Ontologias como novas bases de conhecimento científico

Carlos Henrique Marcondes

Professor do Depto de Ciência da Informação. Mestre e doutor em Ciência da Informação/UFF

Marília Alvarenga R. Mendonça

Professora do Depto de Ciência da Informação/UFF. Mestre em Administração de Sistemas de Informação

Luciana Reis Malheiros

Professora do Depto. de Fisiologia e Farmacologia, Inst. Biomédico/UFF, Doutoranda em Ciência da Informação, PPGCI UFF/IBICT

Leonardo Cruz da Costa

Professor do Depto. de Computação, Instituto de Computação/UFF, doutorando em Ciência da Informação, PPGCI UFF/IBICT

Tatiana Cristina Paredes dos Santos

Bolsista de Iniciação Científica, Curso Biomedicina, UFF

Periódicos científicos publicadosna Web são ainda calcados no modelo impresso. O conhecimento aí contido está em forma textual, não processável por programas. É proposto um modelo de ambiente de publicação na Web que permita a autores publicar seu artigo simultaneamente em formato textual e em formato "inteligível" por programas. Como domínio para avaliação e aperfeiçoamento do modelo, foram analisados 60 artigos de periódicos eletrônicos em Ciências da Saúde. O modelo e os resultados da análise são expostos e discutidos.

Palavras-chave: *Publicações eletrônicas; Metodologia científica; Comunicação científica; Representação do conhecimento; Ontologias; e-Ciência.*

Ontologies as the new bases of scientific knowledge

Scientific journals published on the Web are still based on a print model. Knowledge in those articles is in text format which is not processable by softwares. A Web publishing environment is proposed which enables authors to publish articles both as a text and a software "understandable" format. Sixty articles from Health Science electronic journals were analyzed to test and evaluate the model. The results of the analysis and the model proposed are shown and discussed.

Keywords: *Electronic publishing; Scientific methodology; Scientific communication; Knowledge representation; Ontologies; E-science.*

Recebido em 01.01.2008 Aceito em 13.10.2008

1 Introdução

A sociedade atual tem no conhecimento um de seus pilares econômicos e simbólicos. A reprodução social do conhecimento através de sua guarda, disseminação e uso é uma questão com implicações tão profundas para a sociedade humana, a ponto de, ao longo da sua evolução, ter lhe dedicado grandes esforços e criado instituições especiais para a manutenção e a expansão da Cultura. Museus, Bibliotecas e Arquivos são algumas dessas instituições, criadas com a missão precípua de preservar e disseminar a Cultura. As atividades dessas instituições e as metodologias aí utilizadas compõem o campo prático da Ciência da Informação.

No amplo espectro da Cultura, o conhecimento científico, desde a Modernidade, vem evoluindo segundo uma dinâmica própria, e passando desde então a se imbricar cada vez mais com o sistema produtivo (GONZALÉZ DE GOMEZ, 1987). Hoje a sociedade não produz sem o aporte do conhecimento em geral e, especificamente, do conhecimento científico. Não é por outra razão que a sociedade atual é chamada de sociedade da informação, sociedade do conhecimento ou "modo informacional de desenvolvimento", conforme Castells (1999, p.54). Meios

econômicos, sociais, políticos e tecnológicos são aportados pela sociedade atual para gerir seu acervo de conhecimento. O computador e a *Internet* são, muito justamente, o símbolo da "sociedade da informação" e da "sociedade do conhecimento".

A *Internet* adquire sua face atual com o surgimento do hipertexto e da *Web,* a "teia global", criados por Tim Berneers-Lee no CERN, em 1989. A *Web* transforma a *Internet* num gigantesco sistema de informações em escala mundial. No entanto, o crescimento gigantesco da Web a partir daí colocou novos e inéditos problemas para o acesso à informação aí disponibilizada. É muito fácil para qualquer um publicar na *Internet*, o que fez com que a rede tivesse um crescimento desordenado e caótico, e com que encontrar a informação adequada passasse a ser o principal problema cultural, econômico e científico da atualidade. Nunca a humanidade dispôs de tanta informação e, ao mesmo tempo, nunca foi tão difícil e problemático encontrar a informação relevante. Esse é o principal problema da atual economia da Informação (MARCONDES, 2001).

Para endereçar estes problemas, Tim Berners-Lee propõe a visão da *Web* Semântica, uma extensão da *Web* atual, formada por documentos compreensíveis unicamente por pessoas, para uma Web em que documentos seriam auto-descritíveis, de forma que seu conteúdo possa ser "compreendido" por programas, os agentes de "*software*"[1], que assim poderiam "raciocinar" e fazer inferências sobre o conteúdo de documentos, ajudando as pessoas em diferentes tarefas de recuperação de informações que exigem raciocínio, decisões, inferência de conclusões a partir de informações não explicitamente disponíveis ou de informações contextuais: "The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users" (BERNERS-LEE; HENDLER; LASSILA, 2001, p. 2).

O conhecimento científico, tão importante para nossa sociedade, vinha sendo guardado, preservado e disponibilizado desde a antiguidade nas coleções armazenadas em bibliotecas. O estatuto da guarda, preservação e disponibilização da cultura, dos conhecimentos em geral, e em especial do conhecimento científico, tão caro para as bibliotecas enquanto instituições, vem passando por um desafio com o surgimento da *Web.*

Artigos científicos são o veículo através do qual são disseminados os novos conhecimentos. Desde a publicação das *Philosophical Transactions* da *Royal Society*, na Inglaterra do século XVII, coleções de artigos científicos eram os repositório dos novos conhecimentos. No entanto, hoje as publicações científicas na *Web*, apesar do avanço das tecnologias da informação, são ainda calcadas no modelo impresso.

Artigos científicos são bases de conhecimento, mas através da leitura por seres humanos. Existem dois obstáculos para o acesso e a utilização em larga escala deste conhecimento: o grande número de

[1] Disponível em: <http://en.wikipedia.org/wiki/Software_agent.>. Acesso em: 6 nov. 2008.

publicações, a chamada "explosão informacional", que atinge mais alto grau com o surgimento da *Web* e das publicações eletrônicas; e o fato desse conhecimento estar inserido no texto dos artigos de forma não estruturada, legível somente por pessoas.

