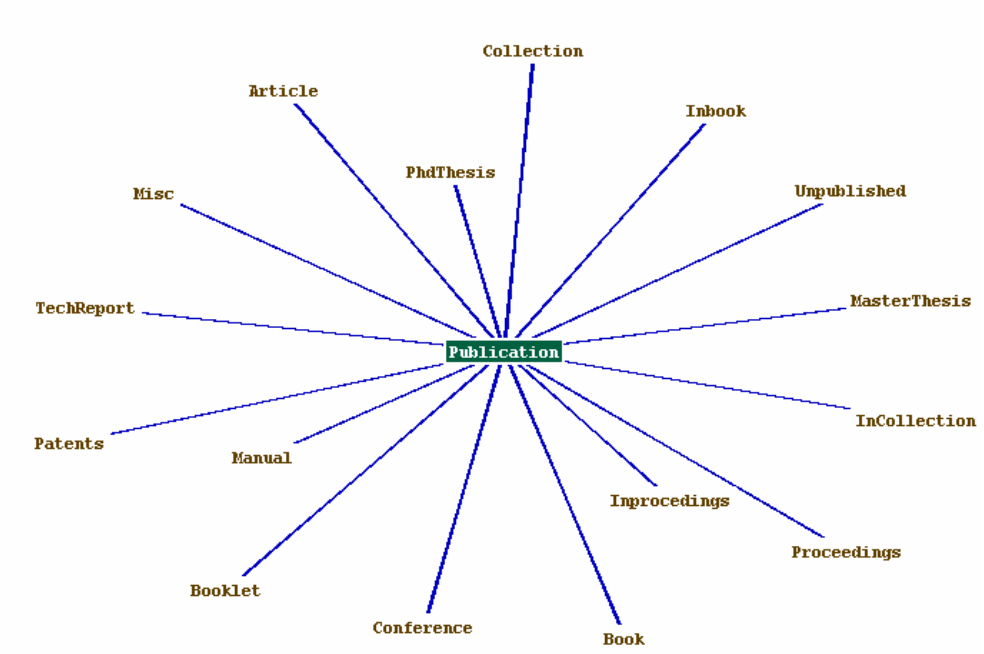

Figura 14. Conceitos referentes a publicações, baseado em Lausen e Stollberg (2004).

A ontologia proposta neste trabalho, denominada OntoArt, se concentra apenas nos conceitos inerentes a representação de artigos. Não há uma metodologia ideal para o desenvolvimento de ontologias, trata-se de um processo interativo de sucessivos refinamentos e detalhamentos. De forma geral, para se criar uma ontologia é necessário: a) definir classes e subclasses presentes na ontologia, b) definir *slots* e sua faixa de valores permitida e c) criar instâncias para classes e *slots*.

A ontologia foi desenvolvida com o auxílio da ferramenta Protégé, versão 3.1.1 (HORRIDGE *et al.*, 2004). Um tutorial que descreve o processo de instalação do Protégé, juntamente com alguns detalhes de criação da ontologia proposta, pode ser consultado no CD anexo a dissertação. A ontologia completa também se encontra no CD. A Figura 15 apresenta uma visão simplificada da ontologia .

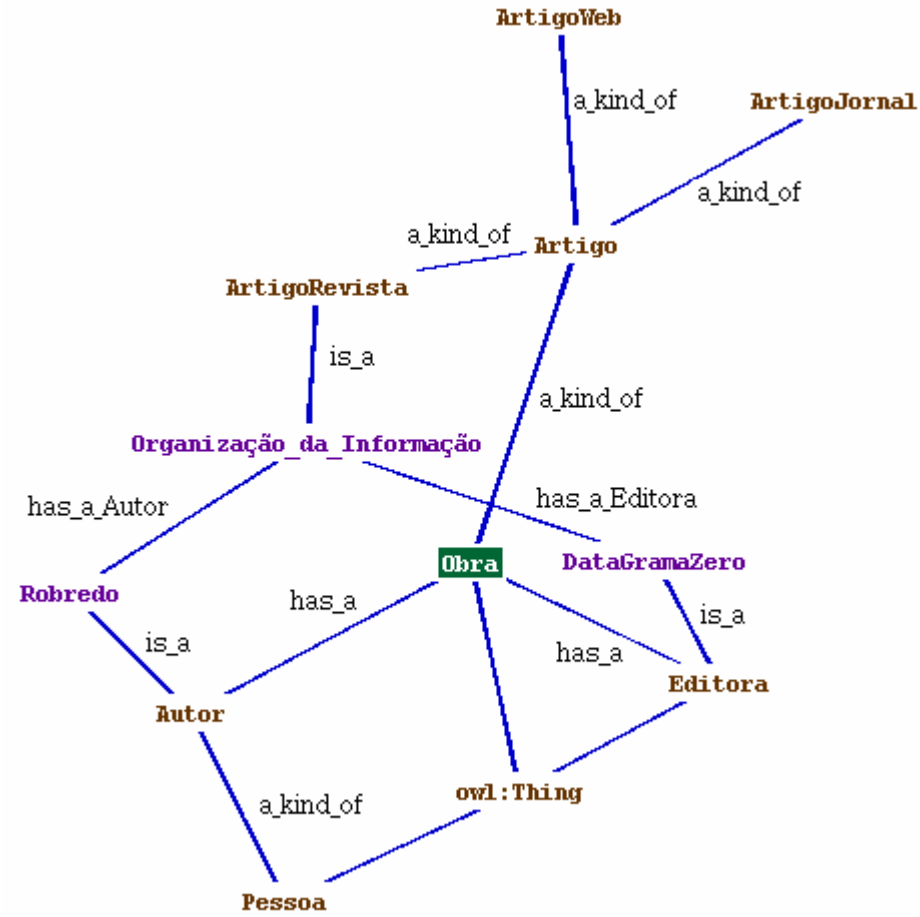

Figura 15. Visão simplificada da OntoArt.

Pela ontologia criada, exposta no diagrama, um agente de *software* pode realizar algumas inferências, tais como:

- Um artigo pode ser de três tipos diferentes, isto é, pode ser publicado em três meios diferentes: revista, jornal ou site da Internet.
- Artigo é um tipo de obra.
- Uma Obra contém Autor e Editora.

- Um Autor é um tipo de Pessoa.
- Robredo é um Autor, isto é, uma Pessoa.
- O Autor do Artigo de Revista chamado Organização da Informação é Robredo.
- O Artigo Organização da Informação foi publicado pela Revista DataGramaZero.

Apesar de não exposto no diagrama OntoArt da Figura 15, o artigo contempla uma série de propriedades (características) que permitem identificá-lo e relacioná-lo com outros artigos. As propriedades definidas para um artigo são apresentadas na Figura 16.

Como pode ser observado pela Figura 16, cada artigo possui uma série de propriedades (metadados) que podem ser usadas para melhor descrevê-lo. As propriedades que possuem as iniciais “ma” foram definidas com base em Marcondes *et al.* (2005), as com “dc” seguem o padrão *Dublin Core*, a com “so” com base em Souza *et al.* (2000); já as que iniciam com “sf” são sugeridas pelo autor deste trabalho.

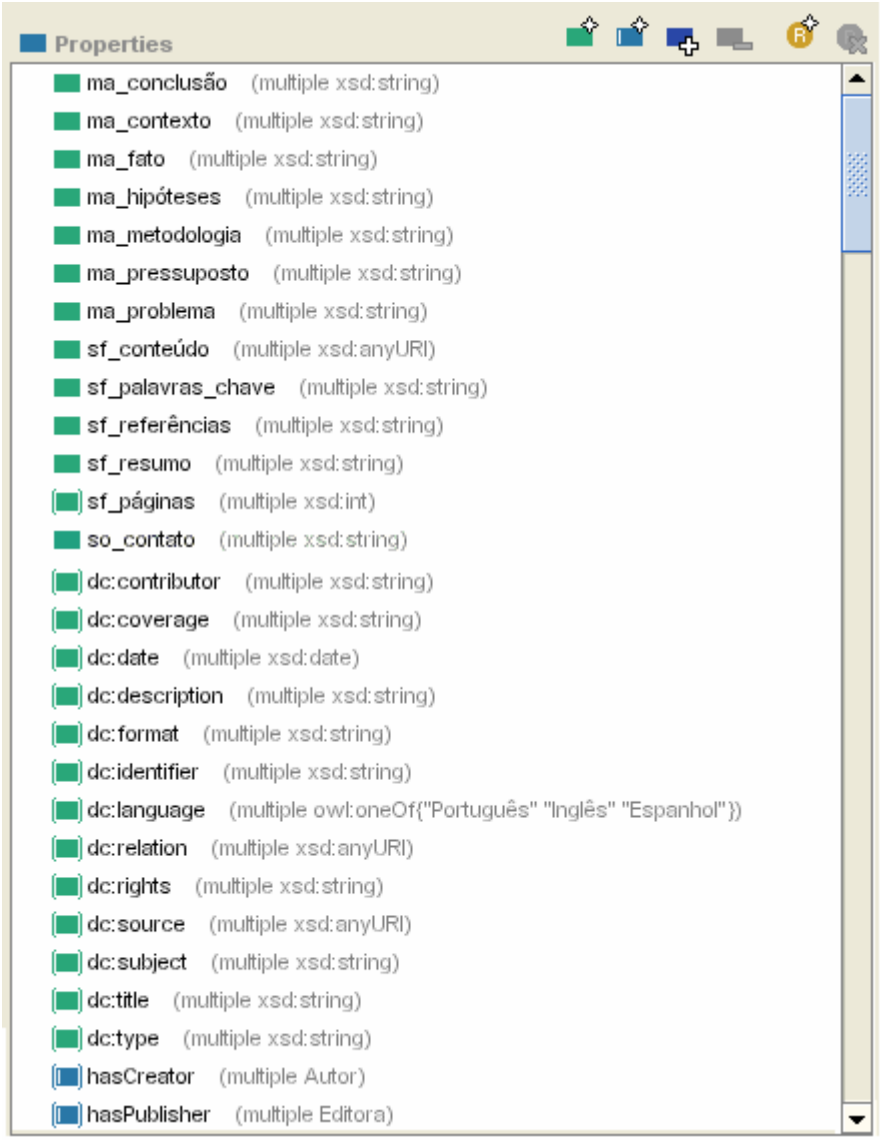

Figura 16. Propriedades definidas para um artigo.

5.4. Estrutura em XML para representação de artigos

As propriedades definidas pela ontologia podem ser usadas (mapeadas) na criação de um arquivo XML. Com isso será possível manter (*on-line*) metadados referentes aos artigos publicados em um site. Uma breve descrição a respeito de cada propriedade encontra-se na Tabela 6.

Tabela 6. Conjunto de propriedades usadas na descrição de artigos.

Propriedade	Descrição
ma_conclusão	Conclusões alcançadas pelo artigo.
ma_contexto	O contexto abordado no artigo, sua abrangência e limitações.
ma_fato	O novo fenômeno, ou descoberta, tratado no artigo. .
ma_hipoteses	Respostas provisórias ao problema proposto.
ma_metodologia	A metodologia usada na realização do artigo.
ma_pressuposto	Os pontos de partida e as limitações da pesquisa presente no artigo.
ma_problema	As questões a serem resolvidas pela pesquisa.
sf_conteudo	Uma referência (endereço web) ao conteúdo da Obra.
sf_palavras_chave	O conjunto de palavras-chave do artigo
sf_referências	As referências bibliográficas usadas na redação do artigo.
sf_resumo	O conteúdo do resumo do artigo.
sf_páginas	O número de páginas que o artigo possui.
so_contato	Indicação para contato, tipicamente o email do responsável.
dc:contributor	Uma entidade que contribuiu com o conteúdo do artigo.
dc:coverage	Define a abrangência do conteúdo do artigo, tipicamente um período de datas ou uma região.
dc:creator	A entidade responsável pela criação do artigo.
dc:date	Uma data associada a publicação do artigo.
dc:description	Uma breve descrição do artigo.
dc:format	Define o tipo da mídia ou dimensões do artigo.
dc:identifier	Contém um identificador único para o artigo (ex. ISBN).
dc:language	A linguagem usada no conteúdo do artigo.
dc:relation	Uma (ou várias) referência a um recurso relacionado.
dc:rights	A entidade que possui direitos autorais sobre o artigo.
dc:source	Uma referência ao local onde o artigo se localiza.
dc:subject	Uma pequena descrição a respeito do assunto do artigo.
dc:title	O título do artigo.
dc:type	Define a natureza ou gênero em que o artigo está inserido (texto, imagem, som, simulação, etc.).
hasCreator	A entidade ou pessoa responsável pela criação do recurso. Podem existir diversos, entretanto todos devem ser uma instância de autor.
hasPublisher	A entidade responsável pela publicação do artigo.

