ORGANIZAÇÃO E REPRESENTAÇÃO DO CONHECIMENTO CIENTÍFICO EM AMBIENTE **WEB:** do formato textual linear aos artigos semânticos

RESUMO - Artigos científicos são ainda hoje publicados eletronicamente segundo o formato textual-linear, cópia digital do formato impresso. Este formato impede que programas possam ser usados para o processamento "semântico" desses conteúdos. O trabalho discute a emergência dos artigos científicos semânticos, utilizando tecnologias da Web Semântica, suas motivações, potencialidades e conseqüências para gestão do conhecimento científico. São levantados requisitos funcionais para artigos semânticos. É apresentado um modelo que serve para ilustrar o atendimento a estes requisitos e suas potencialidades.

Palavras-chave: artigos semânticos; conhecimento científico; representação do conhecimento; organização do conhecimento; gestão do conhecimento, ciência eletrônica.

Carlos Henrique Marcondes
Doutor em Ciência da Informação, pesquisador do CNPq, professor do Departamento de Ciência da Informação e Programa de Pós-graduação em Ciência da Informação, UFF - Universidade Federal Fluminense.

marcon@vm.uff.br

SCIENTIFIC KNOWLEDGE ORGANIZATION AND REPRESENTATION IN THE WEB ENVIRONMENT: FROM THE TEXTUAL-LINEAR FORMAT TO SEMANTIC ARTICLES

ABSTRACT – Scientific articles today are still electronic published according to the textual linear format, a copy of the print format. This prevents the use of programs to semantic processing the content of these articles. The paper discusses the emergence of the scientific semantic articles which use Semantic Web technologies, its motivations, potential and consequences to scientific knowledge management. Functional requirements to semantic articles are posed and a model is proposed which illustrates the fulfillment of these requirements and the potentialities of semantic articles.

Key-words: semantic articles; scientific knowledge; knowledge representation; knowledge organization; knowledge management; e-science.

1. INTRODUÇÃO

Antes do surgimento da Web o acervo de conhecimento da humanidade era armazenado de forma descentralizada, nos acervos das bibliotecas. Apesar da unipresença da Web, das facilidades de acesso imediato a qualquer momento, grande parte dos documentos utilizados pela humanidade ainda hoje, inclusive artigos científicos, mesmo em suas versões digitais criadas e transferidas através da Web, são documentos em formato textual linear, não estruturados como os registros de uma base de dados. As tecnologias básicas da Web atual, o formato HTML para formatação de documentos, o protocolo HTTP para transmiti-los e acessá-los, são tecnologias de *apresentação*, para tornar documentos digitais legíveis por pessoas, o que dificulta a tarefa de agenciar programas para tratar este material, de organizá-lo e gerenciar este conhecimento aí contido com vistas ao acesso, uso e reuso.

Desde a criação do periódico pioneiro "Philosophical Transactions" da "The Royal Society" inglesa em 1665, que o artigo cientifico vem assumindo papel essencial nos quadros institucionais-econômicos da ciência, cumprindo importantes papeis: é o principal registro do conhecimento cientifico, permitindo que fenômenos desconhecidos até então sejam apropriados, validados e integrados ao conhecimento cientifico pré-existente. Exerce ainda papeis, não menos importantes, de indicador da atividade cientifica, tanto a nível coletivo, institucional, quanto individual, e de indicador de mérito do pesquisador por descobertas realizadas.

BJÖRK et al. (2009) estimam que, somente no ano de 2006 foram publicados 1.346.000 artigos em 25.750 periódicos científicos com revisão por pares. Na área Biomédica, bibliotecas digitais como PubMed[1] contém hoje cerca de 22 milhões e 500 mil referências, que crescem à razão cerca de 500 mil por ano. Neste contexto, um grande esforço científico é a busca por melhores ferramentas computacionais que permitam aos cientistas tomar conhecimento, ler e se apropriar do conteúdo deste volume massivo e

[1] http://www.ncbi.nlm.nih.gov/pubmed/.

crescente de informações.

Tenopir et al. (2008) relatam em seus estudos que, como conseqüência, entre outras causas, deste crescimento do número de artigos publicados, cada pesquisador tem aumentado o número de artigos que tem que ler anualmente, simultaneamente a diminuição do tempo despedido com a leitura de cada artigo. Na área científica, em especial nas ciências biomédicas, esta questão assume proporções cada vez mais críticas. Attwood et al. (2009) reclamam da premência de ferramentas computacionais para encontrar conhecimento "perdido" no excesso de literatura científica publicada eletrônicamente. Projetos e instituições de pesquisa chamam a atenção para esta problemática, como pode ser constatado pelos exemplos a seguir:

O projeto **Semantic and Services-enabled Problem Solving Environment for *Trypanossoma Cruzi***, um projeto internacional para criar um ambiente Web integrado de acesso à diferentes recursos e fontes de conhecimento sobre o Trypanossa Cruzi, http://knoesis.wright.edu/research/semsci/application_domain/sem_life_sci/tcruzi_pse/, quando coloca como uma de suas linhas de ação:

> "Semantic text analysis approaches for extraction of knowledge from biomedical literature - Biomedical literature, for example Pubmed, represents a vast and valuable resource for life sciences research. The ability to extract relevant knowledge from biomedical text and its representation in Semantic Web standard formats such as RDF is an important research issue that is being addressed in this project."
> (http://www.bioontology.org/videos/PSE_talk.html).

A **ICBO: International Conference on Biomedical Ontology**, realizada na University at Buffalo, NY, em julho de 2009, com o patrocínio da University at Buffalo College of Arts and Sciences, da National Center for Ontological Research National Center for Biomedical Ontology e da Science Commons, quando coloca entre seus principais focos discutir "*The role of ontology in the future of scientific publishing*".