Hoje, grupos de pesquisadores lançam-se na tarefa de sistematizar e estruturar o conhecimento científico em domínios específicos, e de disponibilizá-lo publicamente na Web, através das chamadas ontologias, de modo a permitir que comunidades científicas compartilhem informações sobre domínios específicos. Ontologias são uma área de pesquisa emergente. Nas Ciências da Saúde, o consórcio *Open Biomedical Ontology*[2] (OBO) congrega mais de 70 diferentes ontologias. As ontologias mais conhecidas e utilizadas em Ciências da Saúde são *Gene Ontology*[3], usada por grupos no mundo inteiro para indexar o trabalho de seqüenciamento genético, e a *Unified Medical Language System* (UMLS) – *Semantic Network*, base terminológica mais conhecida da área médica, mantido pela *National Library of Medicine*. A proposta da UMLS é: "The purpose of NLM's Unified Medical Language System (UMLS®) is to facilitate the development of computer systems that behave as if they "understand" the meaning of the language of biomedicine and health[4]."

O termo Ontologia tem suas origens na Filosofia, como o estudo do ser e de suas condições de existência. Definições com origem na Ciência da Informação, mais ligadas ao foco do uso das ontologias na *Web* Semântica, são as seguintes:

> Ontology is defined as a formal explicit specification of a shared conceptualization[1]. It provides a shared and common understanding of a domain that can be communicated across people and application systems (DING; FOO, 2002b, p. 375).

> A partial conceptualization of a given knowledge domain, shared by a community of users, that has been defined in a formal, machine-processable language for the explicit porpoise of sharing semantic information across automated system (JACOB, 2003, p. 20).

Nestas definições se destacam as noções de "conceitualização" "compartilhada" num determinado "domínio" e de representação "formal", "processável por programas".

Ontologias foram pensadas no contexto da Web Semântica para tornar semanticamente interoperáveis sistemas computacionais distintos, como, por exemplo, um sistema automático de reserva de passagens aéreas com um sistema automático de reserva de hotéis de uma determinada cidade turística. O que, na semântica de um sistema, seria

2 Disponível em: <http://obo.sourceforge.net/>. Acesso em: 6 nov. 2008.
3 Disponível em: < http://www.geneontology.org/>. Acesso em: 6 nov. 2008.
4 Disponível em: <http://www.nlm.nih.gov/pubs/factsheets/umls.html>. Acesso em: 6 nov. 2008.

um *passageiro*, na semântica do outro seria um *hóspede*. Exemplos de interoperabilidade entre sistemas, como o mostrado, vão se tornar uma realidade, na medida em que as potencialidades da *Web* Semântica sejam implementadas: "agentes de *software*" inteligentes vão conseguir navegar pela Web e realizar reservas de passagens e hotéis, de forma a atender a agenda, as necessidades e preferências de seus usuários, automaticamente, dialogando com "web services" (BREITMAN, 2005, p.141), que disponibilizam passagens aéreas e hospedagem na cidade de destino, e "compreendendo" o funcionamento desses serviços através de ontologias. Vários exemplos desse tipo de aplicação são ilustrados no artigo de Tim Berners-Lee, James Hendler e Ora Lassila (2001), já citado.

A Ciência da Informação vem de uma longa tradição teórica e metodológica voltada para a questão da organização de domínios de conhecimento (CAMPOS, 2004), aplicada originalmente à organização de repertórios documentais, com contribuições como as de Outlet (1989), Ranganathan (1967), Dahlberg (1978) e Hjörland (2002a). Estas metodologias têm grande aplicabilidade no desenvolvimento de ontologias, e esta relação é cada vez mais percebida pelos pesquisadores da área (SOERGEL, 2000), (THE SEMANTIC WEB, 2003), (DING; FOO, 2002a; 2002b).

Esta pesquisa parte da constatação de que hoje o conhecimento contido em artigos científicos, mesmo aqueles publicados na Web, se encontra em forma textual, inteligível somente por pessoas. A Comunicação Científica (MEADOWS, 1999) tem sido o mecanismo através do qual novos conhecimentos são incorporados a uma área científica. A apropriação social desse conhecimento requer um longo processo, onde artigos são lidos, avaliados ("*peer review*"), criticados, citados, até que o conhecimento neles contido seja finalmente incorporado ao acervo de "conhecimento público" (ZIMAN, 1979) da humanidade. A forma textual desse conhecimento impede que ele possa ser processado por programas "agentes inteligentes", como é a proposta da Web Semântica, de modo a recuperá-lo de forma semanticamente muito mais rica, e a ajudar cientistas a identificar inconsistências científicas, novas hipóteses, novas descobertas.

A partir desta constatação, nossas questões de pesquisa são as seguintes:

- É possível publicar artigos científicos na Web simultaneamente em forma textual, como são hoje publicados, e em formato "inteligível" por programas "agentes inteligentes"?
- Uma vez que autores estão cada vez mais acostumados a publicar em ambientes de auto-publicação eletrônicos, como tantos que existem hoje[5], em especial os autores que

[5]Em repositórios e bibliotecas digitais, como o Diálogo Científico (Disponível em: <http://www.ibict.br/secao.php?cat=Diálogo%20Científico>. Acesso: 6 nov. 2008) e a Biblioteca Digital de Teses e Dissertações (Disponível em: <http://bdtd.ibict.br/>. Acesso em: 6 nov. 2008), em ambientes de publicação de periódicos eletrônicos, como o SEER – Sistema Eletrônico de Edição de Revistas (Disponível em: <http://www.ibict.br/secao.php?cat=SEER>. Acesso em: 6 nov. 2008), do IBICT, no qual se baseiam tantos periódicos eletrônicos brasileiros.

trabalham em Ciências da Saúde que estão acostumados a usar os chamados "structured abstracts" (BAYLEY; ELDREDGE, 2003), seria possível o desenvolvimento de um ambiente Web capaz de publicar artigos da maneira proposta anteriormente?
- Como representar em formato "inteligível" por programas o conhecimento contido em artigos científicos?

Temos respondido a estas questões, através das seguintes hipóteses:

- É possível estender as funcionalidades dos atuais ambientes *Web* de publicação para permitir que um autor, ao submeter seu artigo, interaja com o sistema e forneça, mais do que simples metadados descritivos do seu artigo, como acontece hoje, elementos que permitam ao sistema extrair, representar e registrar o conhecimento contido no artigo.
- Esse conhecimento tem como base os elementos semânticos do Método Científico, em especial a hipótese, que estabelece uma (nova) relação a fenômenos e poderia ser representado em formato "inteligível" por programas como uma ontologia.
- As ontologias, resultantes da publicação de artigos neste ambiente de publicação, poderiam ser comparadas por programas "agentes inteligentes" com ontologias públicas existentes na *Web*, como a UMLS, *a Gene Ontology*, etc, permitindo a identificação de inconsistências, "*gaps*" no conhecimento existente, ou indícios de novas descobertas.