Evidentemente, nem todas as propriedades são obrigatórias para um mesmo artigo. Dependendo das necessidades de informação, algumas serão essenciais, enquanto que outras, opcionais. Por essa razão, o conjunto de propriedades obrigatórias e opcionais não foi aqui definido. Essas propriedades podem ser analisadas por um mecanismo de busca com o objetivo de retornar resultados mais relevantes. A listagem seguinte apresenta as propriedades mapeadas para *tags* em formato XML. A listagem do Quadro 14 constitui-se numa versão resumida da estrutura necessária para se manter metadados sobre os artigos. A idéia é que para cada artigo publicado no site, existam metadados

correspondentes em formato XML.

Quadro 14. Tags para criação de metadados.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<estrutura type="artigo">
  <dc:creator>                    </dc:creator>
  <dc:contributor>                </dc:contributor>
  <dc:publisher>                  </dc:publisher>
  <dc:subject>                    </dc:subject>
  <dc:description>                </dc:description>
  <dc:identifier>                 </dc:identifier>
  <dc:relation>                   </dc:relation>
  <dc:source>                     </dc:source>
  <dc:rights>                     </dc:rights>
  <dc:format>                     </dc:format>
  <dc:type>                       </dc:type>
  <dc:title>                      </dc:title>
  <dc:date>                       </dc:date>
  <dc:coverage>                   </dc:coverage>
  <dc:language>                   </dc:language>
  <ma:fato>                       </ma:fato>
  <ma:problema>                   </ma:problema>
  <ma:metodologia>                </ma:metodologia>
  <ma:pressuposto>                </ma:pressuposto>
  <ma:contexto>                   </ma:contexto>
  <ma:hipotese>                   </ma:hipotese>
  <ma:conclusao>                  </ma:conclusao>
  <sf:conteudo>                   </sf:conteudo>
  <sf:palavras_chave>             </sf:palavras_chave>
  <sf:referencia>                 </sf:referencia>
  <sf:resumo>                     </sf:resumo>
  <sf:paginas>                    </sf:paginas>
  <so:contato>                    </so:contato>
</estrutura>
```

A estrutura leva em consideração os enunciados de Lourenço (2005), uma vez que as propriedades geram metadados descritivos (que descrevem um recurso), estruturais (suportam *links* entre recursos) e administrativos (permitem controle sobre as publicações). Se todos os artigos de uma base de dados forem elaborados a partir da estrutura do Quadro 14 e disponibilizados na Internet, uma ferramenta de *software* poderá ser usada para pesquisar e recuperar informação de muitas maneiras diferentes. A seguir são listadas algumas buscas possíveis:

- Procure no título do artigo a ocorrência da palavra “marcação”;
- Procure nas palavras-chave a ocorrência da palavra XML;
- Procure no resumo a ocorrência da palavra HTML;
- Retorne todos os artigos que contêm o autor “Bush” nas referências;

- Retorne todos os artigos cuja metodologia usada seja “estudo de caso”;
- Retorne todos os artigos em Português, cujo resumo contenha as palavras “representação do conhecimento”.

Os estudos realizados podem servir de base para a elaboração de um *software* editor de artigos eletrônicos. Esse editor poderia permitir a digitação e publicação dos artigos em XML, conforme a estrutura proposta. Todos os metadados sobre o artigo poderiam ser preenchidos no próprio editor, ou seja, uma só ferramenta poderia ser usada para a edição e documentação dos artigos, isso tornaria a documentação praticamente concomitante com a redação do artigo. Essa ferramenta poderia também auxiliar na melhoria da qualidade dos metadados como, por exemplo, prever erros tipográficos e outros problemas citados por Beall (2005). O diagrama da Figura 17 apresenta a infra-estrutura mínima necessária para se criar um ambiente Internet mais adequado à publicação e recuperação de informações em artigos eletrônicos, conforme a proposta apresentada.

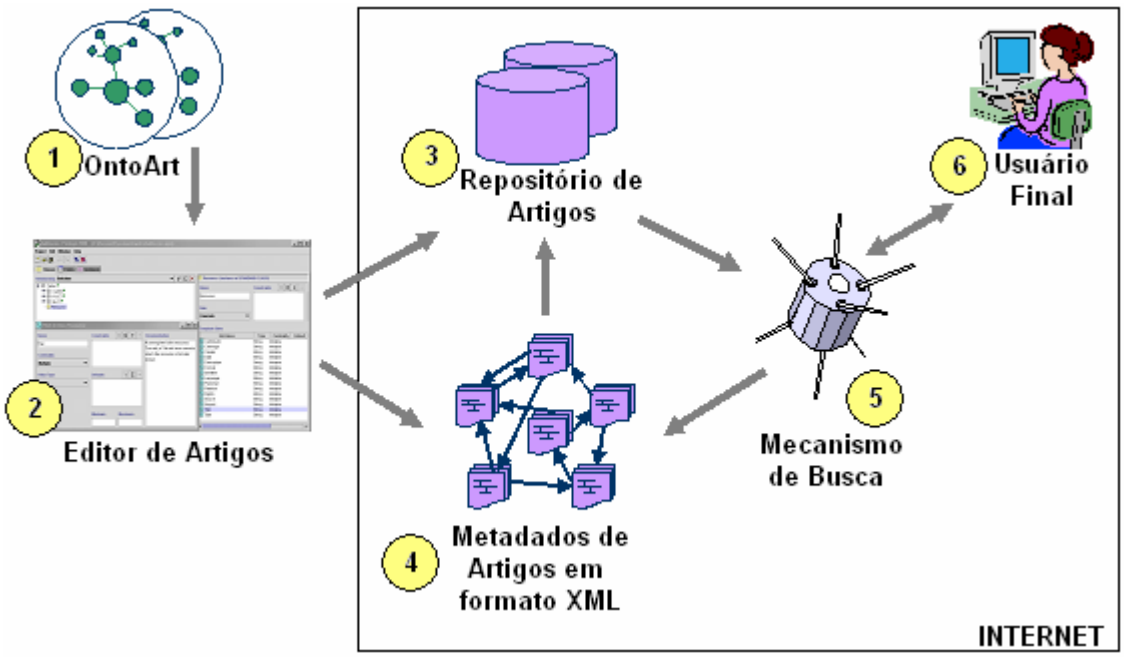

Figura 17. Infra-estrutura para publicação e busca em artigos eletrônicos.

Os principais atores envolvidos no processo, apresentados na Figura 17, são:

1. A ontologia proposta, denominada OntoArt, elaborada a partir da ferramenta Protégé. Ela serve de base para a estrutura a ser gerada pelo Editor de artigos (2).
2. Editor de artigos, uma ferramenta de *software* que pode ser elaborada para a redação de artigos. Ao invés de utilizar editores tradicionais (como Word), o autor de um artigo utilizaria essa ferramenta para redigir seu artigo. O editor suportaria tanto a redação do artigo, quanto sua documentação (as propriedades definidas na OntoArt). Depois de publicado, cada artigo seria armazenado em um repositório (3) disponível a partir da Internet.
3. Os artigos produzidos por (2), assim como seus metadados (4), ficariam armazenados em repositórios disponíveis na Internet.
4. Os metadados elaborados a partir do editor (2) que representam o conhecimento armazenado nos artigos.
5. Mecanismo de busca responsável por receber as solicitações dos usuários e retornar *links* para os artigos relevantes (da mesma forma que Google).
6. Usuário que realiza as consultas em um ambiente Internet acessando o mecanismo de busca (5).

A infra-estrutura apresentada está baseada nos preceitos da Web Semântica e, conforme citado, representa apenas o mínimo necessário para criação de um ambiente adequado para a representação e recuperação de artigos eletrônicos. Estudos futuros podem complementar a estrutura proposta, estabelecendo relações semânticas com RDF e vocabulários controlados com *namespaces*.

A estrutura proposta mantém o conteúdo do artigo separado das partes documentais (resumo, conclusão, hipóteses, etc). Apesar de estarem em locais distintos, os recursos tecnológicos disponíveis permitiriam que sua visualização (em tela) ocorresse de maneira conjunta, tornando esse fato transparente ao

usuário.

Este capítulo apresentou a proposta de uma estrutura de representação, objetivo maior deste trabalho. O próximo capítulo encerra as discussões, apresentando as conclusões alcançadas.

6. CONCLUSÃO

O presente trabalho buscou descrever as principais características de cada uma das formas de representação do conhecimento, realizando uma análise comparativa entre elas, tanto por parte da CI quanto da CC. Apesar de ter sido realizada uma busca exaustiva em referências bibliográficas, nem todas as formas de representação foram contempladas, apenas as mais relevantes.