Uma crítica aos métodos bibliométircos e cientométricos é que: “they do not take into account the semantic content of scientific publications” (Niiniluoto, 2002); o mesmo afirma Latour (2000). A indexação dos artigos é feita por profissionais quando estes são incluídos em bancos de dados e repositórios como Medline e PubMed, e não pelos próprios autores, que sabem mais que ninguém, a importância do que esta sendo reportado em seus artigos. As deficiências de recuperação dos SRIs atuais e os métodos bibliométricos e cientométricos de gestão do conhecimento científico acarretam, como assinala Van Haan (2004), que muitos artigos reportando importantes descobertas científicas não sejam citados por muitos anos a fio após serem publicados; são conhecidos como “sleeping beauties” da ciência.

Novas áreas emergentes de pesquisa surgem para endereçar esta questão, como “Literature-based discovery” (SWANSON et al., 2006), (KOSTOFF et al., 2008), “text mining” (BATH, 2002) se organizam para endereçar esta questão. Toda esta problemática seria grandemente facilitada se artigos científicos digitais tivessem formatos mais suscetíveis de serem tratados por programas.

Desde fins da década de 1980 inicia-se uma evolução em direção ao que seriam hoje artigos semânticos. Gardin (1987; (2001), autor pioneiro, propõe a “escrita logicista”, formalismo para a estruturação do texto de artigos que permitiria evidenciar a estrutura lógica do raciocínio do autor; Murray-Rust e Rzepa, (1999, 2002) propõe, desde fins da década de 90, o uso de linguagens baseadas em XML para estruturar o texto de artigos científicos, o que hoje é uma prática comum e abre o caminho para acessar partes específicas do texto de artigos; com o surgimento da Web Semântica, Marcondes (2005), seguindo as propostas de Murray-Rust/Rzepa e Gardim, propõe um modelo da estrutura de elementos semânticos do artigo, como *questão, hipótese, experimento, resultados* e *conclusão*, codificado em XML; Hunter (2006), inspirada em formatos de metadados para objetos digitais complexos, como METS, propõe artigos científicos como “pacotes”

integrados; (SHUTTON, 2009), propõe uma visão abrangente da problemática das publicações semânticas e seus desafios.

O objetivo deste trabalho é discutir os artigos semânticos e as novas possibilidades para gestão do conhecimento científico que transcendem as possibilidades do atual formato textual-linear dos artigos, cópia eletrônica do formato impresso, como unidade e formato de representação e organização do conhecimento cientifico. Para isto é mostrado um panorama introdutório da emergência das publicações científicas semânticas, as inter-relações do tema e delineando um conjunto de requisitos funcionais para artigos semânticos. O atendimento a estes requisitos é ilustrado a partir da discussão da proposta de um modelo semântico de artigos científicos (MARCONDES, 2005).

A hipótese do trabalho é que as necessidades de leitura otimizada por parte dos pesquisadores das ciências biomédicas, juntamente com a emergência das tecnologias da Web Semântica e das facilidades oferecidas pelo ambiente digital, propiciam que artigos científicos estejam evoluindo no sentido de se tornarem objetos digitais complexos, incorporando "semântica" computacional (MARCONDES, 2012), incluindo, além de texto, comentários, "links" para "datasets", assertivas lógicas relativos ao conteúdo, as referencias citadas, etc.; estes "artigos semânticos" poderiam ser recuperados de forma semanticamente mais eficiente usando as tecnologias e padrões da Web Semântica (BERNERS-LEE et al., 2001), funcionariam como bases de dados, podendo ter seu conteúdo consultado e interligado.

O trabalho esta organizado da seguinte maneira: após esta Introdução, a seção 2 apresenta os materiais e métodos utilizados na pesquisa; a seção 3 discute o cenário atual das publicações científicas eletrônicas seus desafios; a seção 4 propõe um modelo semântico de artigo científico; a seção 5 discute as potencialidades deste modelo; a seção 6 apresenta considerações finais.

2. MATERIAIS E MÉTODOS

O trabalho é desenvolvido como uma revisão, para permitir uma visão do estado da arte da questão dos artigos semânticos. Como será visto a partir dos autores citados, o tema da emergência dos artigos científicos digitais semânticos nas Ciências Biomédicas tem inter-relações com temas como Filosofia da Ciência, Metodologia Científica, Lógica, Retórica, Comunicação Científica, Ciência da Computação, Ontologia Computacional, Bioinformática, Lingüística Computacional, entre outros. Foi com base em aportes teóricos da literatura destas áreas que o modelo proposto foi concebido.

O modelo foi validado e reformulado a partir da análise de 89 artigos das ciências biomédicas, conforme descrito em artigos anteriores (MARCONDES, 2011, 2011b, 2012). A análise utilizou o UMLS, uma grande e amplamente usada base terminológica no domínio da biomedicina.

3. O CENÁRIO ATUAL DAS PUBLICAÇÕES CIENTÍFICAS ELETRÔNICAS

Nesta seção vai-se procurar caracterizar, na forma de subseções, a emergência, modelos adotados, o cenário atual das publicações científicas eletrônicas, dos SRIs e seu uso pela comunidade acadêmica, em especial, das ciências biomédicas, e delinear requisitos funcionais para artigos semânticos.

3.1. Emergência e limitações das atuais publicações eletrônicas

O lançamento do “the Online Journal of Current Clinical Trials”, em 1992 (RENEAR e PALMER, 2009, p. 828) marca o surgimento das publicações eletrônicas acadêmicas. Desde de então os periódicos eletrônicos evoluíram para se tornar o modo corrente de publicação científica. De um modo geral, assumiram o modelo de disseminação vigente nas publicações impressas, o artigo textual linear em formato HTML ou PDF. Neste, novas funcionalidades acrescidas ao modelo de artigo impresso em papel limitaram-se o uso de

“links” para fazer referência a outros artigos ou recursos disponíveis na Web.