Os resultados da pesquisa, até o momento, são apresentados na Seção 4. O objetivo é o de buscar modelos e metodologias que permitam publicar e representar o conteúdo de artigos científicos simultaneamente em formato legível por pessoas e "inteligível" por programas. Isso permitirá a recuperação desse conhecimento de forma semanticamente mais rica, melhorando a comunicação científica, permitindo que programas auxiliem cientistas a avaliar a consistência do conteúdo de artigos e a identificar novas descobertas científicas.

2 Metodologia

Pressupostos epistemológicos estão na base de qualquer proposta de representação do conhecimento (HJØRLAND, 2007). O modelo proposto de representação do conhecimento tem como base o Método Científico (BACON, 1973).

O domínio empírico da pesquisa vem sendo o das Ciências da Saúde, por possuir uma longa tradição de pesquisa sistemática, se constituindo no que Meadows (1999, p.39) chama de "tradições de pesquisa". Esta inclui um o alto grau de formalização e padronização da documentação científica[6] e dos procedimentos de pesquisa aí encontrados, facilitando a busca por um modelo que represente o conhecimento contido em artigos (IONNIDIS et al., 2006). Artigos científicos em Ciências da Saúde têm uma estrutura altamente formalizada, a assim chamada, *Introduction, Methods, Results, and Discussion* (IMRAD), cujo objetivo é, literalmente, refletir o Método Científico[7]:

> The text of observational and experimental articles is usually (but not necessarily) divided into sections with the headings Introduction, Methods, Results, and Discussion. This so-called "IMRAD" structure is not simply an arbitrary publication format, but rather a direct reflection of the process of scientific discovery .

Trata-se de uma pesquisa qualitativa. A partir de um modelo inicial proposto em Marcondes (2005), foram analisados 20 artigos das versões eletrônicas dos periódicos Memórias do Instituto Oswaldo Cruz[8], 20 da versão eletrônica do periódico *Brazilian Journal of Medical and Biological Research*[9], e 20 artigos internacionais sobre o tema células-tronco. A análise serviu para testar, avaliar e aperfeiçoar o modelo original.

Identifica-se o tipo de raciocínio empregado pelo autor. Procura-se a seguir identificar no texto as afirmações científicas, corroboradas ou ainda hipotéticas (a HIPÓTESE do artigo), feitas pelo autor, sob a forma de relações entre fenômenos; procura-se em seguida mapear os conceitos contidos nos elementos da HIPÓTESE, ANTECEDENTE, TIPO DE RELAÇÃO e CONSEQUENTE, com conceitos e tipos de relações da UMLS, para verificar se os mesmos são aí representados.

A análise é feita como se segue. O exemplo ilustrado se refere ao artigo de Camara et al. (2003). É um artigo *experimental-dedutivo*, pois os autores se baseiam, não numa hipótese original, mas sim em hipóteses anteriores formuladas por outros autores, citadas por eles. O que os autores reportam é somente a aplicação da hipótese anterior (HPV esta relacionado com lesões pré-neoplásicas e neoplásicas) a um contexto diferente (Mulheres, DF, Brasil) e avalia a prevalência neste grupo de lesões pré-neoplásicas e neoplásicas.

3 Quadro Teórico

6 Disponível em: <http://www.nlm.nih.gov/mesh/pubtypes2006.html>. Acesso em: 6 nov. 2008.
7 *The International Committee of Medical Journals Editors*. Disponível em: <http://www.icmje.org>. Acesso em: 6 nov. 2008.
8 Disponível em: <http://www.scielo.br/revistas/mioc>. Acesso em: 6 nov. 2008.
9 Disponível em: <http://www.scielo.br/revistas/bjmbr>. Acesso em: 6 nov. 2008.

O modelo proposto se baseou em contribuições teóricas da Ciência da Informação (comunicação científica, representação temática e indexação, representação do conhecimento), Metodologia, Epistemologia e Filosofia da Ciência (Método Científico, natureza do conhecimento científico, mudanças de paradigma, lógica das descobertas científicas), Ciência da Computação (representação do conhecimento) e Ciências da Saúde.

Em que consiste o conhecimento científico? Como chegar a ele? Como ampliá-lo? O conhecimento científico busca identificar a ordem presente na natureza (ALVES MAZZOTTI; GEWANDSZNAJDER, 2002), para controlá-la e prever o futuro: "a exigência de ordem se fundamenta na própria necessidade de sobrevivência" (ALVES, 1987, p. 36). Respostas a estas questões vêm sendo dadas desde os primórdios da cultura humana, mas, de uma forma mais sistemática, desde a Modernidade, a partir da qual a atividade científica passa a ter uma institucionalização crescente. Uma resposta à questão de como chegar ao conhecimento científico foi dada através da institucionalização do Método Científico (BACON, 1973). Ele é a base dos procedimentos de pesquisa nas ciências naturais, em especial nas Ciências da Saúde, domínio empírico da pesquisa, ensinado de forma sistemática durante a formação de pesquisadores e influenciando de forma determinante a Comunicação Científica[10]. Enquanto o Método Científico responde de forma adequada à pratica cotidiana de aquisição do conhecimento científico, as grandes mudanças no corpo da Ciência, chamadas de mudanças de paradigma, são melhor entendidas a partir da Teoria das Revoluções Científicas, de Thomas Kuhn (2003).

O que se entende por conhecimento é questão bastante discutida na Ciência da Informação. A visão corrente é que o conhecimento é um processo individual ocorrendo na mente de pessoas (BARRETO, 1999), visão reafirmada nos paradigmas cognitivos e sócio-cognitivos (ELLIS, 1992), (HJØRLAND, 2002b). No entanto, a Ciência da Informação tem especial interesse no *registro* do conhecimento e nas diferentes formas de representá-lo, para permitir sua apropriação social. Este interesse tem evoluído em direção à representação de conhecimento em formatos legíveis por computador. Vickery (1986), numa revisão sobre o tema, menciona registros de bases de dados e arquivos, estruturas de dados em programas computacionais como tipos diferentes de representações do conhecimento. Buckland (1991) distingue "*information as knowledge*", um processo intangível ocorrendo na mente de indivíduos, de "*information as thing*", conhecimento registrado em textos, registros, imagens etc.