Pelos aspectos estudados, observou-se que tanto a CI quanto a CC buscam representar o conhecimento de maneiras similares, isto é, buscam estabelecer conceitos (gerais e específicos) e seus relacionamentos, dividindo o conhecimento em classes e subclasses. No entanto, cada área está focada em um público alvo diferente: enquanto a CI busca representar o conhecimento para auxiliar seres humanos na busca por informações, a CC o faz para utilização de agentes de *software*, objetivando que a interpretação se torne um processo automático. Isso pode ser observado na criação de tesauros (originários da CI) e nas ontologias (mais presentes na CC).

Para Moreira *et al.* (2004), a maioria dos processos de representação e recuperação de informação ainda se baseia em aspectos sintáticos e estatísticos, considerando a freqüência e distribuição de palavras presentes em documentos. Por essa razão, esses processos não são eficazes quanto à recuperação. Para tentar minimizar esse problema, a Web Semântica busca criar ontologias para permitir a classificação automática de documentos, baseada nos aspectos semânticos das linguagens usadas.

O avanço da Web Semântica pode contribuir para o surgimento de mecanismos de busca que forneçam resultados mais relevantes. Isso também

causará impacto sobre o profissional da informação, modificando suas atividades cotidianas. Ferramentas de *software*, cada vez mais sofisticadas, exigirão uma contínua especialização por parte dos profissionais da CI. Atividades mecânicas, e até mesmo intelectuais, poderão ser substituídas por sistemas cada vez mais capacitados.

Pelos estudos realizados referentes à representação do conhecimento, observa-se que a CC passa a se preocupar cada vez mais com o aspecto semântico da informação. Nota-se que diversas pesquisas estão sendo realizadas a esse respeito, fazendo com que a CC se aproxime cada vez mais da CI.

Em função das novas tecnologias relacionadas à representação e recuperação, torna-se cada vez mais difícil conceber a CI como uma ciência isolada. Nota-se que os preceitos disponíveis na CI estão sendo incorporados e implementados na CC. Observa-se também que as duas áreas são interdisciplinares e muito próximas, não apenas, mas principalmente, no que tange a representação do conhecimento. Por essa convergência das duas áreas, e pela evolução natural do conhecimento, é possível conceber que a CI possa vir a ser uma subárea da CC.

As referências analisadas ajudaram a constatar que a CC poderia desenvolver uma terminologia para aumentar a semântica de diversos modelos, tais como o modelo E-R e o modelo OO. Isso forneceria maior semântica aos modelos, auxiliando humanos e máquinas na interpretação de conteúdos. Esforços nesse sentido podem ser notados na tentativa de se estabelecer vocabulários controlados a partir de *namespaces* com a linguagem RDF.

Apesar de a estrutura proposta neste trabalho não atuar como um tesauro,

isto é, não ter a estrutura mais adequada para ajudar o usuário a aprender sozinho, o estabelecimento de relações entre os artigos pode auxiliar o usuário a entender a evolução do conhecimento. Além disso, é possível conceber que o relacionamento entre os artigos suportado pela estrutura proposta, contribua para aumentar o índice de revocação. Apesar de a enorme quantidade de artigos disponíveis na Internet, é exceção o estabelecimento de relações entre eles, salvo as citações e alguns mecanismos de busca, como Google.

Espera-se que a estrutura proposta possa contribuir para o desenvolvimento de sistemas de recuperação mais eficientes. Muitas universidades poderiam disponibilizar monografias, dissertações, teses e outros materiais, utilizando a estrutura proposta. Alunos dessas universidades poderiam ter a disposição ferramentas que auxiliassem a busca de trabalhos anteriores, realizados por alunos da própria instituição. Em grande parte das universidades, os trabalhos dos alunos ficam praticamente esquecidos, armazenados apenas em formato físico, isto é, são armazenados em um formato em que as pesquisas tornam-se inviáveis.

A estrutura proposta pode auxiliar não apenas seres humanos a recuperar informações de maneira mais efetiva, mas também ser um ponto de partida para que mecanismos de *software* realizem inferências sobre os artigos, baseados na ontologia proposta. Essa característica une o melhor das duas áreas: o auxílio aos seres humanos na recuperação de informações (CI) e o suporte ao tratamento automático (CC). Por esse motivo, acredita-se que este trabalho atingiu os objetivos propostos, a saber, pesquisar e propor uma estrutura de representação para possibilitar a recuperação de informação e conhecimento de maneira mais efetiva, contemplando o melhor das duas áreas.

Para comprovar as conclusões aqui expostas, torna-se necessário desenvolver o sistema proposto de maneira prática. Espera-se que algum pesquisador, principalmente da área da CC, possa implementar a estrutura apresentada.

7. REFERÊNCIAS BIBLIOGRÁFICAS

ABDALLA, K. F. *A Model For Semantic Interoperability Using Xml*. The Technical Resource Connection, Perot Systems, 2003. Disponível em: <http://www.sys.virginia.edu/sieds03/proceed2003/proceedings/B102.pdf>. Acesso em: 14 jan. 2006.

ABREU, M. Ontologias, metadados semânticos e a Web. *Revista dados e Negócios - On-line,* edição 12, mai. 2004. Disponível em: http://www.dadosenegocios.com.br/ontologias_metadados_semanticos_e_a_web_11.html>. Acesso em: 12 fev. 2006.

AGÜERA, J. R. P. Automatización de Tesauros y su utilización en la Web Semântica. 2004. Disponível em:<http://eprints.rclis.org/archive/00004176/01/automatizacion.pdf>. Acesso em: 16 mai. 2005.

ALMEIDA, C. C. *O campo da ciência da informação: suas representações no discurso coletivo dos pesquisadores do campo no Brasil*. Florianópolis, 2005. 395 p. Dissertação (Mestrado em Ciência da Informação) - Universidade Federal de Santa Catarina - UFSC, Florianópolis, 2005. Disponível em: <http://150.162.90.250/teses/PCIN0003.pdf>. Acesso em: 03 jan. 2006.

ALMEIDA, M. B. Uma introdução ao XML, sua utilização na Internet e alguns conceitos complementares. *Ciência da Informação*, Brasília, v. 31, n. 2, p. 5-13, maio/ago.2002. Disponível em: <http://www.eci.ufmg.br/mba/text/art_xml_sub1_WEB.pdf>. Acesso em: 20 dez. 2005.

ALMEIDA, M. B.; BAX, M. P. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. *Ciência da Informação*, Brasília DF, v. 32, n. 3, p. 7-20, 2003. Disponível em: <http://www.scielo.br/pdf/ci/v32n3/19019.pdf>. Acesso em: 30 out. 2005.

ALVARENGA, L. *Representação do conhecimento em tempo e espaço digitais*. Encontros bibli, Florianópolis, 2003. Disponível em: <http://www.encontros-bibli.ufsc.br/Edicao_15/alvarenga_representacao.pdf>. Acesso em: 05 jan. 2006.