Como o estudo de Tenopir et al. (2008) demonstram, a comunidade acadêmica vem aumentando as demandas por mais precisão na seleção dos artigos a serem lidos, que é cada vez menos atendida pelos SRIs convencionais. Acesso por conteúdo a documentos nestes SRIs, incluindo bibliotecas digitais, repositórios, sistemas de publicação de periódicos, ainda é feito por comparação de palavras-chave da consulta feita pelos usuários, unidas através de pouco expressivos operadores booleanos, com palavras-chave que compõe os registros bibliográficos, de maneira semelhante aos primeiros sistemas de recuperação bibliográfica e de automação de biblioteca. Relações expressas por operadores booleanos são processados pelos SRIs como relações extensivas entre conjuntos de documentos que contém determinada palavra-chave e não como relações intensivas entre conceitos. Operadores booleanos não dão conta da expressividade e precisão necessária para a recuperação de conteúdo semântico contido no crescente número de artigos científicos e outras fontes de informação agora disponíveis em toda a Web; são genéricos e se ressentem de expressividade semântica necessária à recuperação de conteúdos em domínios científicos específicos como Biomedicina. Numa busca por “políticas para lidar com AIDS” no PubMed foi recuperado um artigo com o título “A statewide observational assessment of the pedestrian and bicycling environment in hawaii, 2010”, PMID-22172181, que trata de políticas de transito, incluindo “street accommodations (ie, sidewalks and crossing **aids**)”.

Na área biomédica são utilizados crescentemente novos SRIs para identificar na quantidade de literatura publicada, entidades específicas e, em especial, *relações* entre elas, como os sistemas iHOP – Information Hyperlinked over Proteins, que recupera relações genes/proteínas em frases retiradas de resumos do MEDLINE, ou Textpresso, que recupera relações de associação, comparação, co-ocorrência, envolvimento, regulação, contigüidade espacial ou seqüência temporal, entre entidades biomédicas.

Técnicas de mineração de textos também são utilizadas intensivamente para

identificar em resumos e textos de artigos biomédicos relações que são significativas neste domínio; além das relações genes-proteínas, também relações gene-gene, gene-doença, droga-doença (TANABE et al. 1999), (SPASIC et al. 2005), (ZHANG, Y. et al., 2011).

Além dos recursos bibliográficos tradicionais, outras ferramentas para tratamento de dados biomédicos passam a estar disponíveis; exemplos são BLAST, sigla para "Basic Local Alignment Search Tool", um sistema de recuperação de informações que permite entrar com uma seqüência genômica e recuperar seqüências semelhantes em bancos de dados genômicos; o PhenomicDB, um banco de dados/SRI que recupera relações phenótípo-genótipo entre entidades biomédicas. A idéia de "minerar textos" recuperando informações não pela coincidência de palavras-chave isoladas, mas sim por relações entre entidades biomédicas esta presente em muitas experiências (HUNTER et al., 2008).

Exemplos como os citados apontam para um processo ainda mais interessante, a integração de recursos SRI bibliográficos com outros recursos como os mencionados anteriormente. Exemplos são o próprio BLAST, integrado ao SRI do PubMed, permitindo que um uma seqüência genômica mencionada em um artigo recuperado possa ser imediatamente submetida ao BLAST; outro exemplo é o projeto Prospect, dos periódicos científicos publicados pela Royal Society Of Chemistry do Reino Unido, no qual o termos no texto de artigos são identificados, padronizados e marcados como "links", com base em ontologias e terminologias biomédicas, podendo estes "links" serem acionados pelo leitor para acessar definições ou relações entre estes termos. Esta integração é permitida pelas facilidades hoje oferecidas pelo ambiente Web.

É este ambiente Web integrado que vêm permitindo aos cientistas exercitarem cada vez mais as práticas de "leitura estratégica", conforme caracterizado por Renear e Palmer (2009, p. 828):

> Scientists have always read strategically, working with many articles simultaneously to search, filter, scan, link, annotate, and analyze fragments of content. An observed recent increase in strategic reading in the online environment will soon be further intensified by two current trends: (i) the

> widespread use of digital indexing, retrieval, and navigation resources and (ii) the emergence within many scientific disciplines of interoperable ontologies. Accelerated and enhanced by reading tools that take advantage of ontologies, reading practices will become even more rapid and indirect, transforming the ways in which scientists engage the literature and shaping the evolution of scientific publishing.

As práticas de "leitura estratégicas" apontam também para a necessidade de desenvolvimento do próprio formato do texto dos artigos, visando mais objetividade na sua leitura. Tendências nesta direção são uso de seções padronizadas nos artigos biomédicos (modelo IMRAD-Introduction, Method, Results and Discussion[2]); a adoção de resumos estruturados por um número cada vez maior de periódicos biomédicos e sua padronização pelo Medline, com as seções: BACKGROUND, OBJECTIVE, METHODS, RESULTS and CONCLUSIONS[3] é outra iniciativa nesta direção. O uso cada vez mais generalizado de resumos estruturados nos artigos biomédicos aponta para a necessidade do pesquisador acessar diretamente elementos significativos do conteúdo do artigo.

3.2. O Projeto da Web Semântica

O Projeto da Web Semântica (BERNERS-LEE, et al., 2001) abre perspectivas para novas formas de organizar e representar conhecimento na Web. Se propõe endereçar o problema do excesso de informações pela criação de padrões para conteúdos que possam ser "inteligíveis" por máquinas. Nos SOCs – Sistemas de Organização do Conhecimento - atuais, como sistemas de catálogos em arquivos, bibliotecas e repositórios digitais, registros são constituídos de listas de campos e de palavras-chave isoladas; documentos, apesar de digitais, são ainda calcados no modelo textual linear, para leitura por pessoas. Padrões utilizados, como o MARC e ISAD((G), são antigos e exclusivos, não saõ interoperáveis com os novos padrões surgidos com o projeto da Web Semântica. Além disso, metadados e documentos, como um registro MARC referenciando um documento

[2] Ver International Committee of Medical Journals Editors, http://www.icmje.org.