Como representar o conhecimento? Como representar, em especial, o conhecimento científico, em formato "inteligível" por programas? Brookes (1980, p. 131) afirma que: "knowledge is a structure of concepts linked by their relations and information is a small part of such a

[10] Medical Subject Headings (MESH) - Publication Characteristics (Publication Types) - Scope Notes. Disponível em: <http://www.nlm.nih.gov/mesh/pubtypes2006.html>. Acesso em: 6 nov. 2008.

structure". Segundo Sheth, Arpinar e Khasyap (2003, p. 1): "Relationships are fundamental to semantics – to associate meaning to words, items and entities. They are a key to new insights. Knowledge discovery is about discovery of new relationships". Na Ciência da Informação, Farradane (1980, 267) critica os esquemas tradicionais de representação de conteúdo, como vocabulários controlados e tabelas de classificação, propondo uma metodologia que dá ênfase especial às relações: "The usual coordinate indexing systems use a thesaurus to control the vocabulary, selected as unconnected terms". E ainda: "Meaning, considered as relations between terms...". Sua proposta de "*Relational Indexing*" está baseada no papel central das relações para representar o conteúdo de documentos.

Nossa busca é por uma compreensão do que seria o conhecimento, em especial o conhecimento científico. Segundo Miller (1947, p. 306): "The above remarks imply that science is a search after internal relations between phenomena". Hjørland (2007) afirma que: "Today is "knowledge" often regarded as more relative to theories and perspectives, why it would probably be more correct to talk about knowledge claims". Poderíamos então falar de afirmações científicas. Um componente especial do Método Científico, a Hipótese, expressa uma afirmação científica como relação (ainda hipotética, a ser comprovada) entre fenômenos: "As hipóteses científicas geralmente procuram estabelecer relações entre fenômenos" (ALVES MAZZOTTI; GEWANDSZNAJDER, 2002, p. 70). Busca-se, então, uma representação do conhecimento científico que o expresse como uma relação entre fenômenos.

A comparação entre o conhecimento veiculado num determinado artigo com o conhecimento público, já estabelecido numa determinada área, para propiciar a identificação de novas descobertas, se daria, segundo o modelo proposto, comparando a ontologia resultado da publicação do artigo com ontologias públicas como a UMLS. A comparação de ontologias é o campo chamado na literatura de "*ontology mapping*" ou "*ontology aligment*" (DING; FOO, 2002a; 2002b), (DE BRUIJN et al., 2006), (SCHAFFE, 2006).

O ambiente Web Semântica tem aplicabilidade direta nas ciências em geral, especialmente nas Ciências da Saúde (BAKER; CHEUNG, 2007), (W3C Semantic Web Healthcare and Life Sciences interest group – HLCS[11]), nos chamados ambientes de e-Science (DE ROURE; JENNINGS; SHADBOLT, 2001) ou "*knowledge environments*"[12].

4 Resultados

A proposta de um ambiente Web que permita a pesquisadores a auto-publicação de seus artigos simultaneamente em formato legível por pessoas e "inteligível" por programas é delineada na FIG. 1. Esse

[11] Disponível em: <http://esw.w3.org/topic/HCLS>. Acesso em: 6 nov. 2008.

[12] Disponível em: <http://www.esi-bethesda.com/ncrrworkshops/kebr/index.aspx>. Acesso em: 6 nov. 2008.

ambiente pressupõe o desenvolvimento de um "*software*" específico, que registre o conteúdo de um artigo como uma ontologia, e será objeto de pesquisa futura.

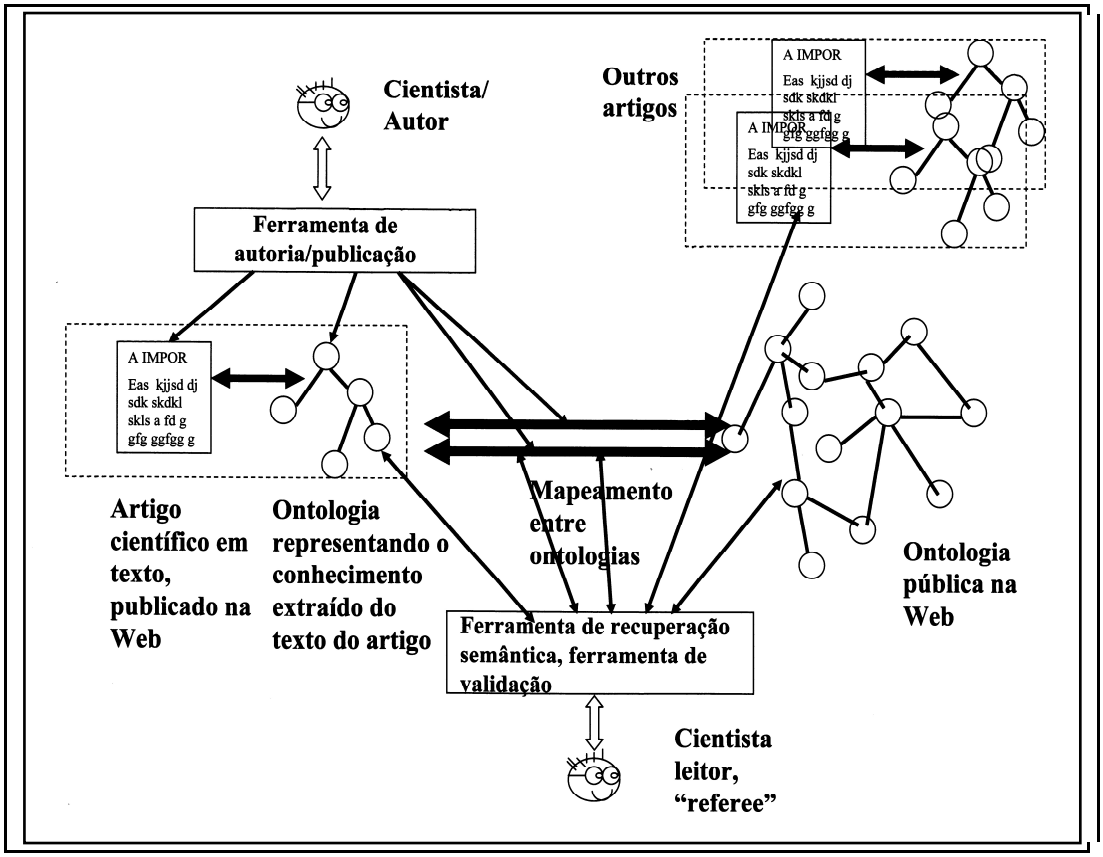

FIGURA 1 – Ambiente de autoria/auto-publicação.