ARANHA, M. L.; MARTINS, M. H. *Filosofando*: *introdução à filosofia*. 2.ed. São Paulo: Moderna, 1993. 395 p.

ÁVILA, B. C. *Representação do Conhecimento Utilizando Frames*. São Carlos, 1991. 102f. Dissertação (Mestrado em Ciências de Comunicação e Matemática Computacional). Instituto de Ciências Matemáticas de São Carlos, USP - Universidade de São Paulo, São Carlos, 1991. Disponível em: <http://www.ppgia.pucpr.br/pesquisa/mining/dissertacoes.htm>. Acesso em: 03 nov. 2005.

BARQUÍN, B. A. R. *et al.* Construção de uma ontologia para sistemas de informação empresarial para a área de Telecomunicações. *DataGramaZero: Revista de Ciência da Informação*, Internet, v. 7, n. 2, abr/2006. Disponível em: <http://www.dgz.org.br/abr06/Art_04.htm>. Acesso em: 17 abr. 2006.

BEALL, J. Metadata and data quality problems in the digital library. *Journal of digital information*, vol. 6, issue 3, 2005. Disponível em: <http://jodi.tamu.edu/Articles/v06/i03/Beall/>. Acesso em: 17 mar. 2006.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web, *Scientific American*, May 2001. Disponível em: <http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2>. Acesso em: 20 jun. 2006.

BOHMERWALD, P. Uma proposta metodológica para avaliação de bibliotecas digitais: usabilidade e comportamento de busca por informação na Biblioteca Digital da Puc-Minas. *Ciência da Informação*, Vol. 34, N° 1 (2005). Disponível em: <http://www.ibict.br/cienciadainformacao/include/getdoc.php?id=1403&article=692&mode=pdf>. Acesso em: 03 jan. 2006.

BUCKLAND, M. K. Information as thing. *Journal of the American Society for Information Science*. (JASIS), v.45, n.5, p.351-360, 1991. Disponível em: <http://www.sims.berkeley.edu/~buckland/thing.html>. Acesso em: 03 fev. 2006.

BURKE, P. *Uma história social do conhecimento: de Gutemberg a Diderot*. Rio de Janeiro: Jorge Zahar, 2003. 241 p.

BUSH, V. *As we may thing*. Atlantic Monthly, p.101-108 (1945). Disponível em: <http://www.ps.uni-sb.de/~duchier/pub/vbush/vbush-all.shtml>. Acesso em: 14 jun. 2005.

CAMPOS, M. L. A. Modelização de Domínios de Conhecimento: uma investigação de princípios fundamentais. *Ciência da Informação*, Brasília, v. 33, n. 1, p. 22-32, 2004.

CAMPOS, M. L. A. Perspectivas para o estudo da área de representação da informação. *Ciência da Informação*, Brasília, v.25, n. 2, 1995.

CAMPOS, M. L. A.; GOMES, H. E. *Elaboração do Tesauro Documentário*: *tutorial*. Disponível em: <http://www.conexaorio.com/biti/tesauro/>. Acesso em: 06 mar. 2006.

CAPES, Comite Multidisciplinar. *Avaliação e Perspectivas*, 2003. Disponível em: <http://www.capes.gov.br/capes/portal/conteudo/MultidisciplinarDoc_Area2003_18jul03.pdf>. Acesso em: 05 mar. 2006.

CÉNDON, B. V. Ferramentas de busca na WEB. *Ciência da Informação*, Brasília, v. 30, n. 1, p. 39-49, 2001. Disponível em: <http://www.scielo.br/pdf/ci/v30n1/a06v30n1.pdf>. Acesso em: 30 jan. 2006.

CORDEIRO, R, I, N**.** Informação Cinematográfica e Textual: da geração à interpretação e representação de imagem e texto*. Ciência da Informação*, Brasília, v. 25, n. 3, p. 461-465, 1996.

DAHLBERG, I. *Teoria da classificação, ontem e hoje*. Mainz Universitat - Alemanha, 1972. Tradução de Henry B. Cox. Disponível em: <http://www.conexaorio.com/biti/dahlbergteoria/dahlberg_teoria.htm>. Acesso em: 29 nov. 2005.

DAVIES, J. I. *Glossary the Terms relevant to Mobile Communications*. Disponível em: <http://homepages.nildram.co.uk/~jidlaw/pages/glossary.html>. 2004. Acesso em: 10 jan. 2006.

DAVIS, M.; Walter, M. Technology Strategies: Next-Wave Publishing. *Technology: Revolutions in Process and Content*. Part 1. Vol. 3, No 23. The Seybold Report, 2003.

DCMI, 2006a. Dublin Core Metadada Initiative. *About the Initiative*. Disponível em: <http://dublincore.org/about/>. Acesso em: 21 fev. 2006.

DCMI, 2006b. Dublin Core Metadada Initiative. *Using Dublin Core*. Disponível em: <http://dublincore.org/documents/usageguide/#whatismetadata>. Acesso em: 21 fev. 2006.

DCMI, 2006c. Dublin Core Metadada Initiative. *Using Dublin Core - the elements*. Disponível em: <http://dublincore.org/documents/usageguide/elements.shtml>. Acesso em: 21 fev. 2006.

DESIGNERZ, web site. *XML Markup Applications*. Disponível em: <http://xml.designerz.com/xml-markup-applications.php>. Acesso em: 18 mai. 2005.

DIAS, E. W. Contexto Digital e Tratamento da Informação. *DataGramaZero: Revista de Ciência da Informação*, Brasília, v.2, n.5, out /2001.

FARIA, C. G.; GIRARDI, R. *Uma análise da Web Semântica e suas implicações no acesso à informação*. UFMA - Universidade Federal do Maranhão. 2002. Disponível em: <http://maae.deinf.ufma.br/orientacoes/orientacao.php?tp=3&sta=1>. Acesso em: 09 abr. 2006.

FERNEDA, E. *Recuperação de Informação: Análise sobre a contribuição da Ciência da Computação para a Ciência da Informação*. São Paulo, 2003. 137f. Tese (Doutorado em Ciências da Comunicação). Escola de Comunicação e Artes, USP - Universidade de São Paulo, São Paulo, 2003. Disponível em: <http://www.teses.usp.br/teses/disponiveis/27/27143/tde-15032004-130230/>. Acesso em: 04 out. 2005.

FREITAS, F. Ontologias e a Web Semântica. In: Renata Vieira; Fernando Osório. (Org.). *Anais do XXIII Congresso da Sociedade Brasileira de Computação*. Volume 8: Jornada de Mini-Cursos em Inteligência Artificial. Campinas: SBC, 2003, v. 8, p. 1-52.

GARCIA, S. C. *O uso de Árvores de Decisão na descoberta de conhecimento na Área da Saúde*. Semana Acadêmicas, 2000. Universidade Federal do Rio Grande do Sul. Disponível em: <http://www.inf.ufrgs.br/pos/SemanaAcademica/Semana2000/SimoneGarcia/>. Acesso em: 15 jun. 2006.

GARCIA, S. S. *Extensões do RDF para Representação do Conhecimento*. 2002.

37f Tese (Doutorado em Sistemas de Computação), UFRJ - Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2002. Disponível em: <http://genesis.nce.ufrj.br/dataware/tebd20021/seminarios/RDF.htm>. Acesso em: 19 set. 2003.

GOMES, H. E. *Classificação, tesauro e terminologia: fundamentos comuns*. Unirio, 1996. Disponível em: <http://www.conexaorio.com/biti/tertulia/ tertulia.htm>. Acesso em: 25 nov. 2005.

GOMES, H. E.; MARINHO, M. T. *Introdução ao Estudo do Cabeçalho de Assunto*. BITI - Biblioteconomia, Informação & Tecnologia da Informação. 1984. Disponível em: <http://www.conexaorio.com/biti/cabecalho/cab_ ass.htm>. Acesso em: 15 jan. 2006.

GRAVES, M. *Projeto de banco de dados em XML*. São Paulo: Makron Books, 2003.

HOLZNER, S. *Desvendando XML*. Tradução Daniel Vieira. Rio de Janeiro: Campus, 2001. 858p.

HORRIDGE, M. *et al. A Practical Guide To Building OWL Ontologies Using The Protégé - OWL Plugin and CO-ODE Tools Edition 1.0*. University of Manchester . 2004. Disponível em: <http://www.coode.org/resources/tutorials/ProtegeOWL Tutorial.pdf>. Acesso em: 15 jun. 2006.