[3] Ver http://www.nlm.nih.gov/bsd/policy/structured_abstracts.html.

digital num sistema de catálogo, são mantidos “prisioneiros” destes sistemas, isolados do resto da Web, só adquirindo significado dentro do contexto destes SOCs, quando são armazenados, recuperados e exibidos.

No contexto do projeto da Web Semântica foi proposto um conjunto de padrões para estruturar metadados e conteúdos, tendo como base a linguagem XML e buscando expressar uma “semântica” computacional. Nas palavras de Berneers-Lee (2001): “In short, XML allows users to add arbitrary structure to their documents but says nothing about what the structures mean. Meaning is expressed by RDF**,** which encodes it in sets of triples, each triple being rather like the subject, verb and object of an elementary sentence”.

Sobre a base constituída pela XML, RDF permite fazer afirmações, formadas por sujeito, predicado e objeto, interligando uma rede de recursos Web. Além do significado intrínseco do sujeito, do predicado e do objeto, uma afirmação pode ter seu significado ampliado, pelo uso de vocabulários específicos disponíveis também na Web, que especificam ainda mais este significado; estes vocabulários são declarados como “name spaces” dentro de um documento RDF (RDF Primer, 2004); exemplos podem ser vistos na Figura 2, na seção 5.

Esse recursos Web podem ser referenciados univocamente em todo o espaço da Web através de outro padrão, os URIs – Uniform Resource Identifier (RFC 2396, 1998) -, capazes de identificar e interligar de forma persistente quaisquer recursos Web.

Conjuntos de afirmações RDF podem ser organizados em ontologias computacionais, expressas na linguagem OWL. Estas incluem incluindo relações classe-subclasse, formando uma ou várias taxonomias, relações todo-parte ou relações funcionais entre diferentes taxonomias, além de regras que especificam a validade de qualquer uma destas relações.

Triplas RDF podem ser armazenadas em bancos de dados e consultadas através da

linguagem SPARQL, transformando assim esta Web em rede interligada – denominada "Linked Open Data" (BIZER; HEALTH; BERNERS-LEE) - numa base de dados totalmente consultável.

3.3. Requisitos funcionais para artigos semânticos

Neste cenário se coloca a demanda por um novo tipo de publicações eletrônicas que possam tirar partido de todas as facilidades oferecidas pelas tecnologias da Web Semântica. O uso das tecnologias descritas permite que artigos possam ter o seu conteúdo tratado por computadores em aplicações que demandem "compreensão" do seu conteúdo (MARCONDES, 2012) superando as imprecisões da linguagem natural e ter termos em seu texto identificados, padronizados e marcados segundo terminologias/ontologias, de modo a servir de "links" para outros termos em outros recursos em bancos de fórmulas de substâncias biomédicas ou de sequenciamento genético; possam também ter elementos semânticos essenciais de seu conteúdo, como objetivo, problema, questões, hipóteses, metodologia, resultados e conclusões, identificados e interligados entre si ou com elementos semelhantes de outros artigos por cadeias de raciocínio/inferência; elementos específicos do conteúdo de artigos semânticos também poderão ser consultados diretamente, como se o artigo fosse uma base de conhecimento.

A seguir são apresentados alguns exemplos de sistemas que avançam neste sentido:

- O Scholarly Ontology Project (SHUM et al, 2003) usa uma ontologia para extrair e estruturar formalmente hipóteses contidas num artigo científico – chamadas de "claims" - e relacioná-las a outros artigos. Estas relações podem ser, por exemplo, "concorda", "discorda", "é evidência a favor", "prova", "refuta", etc.

- Hybrow (RACUNAS et al., 2004), é um sistema para apoio à avaliação de hipóteses no domínio da bioinformática;

- as ferramentas disponibilizadas pelo National Institute of Health, EUA, para tratamento semântico de textos, denominadas Semantic Knowledge Representation Project, http://skr.nlm.nih.gov/.

- MachineProse (DINAKARPADIAN et al. 2006), um sistema que formaliza assertivas

científicas, como hipóteses, com base nos tipos semânticos e relações providos pelo UMLS;

- a ontologia EXPO (SOLDATOVA; KING, 2006) formaliza os elementos de um experimento científico, permitindo com eles anotar artigos;

- SWAN (GAO et al., 2006) é um ambiente Web para a comunidade acadêmica que pesquisa a doença de Alzheimer; inclui um modelo ontológico, ferramentas para suporte à organização de dados científicos, geração e teste de hipóteses científicas e colaboração entre pesquisadores.

- Exemplos do uso da ontologia CITO, usada no periódico PLoS, para formalizar os motivos de uma citação, dentro de um artigos (SHUTTON et al., 2009).

4. REPRESENTAÇÃO E ORGANIZAÇÃO DO CONHECIMENTO CIENTÍFICO EM ARTIGOS SEMÂNTICOS

Trabalhamos há anos (MARCONDES, 2005) na proposta de um modelo semântico de publicações eletrônicas, que tem como objetivo extrair e representar o conteúdo de artigos científicos biomédicos em formato "inteligível" por programas, de modo a permitir que estes realizem "inferências" sobre este conhecimento, permitindo processar o conhecimento assim recuperado de forma semanticamente mais rica que os atuais SRIs.