Até o momento, a pesquisa se concentrou em desenvolver um modelo para representar o conteúdo de artigos em formato "inteligível" por programas. A seguir são descritos: o procedimento de análise desenvolvido, que simula os procedimentos que seriam feitos pela ferramenta de edição/publicação ao interagir com um autor/pesquisador para extrair e representar o conhecimento contido no seu artigo; o modelo, na forma de uma ontologia para representar o Conhecimento Contido em Artigos Científicos; e, finalmente, os resultados quantitativos da análise dos 60 artigos que subsidiaram a elaboração e o aperfeiçoamento do modelo.

4.1 "Ontologia" para representar o Conhecimento Contido em Artigos Científicos:

A análise de um artigo inicia-se por classificá-lo com base no tipo de raciocínio empregado. Esta classificação é baseada em Hutchins (1997) e Gross (1990), e em textos a partir da visão de abdução em Pierce (1977),

como processo de descoberta de novos "*insights*" em Ciência (HOFFMAN, 1997), (MAGNANI, 2001), (PAAVOLA, 2004).

Baseado nestas propostas, considerou-se a seguinte classificação: artigos podem ser teóricos ou experimentais. Artigos teóricos seriam os que propõem novas hipóteses e artigos experimentais testam experimentalmente hipóteses já formuladas ou formulam e testam experimentalmente uma nova hipótese; estes podem usar os métodos de raciocínio dedutivo (no primeiro caso) ou indutivo (no segundo).

Artigos teóricos se caracterizam por discutirem questões de maior abrangência. Analisam criticamente diversas hipóteses anteriores, mostrando suas fragilidades. Estes artigos são os que têm mais potencial de apresentarem contribuições para a Ciência, já que discutem ou questionam o paradigma vigente (KUHN, 2003), (OLIVA, 1994). Sua contribuição é uma nova hipótese, indicando um novo caminho de pesquisa. O tipo de raciocínio empregado é o abdutivo, ou seja, o "*insight*" sobre a solução de questões não explicadas pela Ciência e a formulação de novas hipóteses de solucioná-las.

Artigos experimentais se dividem em dedutivos e indutivos. Ambos se caracterizam por discutirem questões num escopo de abrangência limitado. Não discutem os rumos de uma teoria científica, mas se limitam a confirmá-la ou aperfeiçoá-la. Sempre trazem resultados experimentais.

Os artigos que utilizam o raciocínio dedutivo trabalham a partir de hipóteses já formuladas anteriormente, cujas referências vêm citadas, aplicando-as a um contexto específico. Os artigos que utilizam o raciocínio indutivo se caracterizam por formularem e testarem uma proposta com certo grau de originalidade, dentro do paradigma científico vigente.

Os componentes semânticos identificados são os seguintes: um artigo científico se organiza a partir de um PROBLEMA; um PROBLEMA expressa uma carência, insatisfação ou deficiência conceitual com o atual estado de coisas num domínio de conhecimento. A partir do PROBLEMA, este é inserido numa relação que pode resolver a carência ou deficiência; esta relação é a HIPÓTESE. Uma hipótese enuncia relações entre fenômenos. Uma HIPÓTESE se desdobra em ANTECEDENTE, TIPO-RELAÇÃO e CONSEQUENTE. Um autor, num artigo, pode formular uma hipótese original – HIPÓTESE(o) -, ou tomar a hipótese anterior – HIPÓTESE(a) - de outros autores; neste caso uma ou mais citações referentes à HIPÓTESE(a) – CITAÇÕES(h) - são feitas. Um autor também pode analisar várias HIPÓTESEs (a) para mostrar que elas são insatisfatórias como soluções para o PROBLEMA, e formular sua HIPÓTESE(o). Um artigo teórico se justifica simplesmente por propor uma nova HIPÓTESE(o). Da hipótese, num artigo experimental, deve ser derivado um EXPERIMENTO capaz de ser observável empiricamente. Em um artigo científico, significa ter RESULTADOS observados segundo determinada MEDIDA, em determinado CONTEXTO, segundo determinada METODOLOGIA. Este CONTEXTO, onde os fenômenos relacionados na HIPÓTESE são observados, pode ser desdobrado em AMBIENTE –

comunidade ou instituição onde o fenômeno ocorre -, ESPAÇO - o lugar onde o fenômeno ocorre -, TEMPO, ou época em que o fenômeno ocorre, e GRUPO de indivíduos onde o fenômeno ocorre.

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

- dado um PROBLEMA, com os seguintes aspectos e dados;
- os seguintes Autores/HIPÓTESES anteriores para sua solução não são satisfatórios(as);
- diante disso propõe-se a seguinte HIPÓTESE original.

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

- dado um PROBLEMA, com os seguintes aspectos e dados;
- os seguintes Autores formularam HIPÓTESES anteriores para sua solução;
- diante disso, escolhemos a seguinte (uma das HIPÓTESE anteriores);
- ampliamos e re-contextualizamos esta HIPÓTESE anterior; desenvolvemos o seguinte EXPERIMENTO para testar esta HIPÓTESE anterior;
- o EXPERIMENTO apresentou os seguintes RESULTADOS.

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

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

Os artigos são analisados conforme descrito na Seção 3.- Metodologia.

Estes esquemas descritos acima foram formalizados numa ontologia de domínio, no domínio específico do raciocínio baseado no Método Científico, na forma como ele é expresso no texto de artigos científicos. Apresentamos na FIG.. 2 o modelo da ontologia desenvolvida, representada como um diagrama de classes:

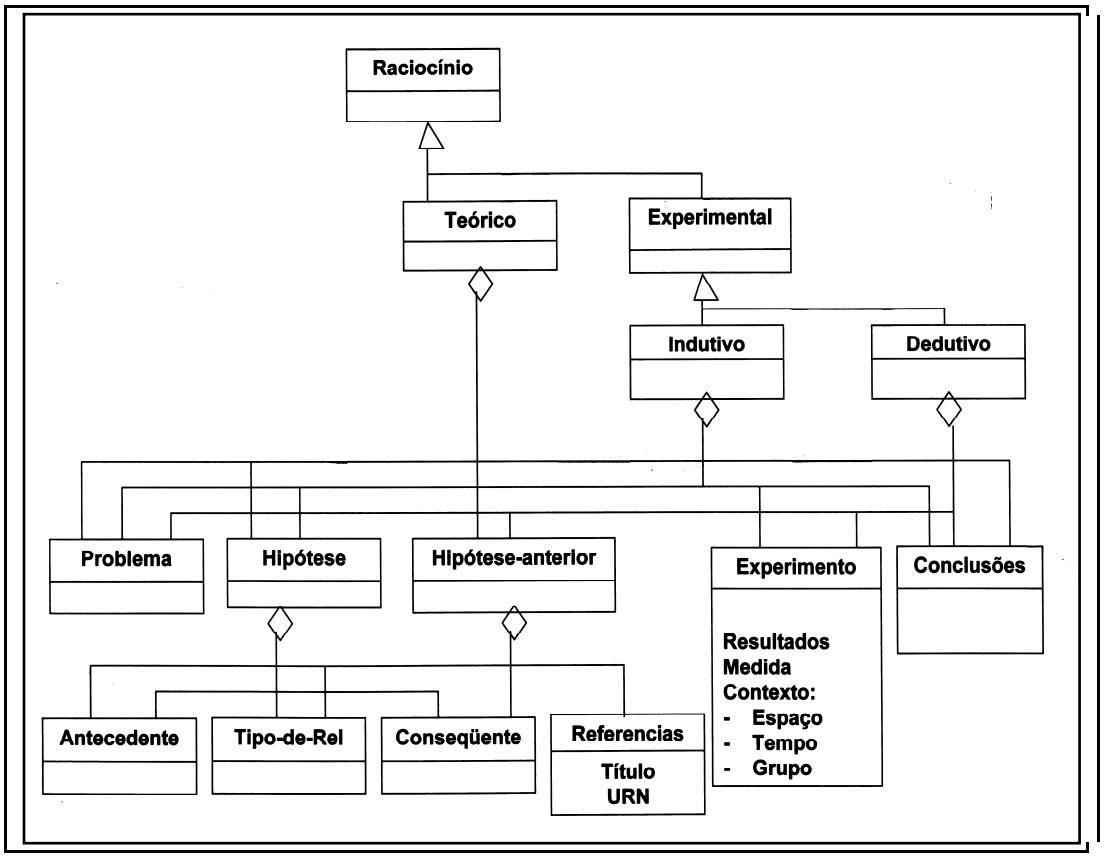

FIGURA 2 – Ontologia do Conhecimento Científico em Artigos.

Uma versão inicial dessa Ontologia implementada na linguagem OWL[13] pode ser encontrada em <http://www.professores.uff.br/marcondes/Scientific_reasoning.owl>.

4.2. Dados quantitativos da análise dos artigos:

Os dados quantitativos, resultado da análise de 60 artigos científicos em Ciências da Saúde, segundo os procedimentos descritos anteriormente, são apresentados a seguir.

TABELA 1 - Artigos por tipo de raciocínio

Artigos analisados	Exp-indutivos	Exp-edutivos	Teóricos abdutivos	TOTAL
MIOC	4	15	1	20
BJMBR	4	14	2	20
CÉLULAS-	10	10	0	20

[13] Ontology Web Language (OWL), um linguagem padrão do W3C para representar ontologias, Disponível em: <http://www.w3.org/2004/OWL/>. Acesso em: 6 nov. 2008.

TRONCO				
TOTAL	18	39	3	60

Fonte: Dados da pesquisa.

Como pode ser visto na tabela, a maior parte dos artigos é experimental (57 em 60) e, dentro desses, a maior parte é experimental dedutiva (39 em 57), caracterizado um padrão de "ciência normal" segundo a proposta de Thomas Kuhn (1970). Os artigos que trazem contribuições significativas para a Ciência, em termos de mudança de paradigma ou "revolução científica", nas palavras de Kuhn, seriam os Teóricos-abdutivos, dos quais só foram encontrados três.

TABELA 2- Resultado do mapeamento dos conceitos encontrados nas hipóteses em conceitos da UMLS

Artigos analisados	**MIOC**	**BJMBR**	**CÉLULAS-TRONCO**	**TOTAL**
Total de artigos	20	20	20	60
Totalmente mapeados	7 (35,0%)	3 (15,0%)	0 (0%)	10 (16,7%)
Parcialmente mapeados	13 (65,0 %)	11(55,0%)	14 (70,0%)	38 (63,3%
Não mapeados	0 (0%)	6 (30,0%)	6 (30,0%)	12 (20,0%)

Fonte: Dados da pesquisa.

O percentual de artigos "não mapeados", ou seja, aqueles em que o conhecimento contido não correspondeu aos conceitos existentes na base de conhecimento utilizada, no caso a UMLS, é maior no grupo células-tronco, quando comparado aos artigos dos periódicos brasileiros, confirmando o fato de que esta é uma área de rápido avanço da Ciência, em que os artigos trazem novidades e novas descobertas.

5 Discussão

Em todos os artigos analisados, uma relação expressando o conhecimento contido no artigo foi identificada. Isso reforça a idéia de que relações entre fenômenos são a essência do conhecimento expresso em artigos científicos. A idéia de verificar o mapeamento de conceitos identificados nas relações expressas em cada artigo, em conceitos da UMLS, mostrou também que essa pode ser uma alternativa promissora para o uso do modelo proposto na identificação de novas descobertas. Certamente muita pesquisa ainda deverá ser desenvolvida nesta direção para que esta metodologia seja viável.

O modelo permitiu identificar claramente um padrão de "ciência normal" (KUHN, 2003) nas publicações analisadas em geral, em especial na área de células-tronco. Nesta área, todos os 20 artigos são experimentais, nenhum é teórico-abdutivo. Em muitos deles, o tipo de relação encontrada é método (UMLS Semantic Network "method" T18314). Os dados indicam que a área, depois da descoberta do potencial terapêutico das células-tronco, trabalha dentro desse paradigma, ainda aperfeiçoando métodos para resolver duas questões-chave: como manter células-tronco indefinidamente sem se diferenciarem (para a criação de bancos de células-tronco) e como controlar a diferenciação de células-tronco em tipos específicos, como células cardíacas, hepáticas, ósseas, pulmonares etc.