JESUS, J. B. M. Tesauro: Um Instrumento de Representação do Conhecimento em Sistemas de Recuperação da Informação. In: *XII SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS*, Recife, 2002.

KAULA, P. N. *Repensando os conceitos no estudo da classificação*. Herald of Library Science, vol. 23, n. 2, jan./apr. 1984, p. 30-44. Disponível em: <http://www.conexaorio.com/biti/kaula/>. Acesso em: 28 nov. 2005.

LARA, M. L. G. O processo de construção da informação documentária e o processo de conhecimento. *Perspectivas em Ciência da Informação*, Belo Horizonte, v. 7, n. 2, p. 127-140, 2002.

LARA, M. L. G. O unicórnio (o rinoceronte, o ornitorrinco...), a análise

documentária e a linguagem documentária. *DataGramaZero: Revista de Ciência da Informação*, Brasília, v.2, n. n.6, p. 1-13, 2001.

LAUSEN, H.; STOLLBERG, M. *Portal Ontology*. Research Report. Digital Enterprise Research Institute. 2004. Disponível em: <http://sw-portal.deri.at/papers/deliverables/ portal_ontologyv0_2.pdf>. Acesso em 14 Jul. 2006.

LE COADIC, Y.-F. *A ciência da Informação*. Brasília: Briquet de Lemos, 1996.

LIMA, G. Â. B. O.. Interfaces entre a ciência da informação e ciência cognitiva. *Ciência da Informação*, Brasília, v. 32, n. 1, p. 77-87, 2003.

LOURENÇO, C. A. *Análise do padrão brasileiro de metadados de teses e dissertações segundo o modelo entidade-relacionamento*. Tese (Doutorado em Ciência da Informação), UFMG – Universidade Federal de Minas Gerais, Escola de Ciência da Informação, Belo Horizonte, 2005.

MACHADO, F. N. R.; ABREU M. *Projeto de banco de dados: uma visão prática*. São Paulo: Érica, 1995.

MARCHAL, B. *XML Conceitos e aplicações*. São Paulo: Berkeley, 2000.

MARCONDES, C. H.; MENDONÇA, M. A. R.; MALHEIROS, L. A estrutura dos elementos de metodologia científica nos textos de artigos de periódicos eletrônicos em Ciências da Saúde. In: *International Conference on Health Information and Libraries*, Salvador, Bahia, Brasil, Proceedings... Salvador, 2005. Disponível em <http://www.icml9.org/program/track5/public/documents/Carlos Henrique Marcondes-112049.doc>. Acesso em: 28 jun. 2005.

MARTINS, J. P. Foundation Of Knowledge representation and reasoning. Universidade Técnica de Lisboa. 2005.

MATOS, A. V. *UML – Prático e Descomplicado.* São Paulo: Érica, 2002. 187 p.

McGARRY, K. *O contexto dinâmico da informação*. Brasília: Briquet de Lemos, 1999.

MEISSNER, S. RDF - Resource Description Framework. 2004. Disponível em:

<http://en.wikibooks.org/wiki/RDF_-_Resource_Description_Framework>. Acesso em 10 ago. 2005.

MELO, T. M. L. Um Sistema para Extração e Classificação de Receitas Culinárias na Web. 2001. Trabalho de Conclusão de Curso (Graduação em Inteligência Artificial), UFPE - Universidade Federal de Pernambuco. Recife. 2001. Disponível em: <http://www.cin.ufpe.br/~tg/2000-2/tmlm.doc>. Acesso em: 03 jul. 2006.

MILES, A. *et al*. *SKOS-Core 1.0 Guide*. Tradução: Thiago Murakami. Disponível em: <http://murakami.objectis.net/artigos/skos>. Acesso em: 09 nov. 2005.

MORAES, A. F.; ARCELLO, E. N. O Conhecimento e sua representação. *Informação e Sociedade*, 2000. Disponível em: <http://www. informacaoesociedade.ufpb.br/pdf/IS1020004.pdf>. Acesso em: 20 nov. 2005.

MOREIRA, A.; OLIVEIRA, A. P. Contribuição da terminologia na modelagem de sistemas computacionais. *DataGramaZero: Revista de Ciência da Informação*, Brasília, v.6, n.5, 2005.

MOREIRA, A. *Tesauros e ontologias: estudo de definições presentes na literaratura das áreas da ciência da informação e da ciência da computação utilizando-se o método analítico-sintético*. 2003. 151f. Dissertação (Mestrado em Ciência da Informação). UFMG – Universidade Federal de Minas Gerais. Belo Horizonte. 2003.

MOREIRA, A. et al. O nível do conhecimento e os instrumentos de representação: tesauros e ontologias. *Revista de Ciência da Informação*, Brasília, v.5, n.6, 2004.

MORENO, F. P. . Requisitos Funcionais para Registros Bibliográficos - FRBR: um estudo no catálogo da Rede Bibliodata. *In: Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação*, 2005, Florianópolis. VI Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação, 2005.

MORGAN, M. *Glossary*. Disponível em: <http://docs.rinet.ru/WebLomaster/ appa.htm>. 2004. Acesso em: 08 jan. 2006.

NAVEGA, S. Técnicas para Representação Computacional de Conhecimento. Intelliwise: research and training. 2005. Disponível em: <http://www.intelliwise.

com/reports/info2005.pdf>. Acesso em: 23 jun. 2006.

NONAKA, I.; TAKEUCHI, H. *Criação de conhecimento na Empresa*. Como as Empresas Japonesas Geram a Dinâmica da Inovação; Tradução de Ana Beatriz Rodrigues, Priscila Martins Celeste. Rio de Janeiro: Campus, 1997.

OLIVEIRA, V. P. Uma informação tácita. *DataGramaZero: Revista de Ciência da Informação*, v.6, n.3. jun/05. Disponível em: <http://www.dgz.org.br/jun05/Art_04.htm>. Acesso em: 13 nov. 2005.

ORTEGA, C. D. Relações históricas entre Biblioteconomia, Documentação e Ciência da Informação. *DataGramaZero: Revista de Ciência da Informação* - v.5 n.5 , out/04.

PEREIRA, S. L. *Estruturas de dados fundamentais: conceitos e aplicações*. São Paulo: Érica, 1996.

PINHEIRO, L. V. R. Processo evolutivo e tendências contemporâneas da Ciência da Informação. *Informação e Sociedade*, João Pessoa, v. 15, n.1, 2005.

PINTO, Á. V. *Ciência e existência: problemas filosóficos da pesquisa científica*. 2.ed. Rio de Janeiro: Paz e Terra, 1979.

RICARTE, I. L. M. *Introdução a Orientação a Objetos*. Unicamp Web Site, 2001. Disponível em: <http://www.dca.fee.unicamp.br/cursos/POOCPP/node3.html>. Acesso em: 19 jun. 2006.

RICCO, Maria Filomena Fontes. *Construindo perfis compartamentais em ambiente organizacional: os estilos de mobilização dos gestores brasileiros*. São Paulo. FEA/USP, 2004. Disponível em: <http://www.teses.usp.br/teses/disponiveis/12/12139/tde-27072004-115228/publico/TeseFilomena.pdf>. Acesso em: 23 nov. 2005.

ROBREDO, J. *Da ciência de informação revisitada aos sistemas humanos de informação*. Brasília: Thesaurus, 2003. v. 1. 242 p.

ROBREDO, J. Organização dos documentos ou organização da informação: uma questão de escolha. *DataGramaZero: Revista de Ciência da Informação*, Internet, v. 5, n. 1, 2004. Disponível em: <http://www.datagramazero.org.br/fev04/

Art_05.htm>. Acesso em: 02 nov. 2004.

RODRIGUEZ, E. M. *Metadatos y recuperación de información: estândares, problemas y aplicabilidad em bibliotecas digitales*. Gijón: Trea. 2002. 432p. Disponível em: <http://eprints.rclis.org/archive/00002888/01/mendez.pdf>. Acesso em: 10 mai. 2006.