Este modelo é descrito a seguir e se apóia em um leque de conceitos, formalismos, "insights" e propostos que abrangem temas como metodologia científica, paradigma científico e raciocínio científico (POPPER, 2001, KUHN, 2003, MAGNANI, 2001, THARGARD 1993, e KLAHR e SIMON, 1999), estrutura profunda, ou semântica, da linguagem (CHOMSKY, 1981), de microestrutura, macroestrutura e superestrutura, (KINTSH, VAN DIJK, 1972), na estrutura retórica e conceitual em geral e de artigos científicos especificamente (HUTCHINS, 1977), BEZERMAN (1988), (GROSS, 1990), (SWALES, 1990), (NWOGU, 1997), (KANDO, 1997, 1999), (FRANKLIN, 2004), (DE WARD, 2009).

em especial em linguagens e sistemas de indexação e recuperação de informação vislumbrou as relações como chave para a representação de significados. Farradane's (1980) na proposta de Indexação Relacional, afirma que: "Meaning, considered as relations between terms…". De acordo com Brookes (1980) "knowledge is a structure of concepts linked by their relations and information is a small part of such a structure". Sheth et al. (2003) afirmam que "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". Miller (1947) afirma que: 'The above remarks imply that science is a search after internal relations between phenomena'.

No modelo proposto o conhecimento científico consiste em afirmações encontradas em *elementos semânticos* chave do texto do artigo, nas quais são identificadas novas relações entre fenômenos, ou entre um fenômeno e suas características. Ao descobrir, colecionar e estabelecer as características de um fenômeno até então desconhecido, este é integrado ao esquema conceitual-classificatório de um domínio científico (DAHLBERG, 1978). Um fenômeno pode ser definido como um 'perceptible fact, a sensible occurrence' (BUNGE, 1998). Fenômenos são aquilo que os cientistas observam, medem e comparam. Para raciocinar sobre fenômenos observados/medidos/comparados, cientistas usam conceitos como a unidade básica de conhecimento científico. Conceitos representam o acordo básico da comunidade científica acerca de fenômenos e são expressos linguisticamente através de termos reunidos em corpus terminológicos, especializados como UMLS e Gene Ontology. Autores fazem afirmações científicas sobre fenômenos observados no texto dos artigos que escrevem, as quais têm a forma de relações entre conceitos.

Relações são, portanto o elemento essencial do esquema de representação do conhecimento proposto. São expressas por três elementos: dois *relata* e um *tipo de relação*. Os dois *relata* – Antecedente e Conseqüente - podem ser: dois fenômenos científicos distintos ou um fenômeno científico e alguma de suas características. O *tipo de*

relação guarda a semântica da relação, por exemplo, causa-efeito, sintoma-doença, método-o que é viabilizado pelo método, etc. As afirmações feitas pelo autor no artigo são representadas como <Antecedente><Tipo de Relação><Conseqüente>. Por exemplo:

- Papiloma Vírus Humano (Antecedente, um fenômeno) causa (tipo de relação) Câncer de Colo do Útero (Consequente, outro fenômeno);

- Encurtamento dos telômeros (Antecedente, um fenômeno) esta associado a (tipo de relação) senescência celular (Conseqüente, outro fenômeno).

- Extremidade dos telômeros (Antecedente, um fenômeno) tem como composição molecular (tipo de relação) 'TTGGG'(Consequente, uma característica do fenômeno expresso pelo Antecedente).

Relações podem aparecer em diferentes elementos semânticos do texto do artigo: no Problema como uma **Questão** – algum dos *relata* ou o tipo de relação são desconhecidos -, como por exemplo: "*To understand the structure of telomerase RNA in vertebrates*" (CHEN, 2000) ou "*we wished to determine whether variation in initial telomere length would account for the unexplained variation in replicative capacity*" (ALLSOPP, 1992) ou "*How could telomeres be involved in nuclear and cell division?*" (GUO-LIANG, 1990). Na **Hipótese**, expressando naturalmente uma relação ainda hipotética, como por exemplo "*we propose that the novel terminal transferase-like activity in the Tetrahymena extracts is involved in the novo elongation step of telomere replication*" (GREIDER, 1985). Nos **Resultados** ou nas **Conclusões**, expressando uma relação validada por um experimento, como por exemplo "*The runaway telomere mutants obtained by altering telomeric DNA sequences have showed that negative telomere-length regulation is associated with optimal cell viability*" (MCEACHERN, 1995). Frequentemente a Conclusão de um artigo também coloca novas **Questões**, como "*the RNA component of telomerase may be directly involved in recognizing the unique three-dimensional structure of the G-rich telomeric oligonucleotide primers*" (GREIDER, 1987).

A análise feita permitiu identificar os seguintes *elementos semânticos*:

Um PROBLEMA expressa uma carência, insatisfação ou deficiência conceitual com o atual estado de conhecimento num domínio. Um PROBLEMA pode se desdobrar em OBJETIVOS de pesquisa e, eventualmente, na formulação mais precisa de uma QUESTÃO que endereça a deficiência conceitual; esta QUESTÃO pode ser referir a um FENÔMENO (nos artigos EXPLORATÓRIOS), ou a dois ou mais FENÔMENOS envolvidos numa RELAÇÃO_ENTRE_FENÔMENOS ou HIPÓTESE. Uma HIPÓTESE relaciona dois ou mais FENÔMENOS através de um TIPO-DE-RELAÇÃO.

Um autor num artigo pode formular uma hipótese original – HIPÓTESE(o) ou tomar a hipótese prévia – HIPÓTESE(p) - de outros autores; neste caso uma ou mais citações referentes à HIPÓTESE(p) – CITAÇÕES(h) - são feitas. Um autor também pode analisar várias HIPÓTESEs(p) para mostrar que elas são insatisfatórias como soluções para o PROBLEMA e formular sua hipótese original - 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 tornar o fenômeno observável empiricamente. Em um artigo científico EXPERIMENTAL, significa ter RESULTADOS observados segundo determinada MEDIDA, em determinado CONTEXTO segundo determinada METODOLOGIA. Este CONTEXTO onde os FENÔMENO(s) 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. Todo artigo também traz uma CONCLUSÃO, na forma de uma proposição sobre um fenômeno ou sobre RELAÇÕES_ENTRE_FENÔMENOS.