O modelo atual, embora testado com 60 artigos, é ainda simples e tosco, fortemente calcado nos elementos de metodologia científica conforme eles aparecem no texto de artigos em Ciências da Saúde. Acredita-se que ele possa ser mais desenvolvido, na medida em que novas situações apareçam em outros artigos científicos a serem analisados. Os elementos do raciocínio científico presentes se apresentam de forma estruturada, viabilizando sua representação como uma ontologia, no sentido comum desse termo, quando usado na engenharia do conhecimento. Isso possibilitará que "agentes de *software*" realizem "inferências" sobre o conhecimento assim representado. Com base na análise desenvolvida para o artigo usado como exemplo na Seção 2, seu conteúdo, registrado segundo o modelo proposto, permitiria as seguintes consultas por um "agentes de *software*" de recuperação semântica de informações:

- Que artigos (também) tem hipóteses relacionando HPV como causa de lesões pré-neoplásicas e neoplásicas em mulheres?
- Que artigos tem hipóteses relacionando outros fatores que não HPV como causa de lesões pré-neoplásicas e neoplásicas em mulheres?
- Que artigos tem hipóteses relacionando HPV como causa de lesões pré-neoplásicas e neoplásicas em outros grupos?
- Que artigos teóricos levantam hipóteses relacionando HPV como causa de outras patologias em mulheres?
- Em que diferentes condições contextuais existem artigos com hipóteses relacionando HPV como causa de lesões pré-neoplásicas e neoplásicas em mulheres?

6 Conclusões

Os métodos convencionais de análise temática em Ciência da Informação têm por objetivo determinar o tema, o "*aboutness*" (HJØRLAND, 2001; LANCASTER, 1993) de um documento, para indexá-lo

[14] Disponível em: <http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html>. Acesso em: 6 nov. 2008.

e permitir sua recuperação num sistema de recuperação de informações para que, aí sim, o conhecimento contido no documento seja processado através de sua leitura por seres humanos. Todos os exemplos citados por Farradane (1980) são neste sentido. Ao contrário, a proposta dessa pesquisa não é a de representar o tema de um artigo, mas sim as afirmações científicas ou relações entre fenômenos num determinado domínio científico, estabelecidas pelo autor em um artigo em formato "inteligível" por programas, para permitir que "agentes de *software*" façam inferências baseadas nessas afirmações. Se em um artigo, determinado autor, baseado nos experimentos por ele realizados, faz uma afirmação relacionando câncer de pulmão com o uso excessivo de refrigerantes tipo cola, e se numa base de conhecimento pública estão registrados como fatores causadores de câncer de pulmão somente o consumo de cigarros e ambientes poluídos, então um "agente de *software*" pode inferir e alertar um cientista para a possibilidade do artigo trazer uma novidade científica. Se vários artigos fazem esta relação, um "agente de *software*" pode inferir e sugerir a um cientista que existam evidências fortes[15] de que o consumo de refrigerantes tipo cola está relacionado com a incidência de câncer de pulmão.

A quantidade de conhecimento científico disponível na *Internet* vem se tornando tão vasta que só poderá ser processada com a ajuda de computadores. É proposto aqui um padrão para representar este conhecimento, de modo a viabilizar seu processamento por programas. A Ciência da Informação pode criar metodologias que irão mais adiante do que simplesmente prover acesso rápido ao texto completo de artigos científicos publicados na Web. O modelo proposto pode ajudar cientistas a processar diretamente o conhecimento contido no texto de artigos científicos, a validar a consistência das afirmações contidas num artigo, a recuperar a linha de raciocínio que levou a uma descoberta. O modelo também aponta para a criação de um novo padrão, uma LMCC – Linguagem de Marcação do Conhecimento Científico –, contendo o conteúdo semântico de artigos científicos publicados na Web. Este artigo destaca os benefícios de um formato semanticamente rico para representar o conteúdo de artigos científicos. Com a ajuda de ferramentas de "*software*" apropriadas, esse conhecimento pode ser extraído como um subproduto do processo de editoração/auto-publicação eletrônica de um artigo pelo autor. Isso abre novas perspectivas para a aquisição, processamento e compartilhamento do conhecimento científico.

Referências

ALVES, R. *Filosofia da ciência*: introdução ao jogo e suas regra. São Paulo: Brasiliense, 1987.

[15] Centro Cochrane do Brasil, Medicina baseada em evidências. Disponível em: <http://www.centrocochranedobrasil.org.br/mbe.asp>. Acesso em: 6 nov. 2008.

ALVES MAZZOTTI, A.; GEWANDSZNAJDER, F. *O Método nas ciências naturais e sociais*: pesquisa quantitativa e qualitativa. São Paulo: Pioneira Thomson Learning, 2002.

BACON, F. *Novum organum*. São Paulo: Abril Cultural, 1973. (Coleção Os Pensadores, 13).

BARRETO, A. A. A oferta e a demanda da informação: condições técnicas, econômicas e políticas**.** *Ciência da Informação*, Brasília**,** v. 28, n. 2, maio/ago. 1999. p.168-142. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19651998000200003&lng=pt&nrm=iso>. Acesso em: 18 jun. 2005.

BAKER, C. J.; CHEUNG, K. (Eds.). Semantic Web. *Revolutionizing knowledge discovery in the Life Sciences*. New York (USA): Springer, 2007.

BAYLEY, L; ELDREDGE, J. The structured abstract: an essential tool for researchers. *Hypothesis 3 Spring*; v.17, n.1, p. 11-13, 2003. Disponível em: <http://gain.mercer.edu/mla/research/hypothesis.html>. Acesso em: 22 jul. 2007.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. *Scientific American*, May, 2001. Disponível em: <http://www.scian.com/2001/0501issue/0501berners-lee.html>. Acesso em: 24 maio 2001.

BREITMAN, K. *Web Semântica*: a internet do futuro. Rio de Janeiro: LTC, 2005.

BROOKES, B. The foundations of Information Science: Part I: philosophical aspects. J*ournal of Information Science,* v. 2, p.125-133, 1980.

BUCKLAND, M. *Information and information systems*. Westport (CT): Praeger/Greenwood, 1991.

CAMARA, G. N. L. et al. Prevalence of human papillomavirus types in women with pre-neoplastic and neoplastic cervical lesions in the Federal District of Brazil. *Mem. Inst. Oswaldo Cruz*, v.98, n. 7, p. 879-883, 2003.

CAMPOS, M. L.A. Modelização de domínios de conhecimento: uma investigação dos princípios fundamentais. *Ciência da Informação*, v. 33, n. 1, p. 22-32, jan./abr. 2004.

CASTELLS, M. *A sociedade em rede*. São Paulo: Paz e Terra, 1999. v. 1.

DAHLBERG, I. *Ontical structures and universal classification*. Bangalore: Sarada Ranganthan Endowment, 1978. 64 p.

DE BRUIJN, J. et al. Ontology Mediation, Merging and Aligning. In: DAVIES, J.; STUDER, R.; WARREN, P. *Semantic Web Tenologies*: trends and research in ontology-based system. West Sussex (UK): John Wiley, 2006.