ROSA, I. S. *Soluções para EAD online numa perspectiva construtivista*. Universia, 2005. Disponível em: <http://www.universiabrasil.net/materia/materia.jsp?id=6354>. Acesso em: 28 nov. 2005.

RUCAVINA, Peter. *Creating an RSS feed of the books you have checked out of the library*. Disponível em: <http://ruk.ca/article/2290>. Acesso em: 07 mai. 2005.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. *Perspectivas em Ciência da Informação*, Belo Horizonte, v. 1, n. 1, p .41-62, jan./jun. 1996.

SILVA, O. J. *XML - aplicações práticas*, São Paulo: Érica, 2001.

SOUZA, M. I. F.; VENDRUSCULO, L. G.; MELO, G. C. Metadados para a descrição de recursos de informação eletrônica: utilização do padrão Dublin Core. *Ciência da Informação*, Brasília, DF, v. 29, n. 1, p. 93-102, 2000.

SOUZA, R. R.; ALVARENGA, L. A Web Semântica e suas contribuições para a ciência da informação. *Ciência da Informação*, Brasília, v. 33, n. 1, p.132-141, jan./abr. 2004. Disponível em: <http://www.ibict.br/cienciadainformacao/include/getdoc.php?id=356&article=71&mode=pdf>. Acesso em: 02 jun. 2006.

SYMANTEC, 2005. *Maximize o Seu Site para Mecanismos de Busca*. Disponível em: <http://www.symantec.com/region/br/smallbusiness/howto/max.html>. Acesso em: 25 jan. 2006.

TEIVE, R. C. G. *Planejamento da Expansão da Transmissão de Sistemas de Energia Elétrica Utilizando Sistemas Especialistas*. 1997. Tese (Doutorado em Engenharia de Produção) - Universidade Federal de Santa Catarina, Conselho Nacional de Desenvolvimento Científico e Tecnológico. Florianópolis. 1997.

TOLMASQUIM, A. T.; COSTA, A. M.; MIRANDA, M. L. C. Sistema de produção

do conhecimento para representação/recuperação da informação em história da ciência. In: *V CONGRESSO LATINO-AMERICANO DE HISTÓRIA DAS CIÊNCIAS E DA TECNOLOGIA*, 1998, Rio de Janeiro.

TRISTÃO, A. M. D.; Fachin, R. B. G.; Alarcon, O. E.. Sistemas de classificação facetada e tesauros: instrumentos para a organização do conhecimento. *Ciência da Informação*, Brasília, v. 33, n. 2, p. 161-171, 2004.

VILLAS, M. V. *et al. Estrutura de Dados: Conceitos e Técnicas de Implementação*. Rio de Janeiro: Campus, 1993.

VISUAL PARADIGM. *Visual Paradigm for the Unified Modeling Language 5.2 User's Guide Part1*. 2006. Disponível em: <http://www.visual-paradigm.com/product/vpuml/vpumluserguide.jsp?format=pdf>. Acesso em: 20 jul. 2006.

W3 CONSORTION. *About the World Wide Web Consortium*. Disponível em: <http://www.w3.org/>. Acesso em: 07 ago. 2005.

W3 CONSORTION. *RDF Vocabulary Description Language 1.0. RDF Schema.* 2004. Disponível em: <http://www.w3.org/TR/rdf-schema/>. Acesso em: 11 ago. 2005.

W3 CONSORTION. *Resource Description Framework (RDF)*. 2004. Disponível em: <http://www.w3.org/RDF/>. Acesso em: 11 ago. 2005.

WERSIG, G. Information science: the study of postmodern knowledge usage. *Information Processing & Management*. v. 29, n. 2, p. 229-239, 1993.

WIDMAN, L. E. Sistemas Especialistas em Medicina. *Revista Informática Médica*, vol 1 n.5. set/out 1998. Disponível em: <http://www.informaticamedica.org.br/informaticamedica/n0105/widman.htm>. Acesso em: 03 jul. 2006.

WOLF, M. *Teorias da Comunicação*. 5.ed. Lisboa: Editora Presença, 1999. p. 110-122.

WRIGHT, A. *Forgotten Forefather: Paul Otlet*. Disponível em: <http://www.boxesandarrows.com/archives/forgotten_forefather_paul_otlet.php>. Traduzido por Moreno Barros, disponível em: <http://tecnologica.extralibris.info/internet/o_antepassado_esquecido_paul_o.html>. Acesso em: 10 nov. 2005.

ANEXO A – TUTORIAL PROTÉGÉ

(por Sérgio Furgeri)

Este tutorial apresenta noções básicas da ferramenta Protégé para criação de ontologias. Considera-se que o leitor já possui conhecimentos sobre ontologias. São descritas algumas etapas da criação de uma ontologia chamada **Obra**, apresentada na dissertação denominada **REPRESENTAÇÃO DE INFORMAÇÃO E CONHECIMENTO: ESTUDO DAS DIFERENTES ABORDAGENS ENTRE A CIÊNCIA DA INFORMAÇÃO E A CIÊNCIA DA COMPUTAÇÃO**. Seguem os procedimentos a partir do download do Protégé, considerando-se a instalação para a plataforma Windows 2000/XP.

1. Faça o down*l*oad do protégé a partir do endereço <**http://protege.stanford.edu/download/download.html**>.

    a. Será necessário fazer um registro padrão para novos usuários no link "**register"**. Caso já tenha registro no site basta acessar o link "**download"**.

    b. Realizando o registro, você será direcionado ao link para baixar a ferramenta Protégé. Selecione o link semelhante à "**Download full Protégé release version...**", um o link com a versão mais atual da ferramenta na sua versão completa.

    c. Na página com os links para download, selecione o link referente a plataforma instalada em seu computador, ou seja, Windows, Linux etc.. Verifique se o arquivo a ser baixado contém a instalação do Java VM (Java Virtual Machine), pré-requisito para a execução do Protégé.

2. Após baixar o arquivo, dê duplo clique sobre ele para iniciar o processo de instalação. Surgirá uma janela de Introdução (Figura 1). Pressione o botão "**Next**".

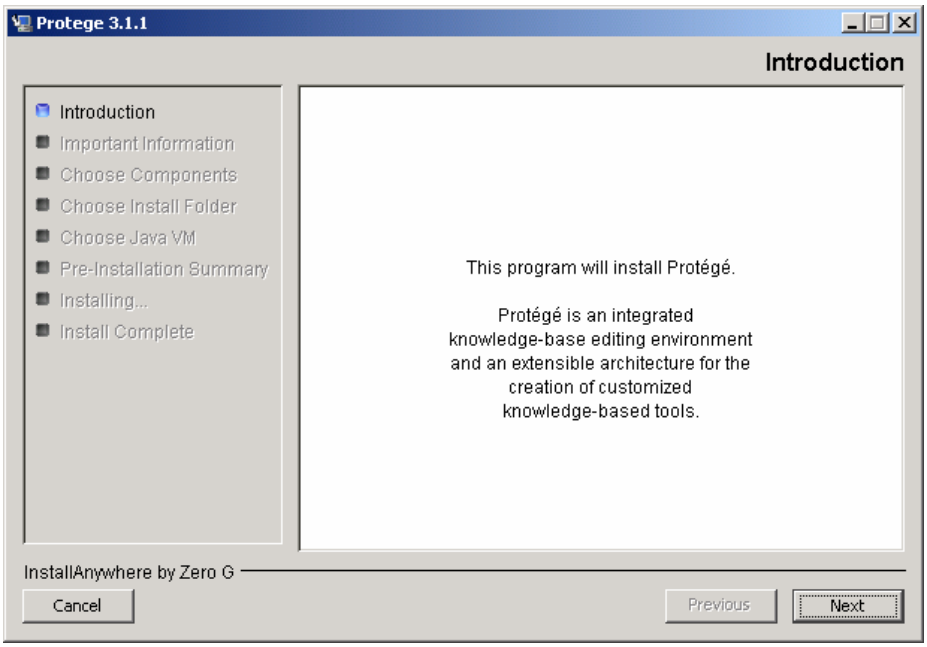

Figura 1. Janela de Introdução

3. Na seqüência, surgirá uma outra janela (Figura 2) com informações adicionais. Pressione o botão "**Next**".

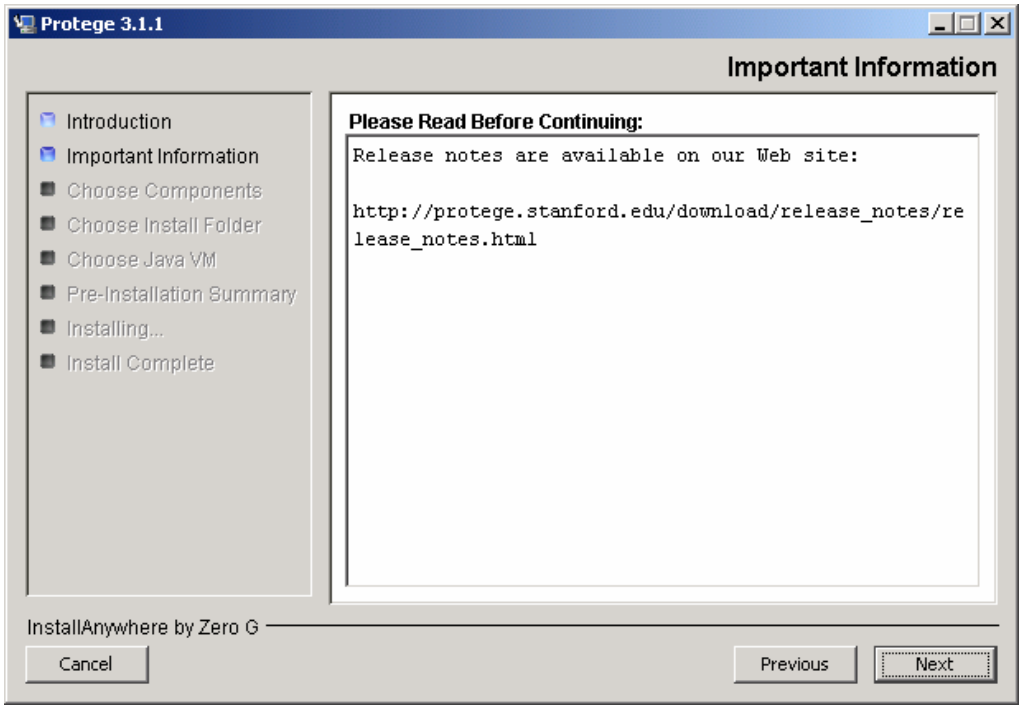

Figura 2. Informações adicionais

4. A próxima janela (Figura 3) permite selecionar os componentes que serão instalados. Aqui será escolhida a opção "**Basic + OWL**" que contém as funções básicas de uma aplicação Protégé e também suporte a linguagem OWL. Pressione o botão "**Basic + OWL**" e, a seguir, o botão "**Next**".

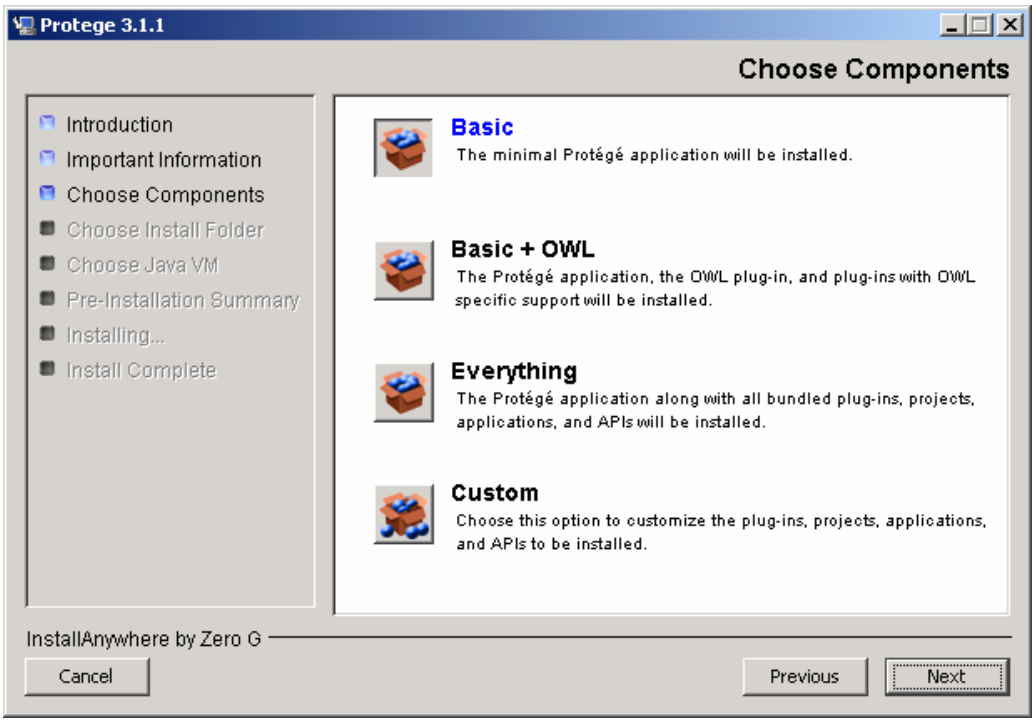

Figura 3. Escolha do tipo de instalação

5. Na janela seguinte (Figura 4), será definido o caminho onde ficará a instalação do Protégé no computador. Você poderá selecionar o diretório desejado através do botão "**Choose**". Em nosso caso, deixaremos o caminho padrão sugerido na instalação. Pressione o botão "**Next**".

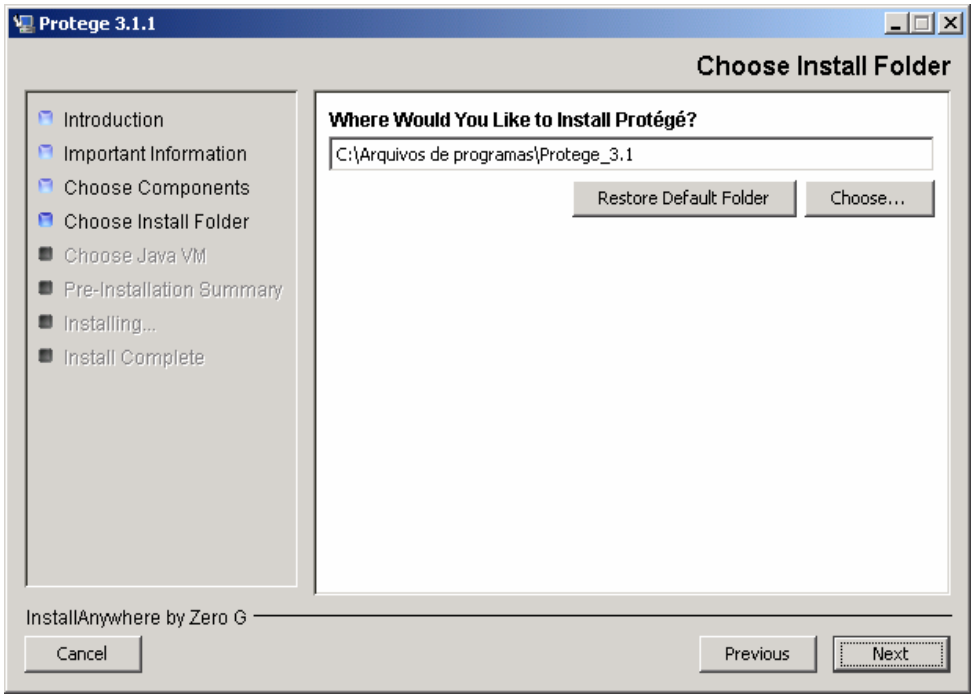

Figura 4. Caminho da instalação

6. Na seqüência, surgirá uma janela (Figura 5) para selecionar o caminho da JVM (Java Virtual Machine), pré-requisito para a execução do Protégé. Caso já exista uma versão da JVM instalada na máquina, seu caminho será localizado automaticamente. Selecione a JVM e, em seguida, pressione o botão "**Next**".

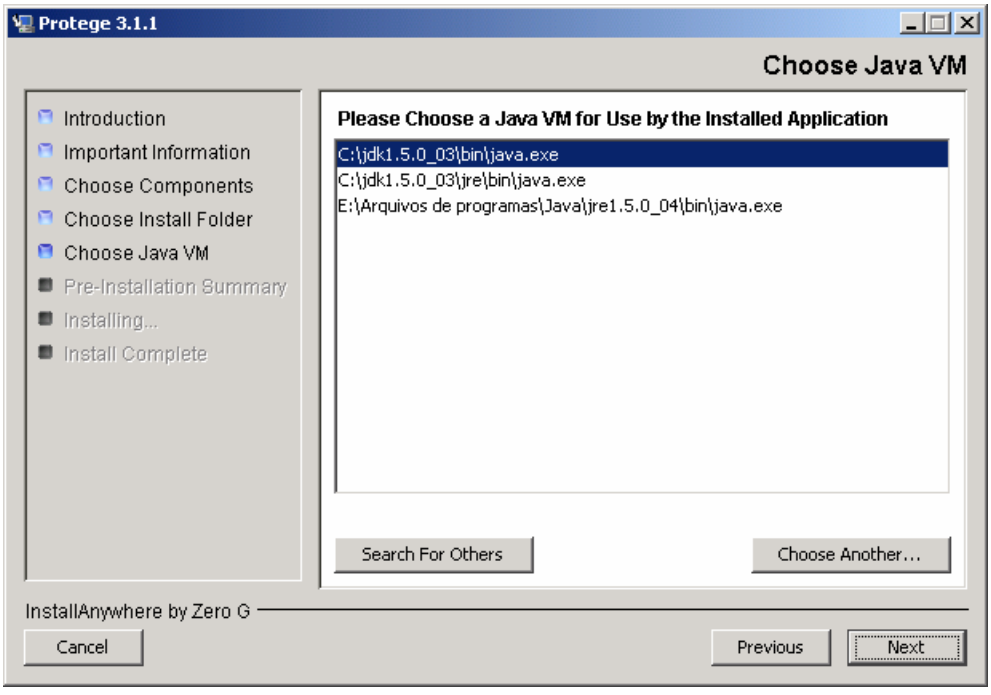

Figura 5. Escolha da JVM

7. Na janela seguinte (Figura 6) é apresentado um pequeno resumo de tudo que foi selecionado nas janelas anteriores. Pressione o botão "**Install**" para que os arquivos sejam instalados em seu computador.

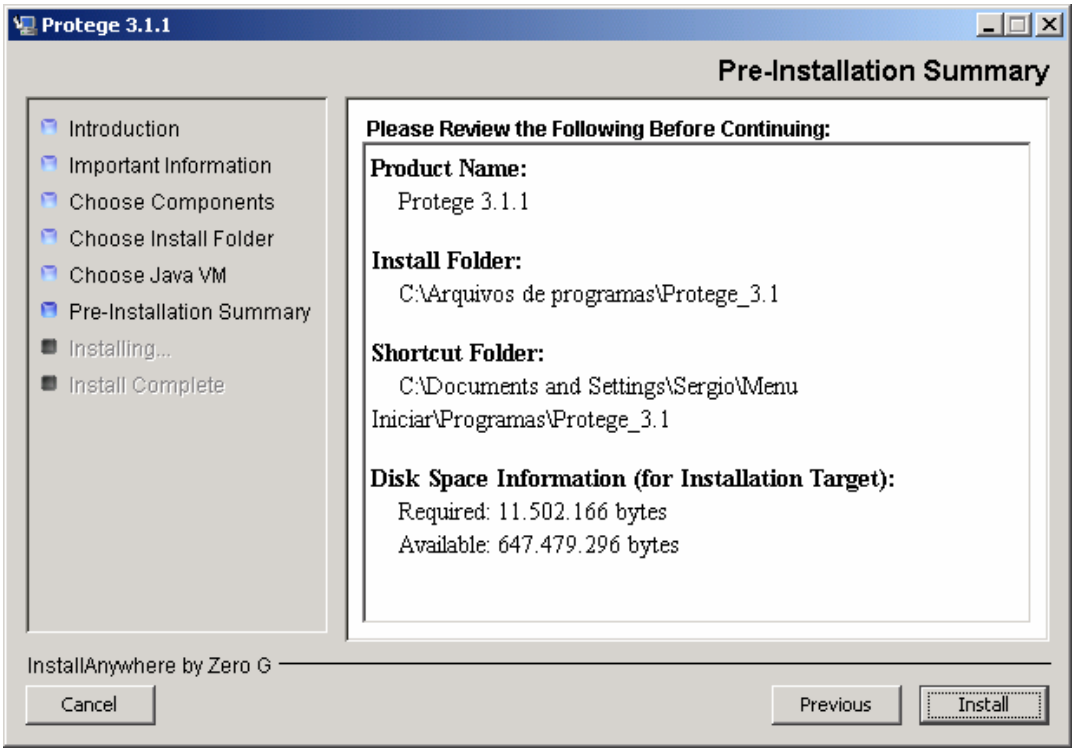

Figura 6. Resumo da instalação