Estes elementos semânticos, não se apresentam de forma uniforme em diferentes artigos. Outro resultado obtido é a proposta de uma *tipologia de artigos*, elaborada a partir de propostas como as de (KINTSH, VAN DIJK, 1972), (HUTCHINS, 1977), BEZERMAN (1988), (GROSS, 1990), (SWALES, 1990), (NWOGU, 1997), (KANDO, 1997, 1999),

(FRANKLIN, 2004).

Os diferentes tipos de artigos identificados – Teóricos, Experimentais-exploratórios, Experimentais-indutivos e Experimentais-dedutivos – organizam os elementos semânticos em diferentes padrões de encadeamento, que expressam diferentes raciocínios, estratégias de argumentação e pressupostos para se chegar às conclusões de um artigo. A tipologia de artigos inicialmente proposta, com base na literatura citada, foi sendo reformulada/aperfeiçoada a partir da análise dos artigos.

Artigos teórico-abdutivos 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). Sua contribuição é uma nova hipótese, indicando um novo caminho de pesquisa. O tipo de raciocínio empregado é o abdutivo (MAGNANI, 2001) ou seja, o "insight" sobre a solução de questões não explicadas na Ciência e a formulação de novas hipóteses de solucioná-las. Esta tipologia é provisória, uma vez que só foram encontrados somente 3 artigos deste tipo entre os 89 analisados. O famoso artigo de Watson e Crick (1953) que propôs a estrutura helicoidal para a molécula do DNA seria um típico artigo deste tipo.

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 satisfatórias,
- diante disso, propomos a seguinte HIPÓTESE original

Artigos experimentais constam necessariamente de um experimento empírico. Se dividem em exploratórios, dedutivos e indutivos. 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.

Artigos experimentais-exploratórios tem um caráter exploratório ao desvendar um fenômeno ainda desconhecido pela ciência (FRANKLIN, 2004), geralmente não são guiados por uma hipótese e buscam descrever/caracterizar (MILLER, 1947) este fenômeno como primeiro estágio para integrá-lo/classificá-lo ao esquema de um domínio científico, trabalhando na direção proposta por Dahlberg (1995) de formular e provar proposições que descrevem/caracterizam um fenômeno. Este tipo de artigo pode vir a ganhar importância em função da emergência de ferramentas automatizas que permitem aos cientistas identificar padrões nos dados sem o auxílio de hipóteses prévias. (THE FOURTH PARADIGM, 2009).

O desenvolvimento do raciocínio num artigo **experimental-exploratório** segue o seguinte padrão:

- dado um PROBLEMA ou FENÔMENO ainda não bem caracterizado,
- desenvolvemos o seguinte EXPERIMENTO que permite identificar a(s) seguinte(s) CARACTERÍSTICA(s) desse FENÔMENO.

Artigos experimentais-dedutivos trabalham a partir de relações entre fenômenos já formuladas anteriormente, cujas referências vêem citadas, aplicando-as a testando-as e validando-as um contexto específico. Os **artigos experimentais-indutivos** se caracterizam por proporem e testarem novas relações entre fenômenos.

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 HIPÓTESE(s) anteriores para sua solução,
- diante disso, escolhemos a seguinte (uma das HIPÓTESE(s) anteriores).
- ampliamos e recontextualizamos esta HIPÓTESE anterior; desenvolvemos o seguinte EXPERIMENTO para testar esta HIPÓTESE anterior;

- o EXPERIMENTO apresentou os seguintes RESULTADO(s).

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 RESULTADO(s).

Os elementos do modelo podem ser representados como na seguinte Figura.