DE ROURE, D.; JENNINGS, N.; SHADBOLT, N. *Research agenda for the Semantic Grid:* a future s-Science infraestructure. [s.l.: s.n.], 2001. (Report commissioned for EPSRC/DTI Core e-Science Programme).

DING, Yin; FOO, S. Ontology research and development. Part 1 - a review of ontology mapping and evolving. *Journal of Information Science*, v.28, n.10, p. 123 – 136, 2002a.

DING, Y.; FOO, S. Ontology research and development. Part 2 - a review of ontology generation. *Journal of Information Science*, v.28, n.4, p. 375-388, 2002b.

ELLIS, D. Paradigms and proto-paradigms in information retrieval research. In: VAKKARI, P.; CRONIN, B. (Eds.). *Conceptions of library and Information Science:* historical, empirical and theoretical perspectives. London: [s.n.], 1992. p.165-186.

FARRADANE, J. Relational Indexing. Part I. *Journal of Information Science*, v.1, p.267-276, 1980.

GONZÁLEZ DE GOMEZ, M. N. O papel do conhecimento e da informação nas formações políticas ocidentais. *Ciência da Informação*, Brasília, v.16, n.2, p.157-167. jul./dez., 1987.

GROSS, A. G. *The Rhetoric of Science*. Cambridge; Londres: Harvard University Press,1990.

HOFFMANN, M. Is there a "Logic" of Abduction? In: CONGRESS OF THE IASS – AIS, 6. International Association for Semiotics Studies, Guadalajara, Mexico, 1997. *Proceeedings…* Disponível em: <http://www.unibielefeld.de/idm/personen/mhoffman/papers/abduction-logic.html>. Acesso em: 14 dez. 2005.

HJØRLAND, B. Towards a theory of aboutness, subject, topicality, theme, domain, filed, content… and relevance. *Journal of the American Society for Information Science and Technology*, v. 52, n. 2, p. 774-778, 2001.

_____. Epistemology and the socio-cognitive perspective in information science. *Journal of the American Society for Information Science and Technology,* v. 53, n.4, p. 257-270, 2002a.

_____. Domain analysis in information science: eleven approaches – traditional as well as innovative. *Journal of Documentation*, v.58, n.4, p. 422 – 462, 2002b.

_____. *Units or entities in knowledge organization (KO)*: what is being rganized? 2007. Disponível em: <http://www.db.dk/bh/lifeboat_ko/HISTORY%20&%20THEORY/units_in_knowledge_organization.htm>. Acesso em: 4 ago. 2007.

HUTCHINS, J. On the structure of scientific texts. *Papers in Linguistics*, Norwich, UK: University of East Anglia, 1977. p. 18-39. Disponível em:

<http://ourworld.compuserve.com/homepages/wjhutchins/UEAP/L-1977.pdf>. Acesso em:20 mar. 2006.

IONNIDIS, J. at al. A road map for efficient and reliable human genome epidemiology. *Nature Genetics,* v.38,p.3-5, 2006.

JACOB, E. K. Ontologies and the Semantic Web. *Bulletin of the American Society for Information Science and Technology*, v. 29, n. 4, p. 19-22, April/May, 2003.

KUHN, T. S. *A estrutura das revoluções científicas.* São Paulo: Perspectiva, 2003. (Série Debates Ciência).

LANCASTER, F. W. *Indexação e resumo*: teoria e prática. Brasília: Briquet de Lemos, 1993.

MAGNANI, L. *Abduction, reason, and scienc*e: processes of discovery and explanation. New York: Kluwer Academic; Plenun Publishers, 2001.

MARCONDES, C. H. From scientific communication to public knowledge: the scientific article Web published as a knowledge base. In: ICCC ELPUB - INTERNATIONAL CONFERENCE ON ELECTRONIC PUBLISHING, 9., Leuven, Bélgica, 2005. *Proceedings*... Leuven, Bélgica, 2005. p.119-27. Disponível em <http://elpub.scix.net>. Acesso em: 28 ago. 2006.

MARCONDES, C. H. Representação e economia da informação. *Ciência da Informação***,** Brasília, v. 30, n. 1, p. 61-70, 2001.

MEADOWS, A. J. *A comunicação científica*. Brasília: Briquet de Lemos, 1999.

MILLER, D. L. Explanation versus description. *Philosophical Review*, v.56, n.3, p. 306-312, 1947.

OLIVA, A. Kuhn: o normal e o revolucionário na reprodução da racionalidade científica. In: PORTOCARRERO, V. (Org). *Filosofia, história e sociologia das ciências*. Rio de Janeiro: Ed. FIOCRUZ, 1994. p. 67-102.

OTLET, P. *Traité de Documentation***:** le livre sur le livre. Théorie et pratique. Liège: Centre de Lecture Publique de la Communauté Française de Belgique, 1989. 432 p.

PAAVOLA, S. Abduction as a Logic and methodology of discovery: the importance of strategies. *Foundations of Science*, v.9, n. 3. Nov., p. 267-283, 2004.

RANGANATHAN, S. R. *Prolegomena to Library Classification.* New York: Asia Publishing House, 1967.

PIERCE, C. S. *Semiótica*. São Paulo: Perspectiva, 1977. (Série Estudos).

SCHAFFE, F.; WEITEN, M. Ontology Mediation, Merging and Aligning. In: DAVIES, J.; STUDER, R.; WARREN, P. *Semantic Web Tecnologies*: trends and research in ontology-based system. West Sussex (UK) : John Wiley, 2006.

SHETH, A; ARPINAR, I. B.; KHASYAP, V. Relationships at the heart of semantic web: modeling, discovering and exploiting complex semantic relationships. In: NIKRAVESH, M. et al. *Enhancing the power of the internet studies in fuzziness and soft computing*. [s.l.: s.n.], 2003. Disponível em: <http://cgsb2.nlm.nih.gov/~kashyap/publications/relations.pdf>. Acesso em: 18 set. 2006.

SORGEL, D. The rise of ontologies or the reinvention of classification. *Journal of the American Society for Information Science*, v.5, n.12, p.1119-1120, 2000.

THE SEMANTIC WEB. Bulletin of The American Society for Information Science and Technology, v. 29, n. 4, Apr./May 2003. (Special Section).

VICKERY, B. C. Knowledge Representation: a brief review. *Journal of Documentation*, v.42, n.3, p.145-59, 1986.

ZIMAN, J. *Conhecimento público*. São Paulo: Ed. da Universidade de São Paulo, 1979.