8. A seguir, será apresentada uma janela informando que a instalação foi completada. Pressione o botão "**Done**" para sair da instalação.

9. Acesse novamente o endereço **http://protege.stanford.edu/download/download.html**. Esse endereço permite baixar diversos plug-ins para o Protégé. Iremos baixar o plug-in "TGVizTab", um item adicional que permite gerar a visualização gráfica de ontologias criadas pelo Protégé. Siga os procedimentos seguintes:

   a. Selecione o menu lateral "**PLUG-INS**".

   b. Escolha a opção "**view by topic**" para que uma nova janela do navegador se abra com a relação de links disponíveis.

   c. Selecione o tópico "**Visualization**".

   d. Escolha o link "**TGViz Tab Widget**", você será direcionado para uma nova página. Esta página contém informações para download do "TGVizTab", o gerador de visualização gráfica de ontologias. Mais adiante veremos a geração gráfica da ontologia a ser criada.

   e. Procure a seção de download da página e encontre a indicação "**Latest version**" ou a versão mais recente; na seqüência existe um link com o nome do plug-in (TGVizTab) e versão.

f. Clique sobre o link para realizar o download do TGVizTab. O item 26 apresenta os procedimentos necessários para instalar esse plug-in no Protégé.

10. Em seguida, execute a ferramenta Protégé através de Iniciar | Programas | Protégé. O programa será aberto e surgirá a caixa de diálogo (Figura 7) para que você selecione uma ontologia já existente, ou crie uma nova ontologia. Clique no botão "**Create New Project...**" para criar um novo projeto.

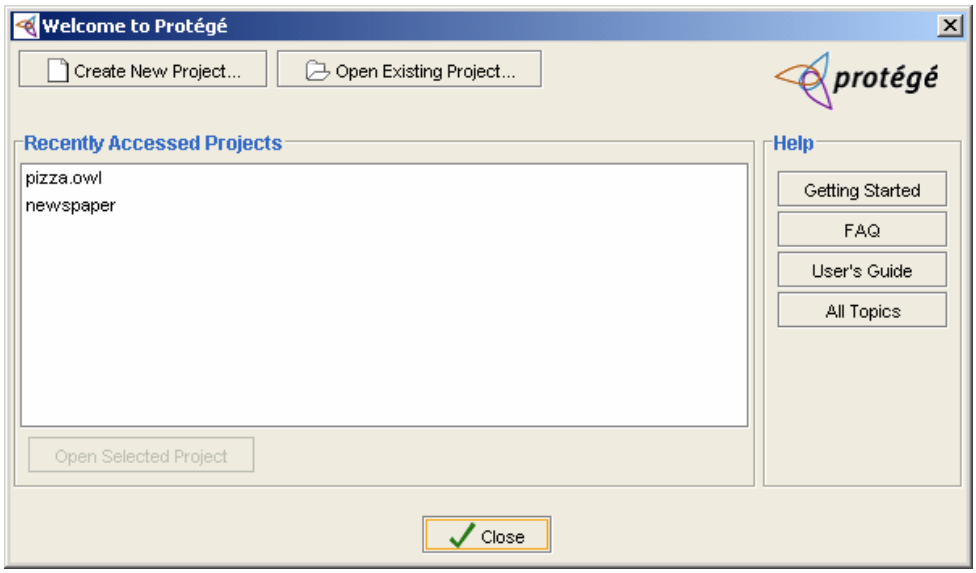

Figura 7. Criando um novo projeto

11. Na janela "**Create New Project**" (Figura 8), selecione a opção "**OWL Files (.owl or .rdf)**" e clique no botão "**Finish**".

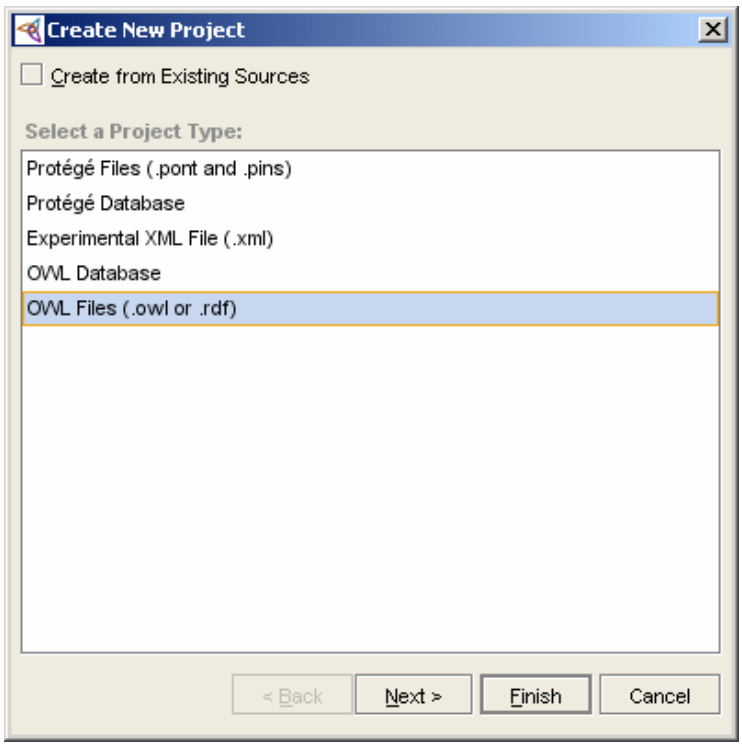

Figura 8. A janela Create New Project

12. A seguir serão definidas as propriedades do projeto (nome, localização etc.). Através do menu File | Save Project, salve o projeto com o nome "**Obra**", conforme as indicações apresentadas nos campos da Figura 9. Obs: o caminho para a gravação do projeto pode ser definido segundo a sua preferência, não necessitando seguir o mesmo caminho indicado na figura. Em seguida pressione o botão "**OK**".

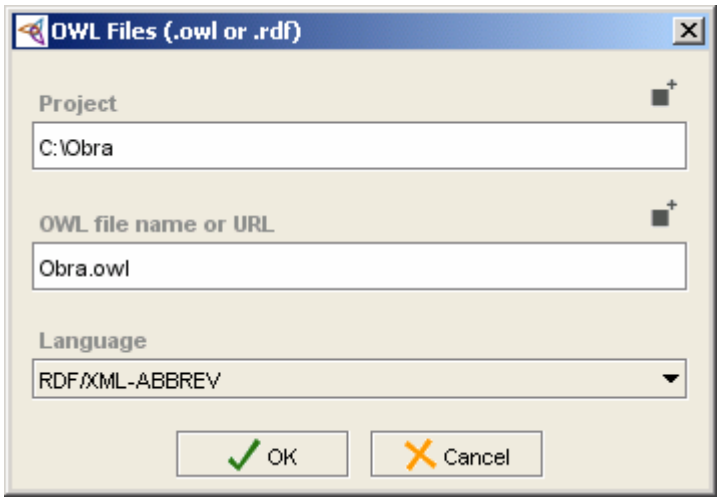

Figura 9. Propriedades do projeto

13. Baseado na Figura 10, clique na guia "**OWL Classes**", na janela "**SUBCLASS RELATIONSHIP**". Selecione o item "**owl:Thing**", a raiz da ontologia. Clique com o botão direito do mouse sobre ele e escolha a opção "**Create SubClass**" do menu.

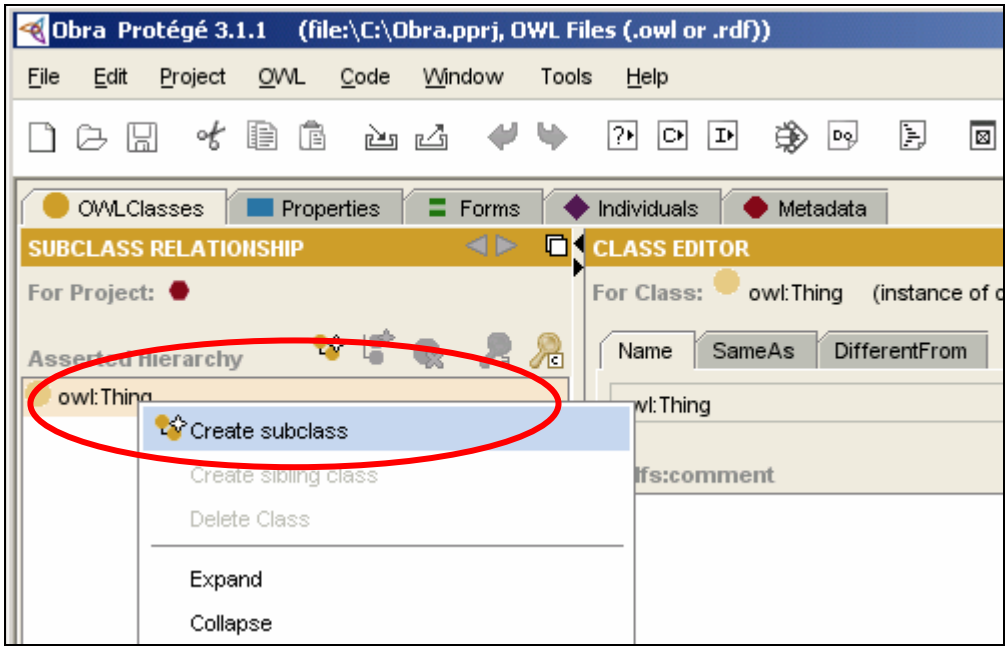

Figura 10. Criando uma nova classe

14. Será criada uma classe com o nome "Class_1". Em seguida podem ser definidas diversas propriedades da classe criada através da janela "**CLASS EDITOR**" (Figura 11). No campo "**Name**" desta janela renomeie a classe recém criada para "**Obra**".

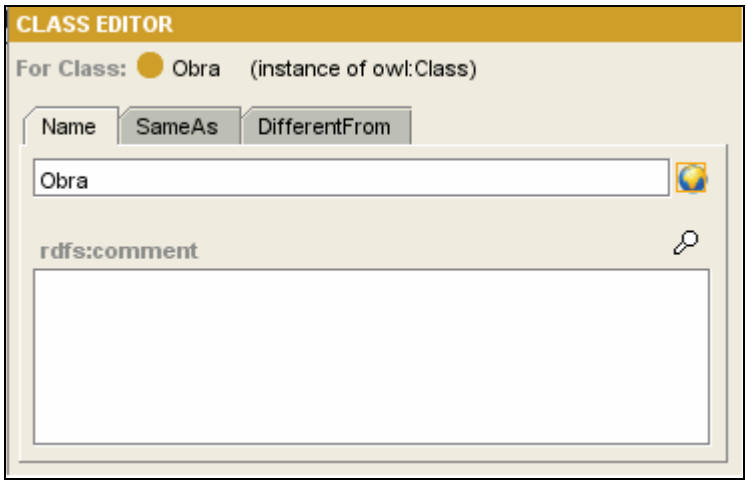

Figura 11. A janela CLASS EDITOR

15. Em seguida repita os procedimentos 13 e 14, criando duas novas subclasses chamadas "**Editora**" e "**Pessoa**", respectivamente. A hierarquia de classes deve estar de acordo com a Figura 12.

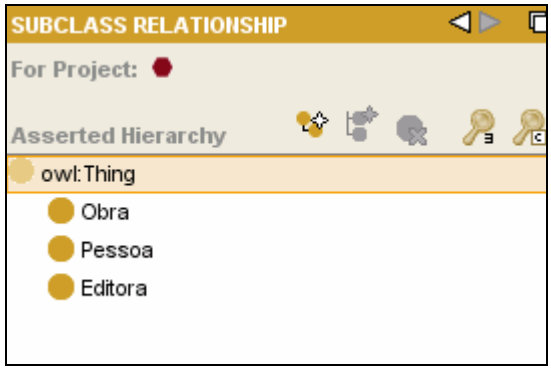

Figura 12. Hierarquia com três classes

16. Na janela "**SUBCLASS RELATIONSHIP**", selecione a classe "**Pessoa**", pressione o botão direito do mouse sobre ele, escolha a opção "**Create SubClass**" do menu. Nomeie a classe criada para "**Autor**". A hierarquia da janela "**SUBCLASS RELATIONSHIP**" deverá ficar de acordo com a Figura 13.

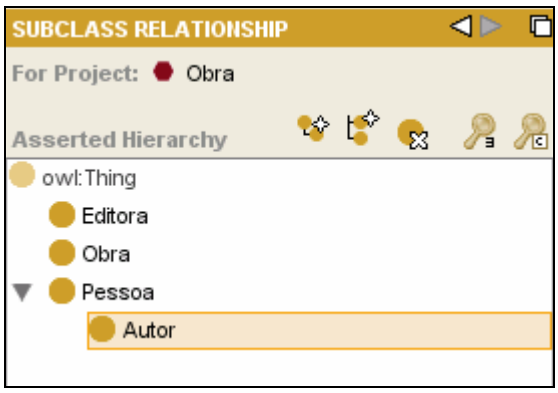

Figura 13. A classe Pessoa com a subcasse Autor

17. Selecione novamente a classe "**Pessoa**". No item "**Properties**" da janela "**SUBCLASS RELATIONSHIP**", haverá vários ícones. Selecione o

primeiro ícone, chamado “**Create datatype property**”, referente à criação de propriedades por tipos de dados.

18. Será aberta a janela para definição de propriedades (Figura 14), denominada “**Property Editor**”. No campo “**Name**” substitua “**DatatypeProperty_1**” por “**nome**”. No campo “**Range**” selecione a opção “**xsd: string**”. Isso define que a propriedade “nome” da classe “Pessoa” poderá armazenar valores do tipo string (texto).

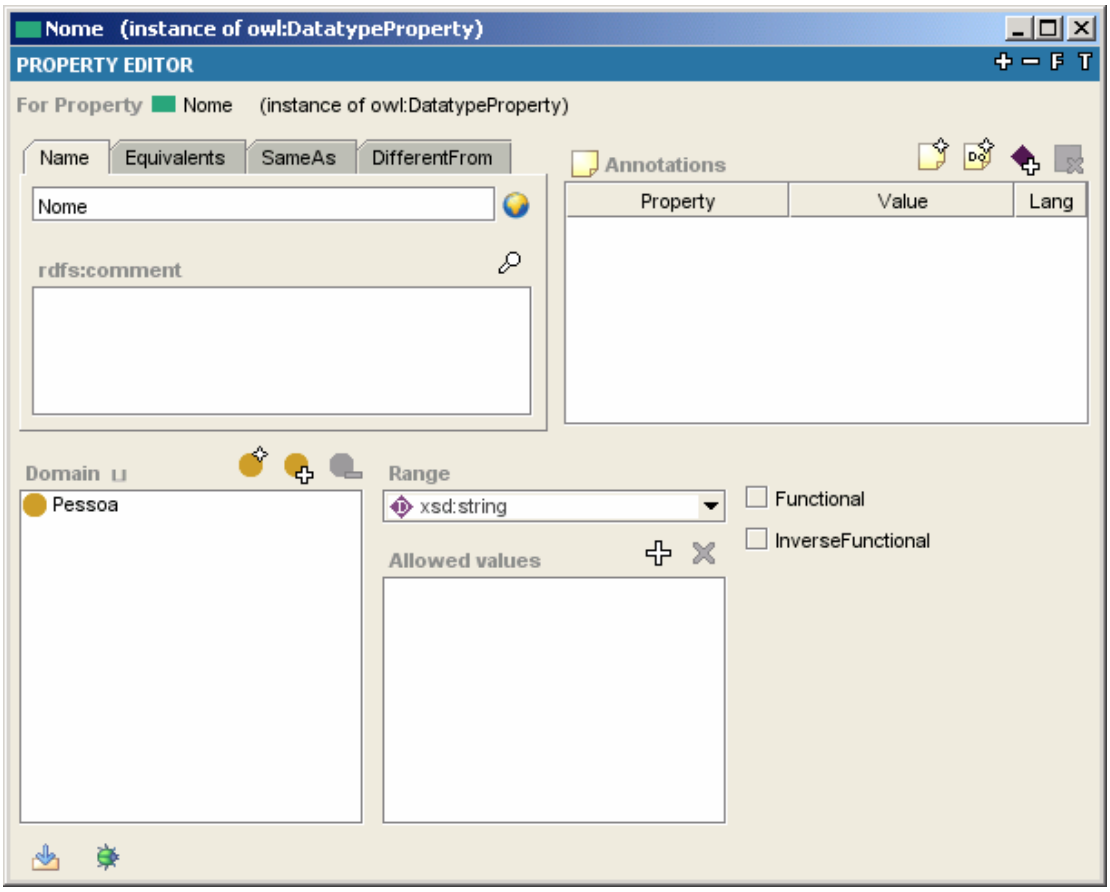

Figura 14. Definição de propriedades

Realize os mesmos procedimentos para a criação das propriedades cpf e sexo, conforme as definições seguintes:

- Cpf – Defina o campo Name como “**cpf**” e o campo Range como “**xsd: long**” (valores inteiros).

- Sexo - Defina o campo Name como “**sexo**” e o campo Range como “**xsd: string**”. Nesta propriedade, vamos definir mais uma característica. Em “**Allowed Values**” podemos definir o domínio de valores para “**sexo**”, isto é, os valores permitidos. No caso, só poderão ser fornecidos os caracteres “**m**” (masculino) ou “**f**” (feminino). Pressione o ícone com o símbolo de adição denominado “**Create value**” e, em seguida, digite o caractere “**m**” na caixa de diálogo. Repita o mesmo procedimento para incluir o caractere “**f**”. A Figura 15 ilustra esse processo.

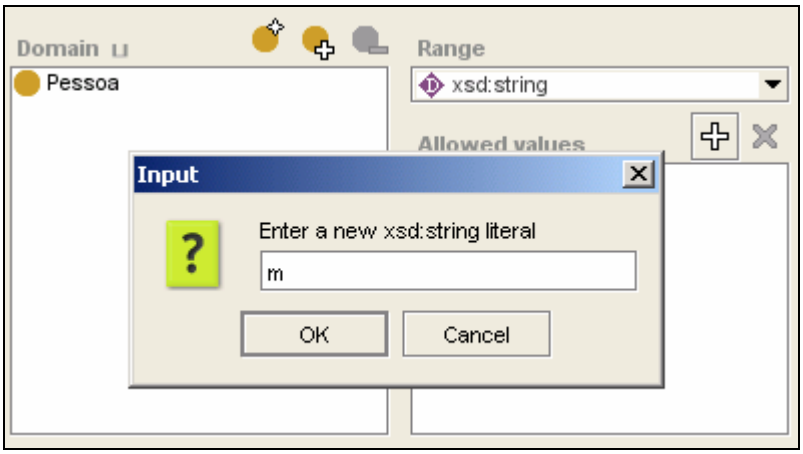

Figura 15. Adicionando valores permitidos para uma propriedade

19. Selecione a subclasse “Autor” da classe “Pessoa”. Perceba que as propriedades criadas na classe “Pessoa” foram herdadas pela sua subclasse “Autor”. Seguindo um procedimento semelhante ao da propriedade “sexo”, vamos definir o domínio para a propriedade “nível”.

    a. Defina o campo “Name” como “**nível**”.

    b. Defina o campo “Range” como “**xsd: string**”.

    c. Defina o campo “Allowed Values” com os valores “**Doutor**”, “**Mestre**”, “**Graduado**” e “**Técnico**”.

20. Da mesma forma que foi criada a subclasse “Autor” da classe “Pessoa”, selecione a classe “Obra” e crie uma subclasse chamada “Artigo”. A subclasse “Artigo” terá também três subclasses chamadas “ArtigoWeb”, “ArtigoRevista” e “ArtigoJornal”. A estrutura deve ficar como a demonstrada na Figura 16.

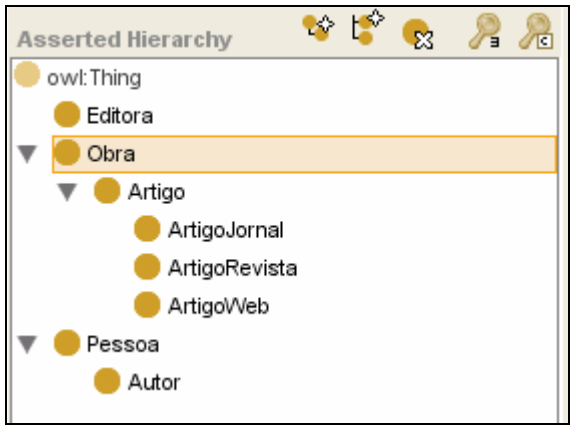

Figura 16. Subclasses da classe Artigo

A classe “Obra” possui ainda as propriedades listadas na Figura 17. Essas propriedades serão herdadas pela subclasse “Artigo”, uma vez que “Obra” é a superclasse de “Artigo”. Para criar as propriedades basta seguir os mesmos procedimentos apresentados anteriormente.

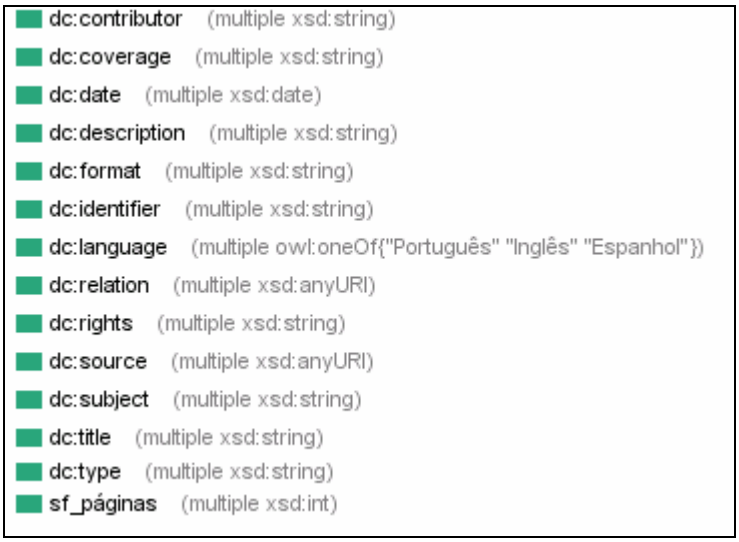

Figura 17. Propriedades da classe Obra

Além das propriedades herdadas através da classe “Obra”, na subclasse “Artigo” foram definidas também as propriedades relacionadas na Figura 18.

Figura 18. Propriedades adicionais da classe Artigo