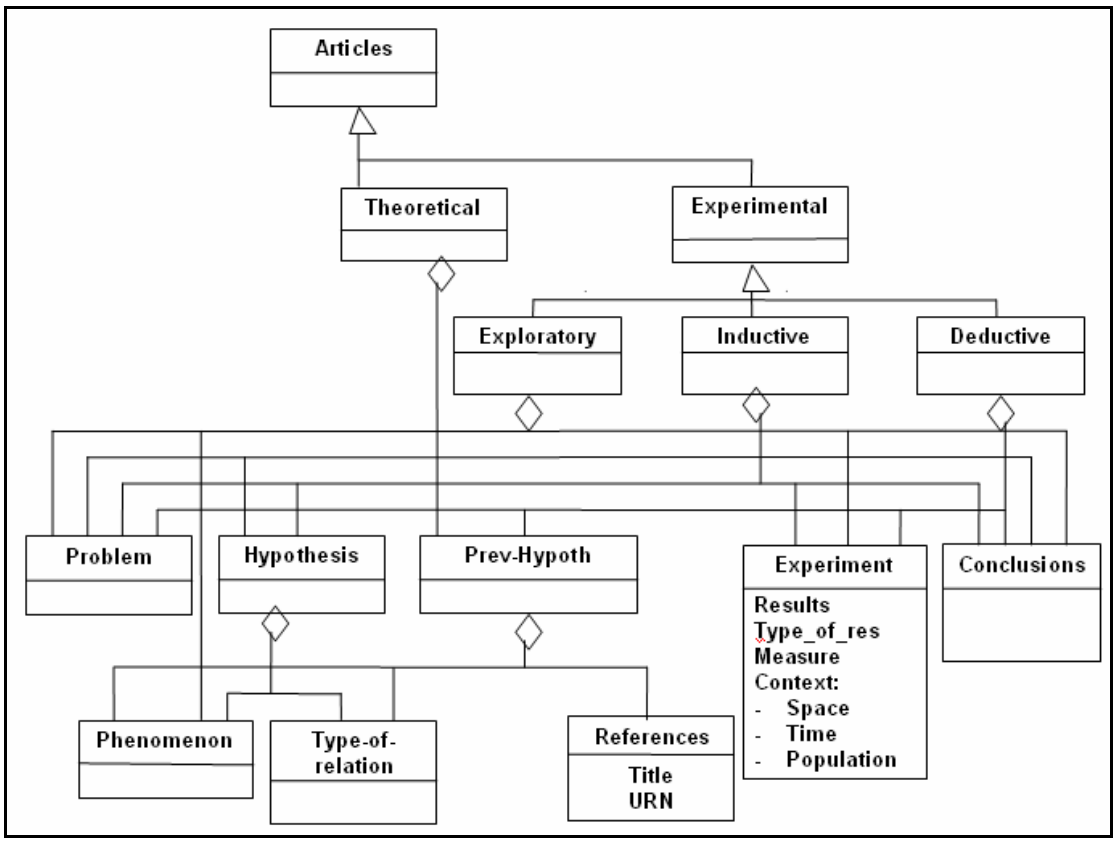

Figura 1 - Modelo de Representação do Conhecimento

A potencialidade de um modelo semântico de representação de conhecimento como o proposto é que o mesmo permite a recuperação do conhecimento científico de uma forma semanticamente mais rica. Programas podem fazer "inferências" sobre o conhecimento representado segundo o modelo, como nos seguintes exemplos:

- O papiloma vírus (Antecedente) causa (Relação) que outros tipos de câncer (Conseqüente?)?
- Que outras (Antecedente?) causas (Relação) pode ter o câncer de colo de útero (Conseqüente?) além do papiloma vírus?

5. ARTIGOS SEMÂNTICOS E SUAS POTENCIALIDADES

Um desafio considerável para a implementação do modelo proposto é a obtenção dos elementos semânticos dos artigos. No contexto da produção de artigos científicos nas ciências biomédicas, estes são submetidos pelos próprios autores em sistemas Web de submissão com os quais contam praticamente todos os periódicos eletrônicos hoje. Questões como enriquecimento, anotação e marcação dos elementos semânticos do artigo só poderá ser endereçada com sucesso com o apoio dos próprios autores, a exemplo do que já fazem quando formulam resumos estruturados. Consideramos em nossa proposta o momento da submissão de artigos a sistemas de periódicos eletrônicos ou a bibliotecas e repositórios digitais como um momento *privilegiado*, em que autores estão especialmente motivados a realizarem estas tarefas. Será necessário criar ferramentas que apóiem este processo, editores semânticos de artigos científicos e sistemas de submissão, apoiados em ontologias biomédicas.

Nossa proposta é o sistema de auto-submissão de artigos a periódicos eletrônicos (COSTA, 2010), no qual autores, além dos metadados convencionais que descrevem seu artigo, entram também com as conclusões do artigo. O sistema processa lingüisticamente o texto das conclusões, formatando-as como relações segundo o modelo proposto, além de mapear termos da conclusão em termos do UMLS; o processo é todo validado pelo próprio autor, que verifica se os termos e relação sugeridos pelo sistema equivalem aos da conclusão do seu artigo. Obtém-se assim um *registro semântico* do artigo, como ilustrado na figura a seguir, em que a seguinte conclusão: “telomere replication (Antecedent) involves (Type_of_relation) a terminal transferase-like activity (Consequent),” encontrada em Segundo et al. (2004), pode ser representada em RDF. Observa-se o uso de 3 “names

spaces", vocabulários específicos que agregam semântica às declarações RDF: dc (Dublin Core), sa (Semantic Article, proposta nossa) e UMLS (Unifyed Medical Language System).

```
<rdf:RDF
      xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:sa="http://example.org/semarticles/"
xmlns:umls="http://www.nlm.nih.gov/research/umls/">
<rdf:Description rdf:about="http://art_id/">
<dc:title>title</dc:title>
<dc:creator>creator</dc:creator>
<dc:subject>subject</dc:subject>
 <dc:date>date</dc:date>
<sa:conclusion>
<rdf:Description rdf:about="http://art_id/conclusion">
<sa:antecedent>telomere replication</sa:antecedent>
<sa:type_rel>involves</sa:type_rel>
<sa:consequent>a terminal transferase-like activity</sa:consequent>
<sa:antecedent_mapping>http://www.nlm.nih.gov/research/umls/CUI01</sa:antecedent_mapp
ing>
<sa:type_rel_mapping>http://www.nlm.nih.gov/research/umls/CUI02</sa:type_rel_mapping>
<sa:consequent_mapping>http://www.nlm.nih.gov/research/umls/CUI03</sa:consequent_map
ping>
</rdf:Description>
</sa:conclusion>
</rdf:Description>
</rdf:RDF>
```

Figura 2. Conclusão do artigo representada em RDF. CUI significa Identificador Único do Conceito, do UMLS.

O UMLS é uma base terminológica no domínio das ciências biomédicas, englobando mais de 100 fontes[4]. É composta de três bases de conhecimento integradas: o Metathesauros, contendo 2.886.423 termos[5]; a "Semantic Network", que estrutura termos biomédicos em 154 categorias, denominadas "*semantic types*", relacionadas entre si por

[4] http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/source_vocabularies.html

[5] http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html

54 “*semantic relations*”; e o Specialist Lexicon, com informações sintáticas, morfológica e ortográficas sobre termos em inglês encontrados no UMLS. Rotinas em linguagem de programação Java usam o Specialist Lexicon para tratar textos biomédicos. Em relação a uma base terminológica convencional como o MEsH, a extensão do UMLS com o “Semantic Network” lhe confere maior potencial semântica; segundo seus criadores: *“*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*”*.