21. Selecione o item referente à classe “Obra”. Vamos criar duas propriedades para ela, mas ao invés de serem propriedades de tipos de dados (como as criadas até aqui), iremos criar propriedades do tipo objeto. Pressione o segundo ícone da seção “**Properties**” chamado “**Create object property**” onde será aberta uma nova janela. Nesta janela definimos o campo “Name” como “**hasCreator**”.

    De forma semelhante às propriedades de tipos de dados, onde se define o tipo de dado e seus valores permitidos (domínio), a propriedade “hasCreator” (do tipo objeto) será usada garantir que uma obra só possua autores válidos, isto é, os que já tiverem sidos cadastrados no sistema. Isso é feito relacionando-se a classe “Obra” com a classe “Autor”. Siga os procedimentos seguintes.

    a. Na janela “**PROPERTY EDITOR**” da propriedade de objeto “**hasCreator**” no campo “**Domain LI**” a classe de destino dos dados já está definida, isto é, será mostrada a classe “**Obra**”.

b. No campo “**Range LI**”, selecione o terceiro ícone chamado “**Add Named Class...**” (Figura 19).

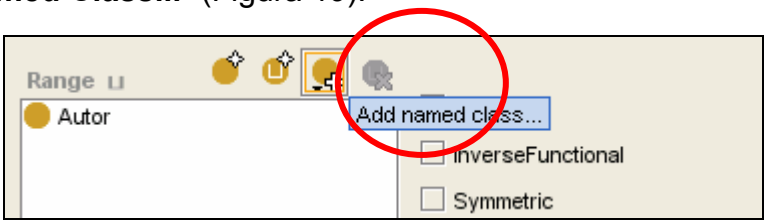

Figura 19. Relacionando classes

c. Será aberta uma nova caixa de diálogo com a estrutura das classes para que você possa selecionar a classe de origem dos dados. Selecione a classe “**Autor**” e pressione o botão “**OK**” como apresenta a Figura 20.

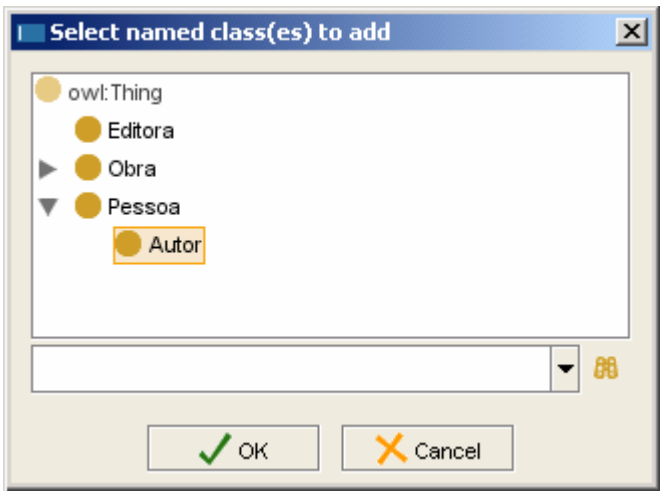

Figura 20. A subclasse Autor

d. Perceba que enquanto as demais propriedades foram criadas com os tipos de dados tradicionais como string, inteiro, inteiro longo, a propriedade de objeto “hasCreator” está relacionada a classe “Autor” e, por conseguinte, as propriedades definidas para essa classe. Observe na seção “Properties” que a propriedade “hasCreator” aparece definida como sendo do tipo Autor. Com isso, as duas classes ficam relacionadas e apenas nomes de autores cadastrados poderão estar vinculados a uma obra.

e. Na seqüência, crie uma propriedade de objeto usando os mesmos procedimentos anteriores. No campo “Name” insira o nome “hasPublisher”. No campo “Domain LI” mantenha o default já definido, e no campo “Range LI” selecione o item referente classe “Editora” e pressione o botão “OK”.

22. A seguir será usada a guia “**Forms**”. Essa guia permite definir o layout dos formulários a serem utilizados para a entrada de dados. Selecione a subclasse “**Autor**”, na estrutura de árvore à esquerda da janela. Á direita da janela encontra-se o “**FORM EDITOR**”. O real objetivo de demonstrar esta parte está relacionado ao campo “**Selected Widget Type**”.

    Além de toda modificação que podemos realizar no layout dos formulários para a inserção de dados, é especificamente no campo “Selected Widget

Type" que definimos como as propriedades poderão ser preenchidas, de acordo com o tipo de dado utilizado. As opções são:

a. MultiLiteralWidget: permite inserir vários valores. Por exemplo, se fosse necessário inserir diversos números de telefone para o mesmo autor. Não utilizaremos essa opção.

b. SingleLiteralWidget: (Figura 21): permite inserir um dado único, uma caixa de texto para o usuário digitar o conteúdo do campo. Em nosso exemplo, o preenchimento do campo cpf.

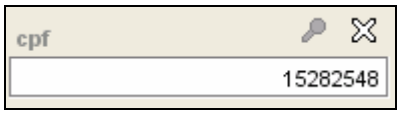

Figura 21. Campo com um único valor

c. DataRangeFieldWidget (Figura 22): podendo selecionar uma determinada opção disponível em um combo.

Figura 22. Lista de opções com DataRangeFieldWidget

23. Acesse a guia "**Individuals**". Neste local vamos incluir os conteúdos das classes da ontologia.

    a. Selecione a classe "**Editora**" no "**CLASS BROWSER**", e depois acesse a seção chamada "**INSTANCE BROWSER**" (no centro da janela).

    b. Pressione o ícone chamado "**Create Instance**". Será criada uma instância com o nome padrão de "Editora_1".

    c. Renomeie "Editora_1" para "DataGramaZero", dando duplo sobre a instância recém criada (abre a caixa de diálogo "**INDIVIDUAL EDITOR**") ou no campo "**Name**" da seção "**INDIVIDUAL EDITOR**" à direita da janela.

24. Na janela "**CLASS BROWSER**" selecione a classe "**ArtigoRevista**" e crie novamente uma instância repetindo o mesmo procedimento descrito no item anterior. Para esta instância daremos o nome de "**Organização_da_informação**". Para o preenchimento das demais propriedades pertencentes à classe defina os conteúdos que você achar mais apropriado.

25. Selecione a classe "**Autor**", crie uma instância, definindo o campo "**Name**" como "Robredo" e preencha as propriedades conforme abaixo:

    a. "**nome**": Jaime Robredo.

    b. "**nível**": doutor.

    c. "**sexo**": m.

    d. "**cpf**": 456456456

Os três últimos procedimentos listados (23, 24, 25), mostram como é possível criar instâncias.

26. O início do tutorial apresentou os passos para download do plug-in "**TGVizTab**". A seguir serão apresentados os procedimentos para instalá-lo no Protégé:

    a. Salve as alterações da ontologia criada e feche o Protégé;

    b. O arquivo **TGVizTab** está compactado. Dê duplo clique sobre ele para abri-lo;

    c. Extraia os arquivos na pasta "**plugins**", localizada na pasta de instalação do Protege;

    d. Execute o Protégé e abra a ontologia criada (Obra).

    e. Acesse o menu "**Project**" seguido do submenu "**Configure...**".

    f. Na janela que se abriu, marque a opção "**TGVizTab**" e pressione "**OK**".

    g. Note que foi inserida uma nova janela no Protégé, uma guia chamada "**TGVizTab**" que acabamos de adicionar. Selecione a guia "**TGVizTab**".

    h. Na janela do "**CLASS BROWSER**" selecione a classe "**Obra**".

    i. Pressione o ícone "**add class**" (parte superior da guia).

    j. Pressione o ícone "**configuration**". Na janela que se abriu, selecione a guia "**Settings**" e marque as opções apresentadas na Figura 23.

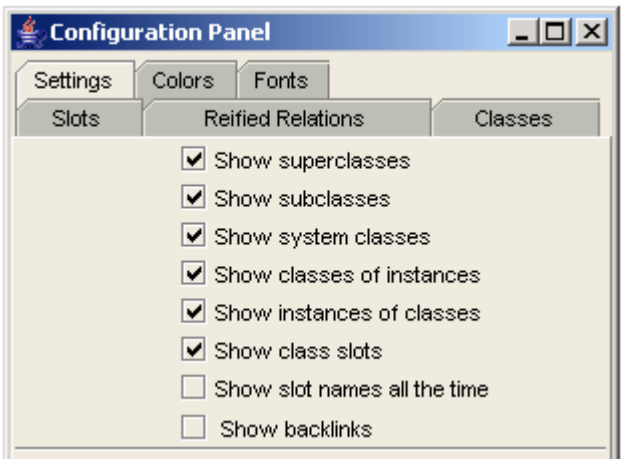

Figura 23. Configuração do gráfico

    k. Feche a janela e pressione o ícone "**create graph**". Após alguns segundos surge o gráfico demonstrando a estrutura da ontologia criada.

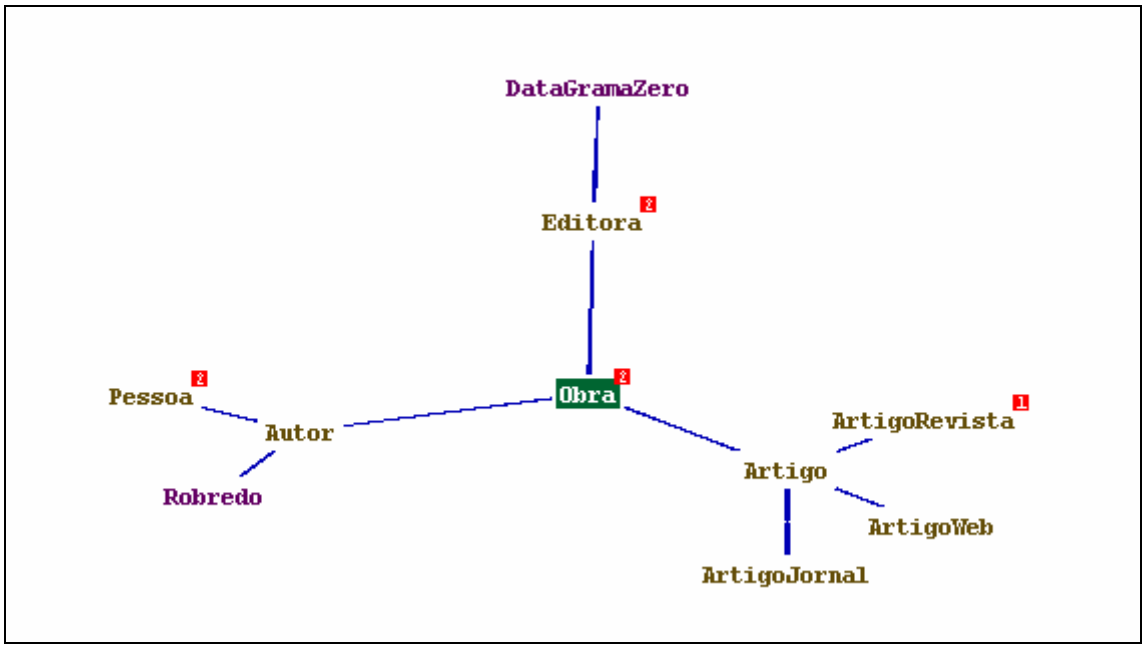

Figura 24. A estrutura da ontologia Obra

Algumas considerações a respeito do gráfico:

- Clicando e arrastando o cursor sobre os itens do gráfico é possível movimentá-lo.
- Com um duplo clique sobre os itens é possível modificar a visualização do gráfico, na perspectiva de outra classe ou instância.
- Através da barra de rolagem (parte superior da guia), é possível aumentar ou diminuir o zoom do gráfico.

Conforme citado anteriormente, esse tutorial apresentou apenas algumas etapas a serem realizadas na construção de uma ontologia. Para uma referência completa consulte a manual do Protégé.