O UMLS, ao contrário de vocabulários que fazem parte dele e são utilizados para indexar a literatura científica como o MeSH, onde um artigo é indexado por termos sem qualquer relação entre eles, incorpora *relações semânticas* na “Semantic Network”, nas quais são especificados tipos semânticos que podem fazer parte de cada uma delas. Num modelo de representação do conhecimento como o proposto, esta característica é essencial, uma vez que o objetivo do processamento é que as conclusões do artigo sejam formatadas em relações, onde cada um dos relata é mapeado para termos do “MetaTheaurus” e a relação é *mapeada* pelo sistema (COSTA, 2010) para uma das “semantic relations” do “Semantic Network”. Posteriormente, o resultado deste mapeamento, registrando como um autor vê representados, no momento sua publicação, os termos e relações da conclusão do seu artigo no UMLS, ou se não os vê representados, ou os vê somente representados parcialmente, poderá ser usado para a identificação de novas descobertas reportadas em artigos, como explicado a seguir.

Outro desdobramento promissor da pesquisa é a utilização do modelo para identificação de novas descobertas, através da comparação do seu conteúdo (expresso pelas suas conclusões, por exemplo), com o conteúdo de terminologias biomédicas como o UMLS. Esta é a hipótese da tese de Malheiros (2010) e baseia-se no seguinte.

Thomas Kuhn (2007), um dos mais proeminentes autores em Filosofia da Ciência, teorizou sobre a evolução e mudança na Ciência. Uma característica do que Kuhn chama

de período pré-paradigmático num domínio científico é a falta de uma terminologia precisa e consensada. O discurso científico no período pré-paradigmático necessita assim de estabelecer de forma precisa o significado dos conceitos que utiliza. Um indicador de que um domínio científico atingiu um estágio paradigmático é o estabelecimento de um sistema conceitual consensado no qual conceitos utilizados para descrever o paradigma possuem um significado preciso. No capítulo, Kuhn (2007, p. 149) ressalta que, de um ponto de vista cognitivo, novos conceitos são necessários para lidar com a mudança de paradigma; um novo paradigma vai requerer assim um novo sistema conceitual para descrevê-lo; novos conceitos implicam em novos termos para representá-los

Descobertas científicas necessitam de um período de tempo para que sejam avaliadas, criticadas, reformuladas, aceitas ou rejeitadas pela comunidade num domínio científico. Terminologias/ontologias/biomédicas mantém o conhecimento consensado num domínio (e não o conhecimento revolucionário ou controverso resultante das novas descobertas que ainda não atingiram o consenso da comunidade científica), uma vez que seu principal objetivo até hoje tem sido indexar a literatura científica ou experimentos científicos como a GO.

Altamente significativo é o fato observado de que existe um *intervalo de tempo* entre uma nova descoberta científica, sua integração à sistema de conceitos de um domínio científico e o processo de definir um termo para representá-lo. A enzima telomerase foi descoberta em 1985; um termo MeSH para representá-la foi estabelecido somente em 1995, dez anos após sua descoberta. Também, um relatório de 1981 do Centers of Disease Control and Prevention, EUA (CDC 1981), relatando cinco casos de pneumocystis carinii pneumonia (PCP) entre homens jovens saudáveis em Los Angeles, uma doença que viria a ser conhecida como AIDS; de acordo como a National Library of Medicine[6], um termo para a AIDS foi estabelecido no MeSH somente em 1983. Mudanças científicas necessitam de um novo sistema conceitual é este necessita de um intervalo de

[6] http://www.nlm.nih.gov/mesh/MBrowser.html.

tempo para sua representação nas terminologias científicas

Em sua tese Malheiros (2010) analisou artigos que compõe as “key publications” indicadas pelos pesquisadores ganhadores do Prêmio Lasker de Medicina de 2006, onde são relatados na seqüência de artigos os passos que levaram a descoberta da enzima telomerase, fundamental para a reprodução celular e que tem conseqüências na compreensão do desenvolvimento de doenças como o câncer. Segundo esta autora, indícios de novidades científicas poderiam ser identificados comparando o conteúdo de conclusões de artigos com terminologias biomédicas. A análise efetuada cronologicamente, mostrou que nos primeiros artigos a taxa de *mapeamento* dos termos das conclusões em termos do UMLS *era baixa* mas crescia ao longo do tempo, à medida que este novo fenômeno científico começa a se refletir em novos termos do UMLS. Detalhes podem ser encontrados em Malheiros e Marcondes (2011).

Além disso, verificou-se também que os artigos analisados seguiam um padrão classificatório derivado do modelo proposta: os primeiros artigos em ordem cronológica eram **experimentais-exploratórios**, refletindo a caracterização inicial e apropriação de um novo fenômeno pela ciência; à medida que a telomerase vai sendo descrita e suas propriedades identificadas, os artigos passam a ser do tipo **experimental-indutivo** ou **dedutivo**, nos quais o novo fenômeno passa a ser relacionado com outros, conforme a descrição deste tipo de artigos feita anteriormente.

Os resultados, apesar de serem iniciais, mostram que um modelo semântico de representação de artigos científicos, no qual o conteúdo é considerado e *explicitado*, representado em padrões da Web Semântica, mapeado/correlacionado também explicitamente com termos de uma terminologia biomédica, pode trazer muitas potencialidades. A figura a seguir ilustra como a representação de conclusões em RDF, o mapeamento de termos e relações destas conclusões para termos do UMLS e o registro de todas estas informações, juntamente com metadados bibliográficos convencionais e o

texto do próprio artigo, num *registro bibliográfico ampliado*, podem apoiar a identificação de novas descobertas. No exemplo a seguir, o relacionamento, entre os dois artigos - *causado_por* é um caso específico do relacionamento *associado_a* -, indica que, segundo o relatado em ambos os artigos, o "encurtamento dos telômeros" poderia estar ligado ao "câncer".

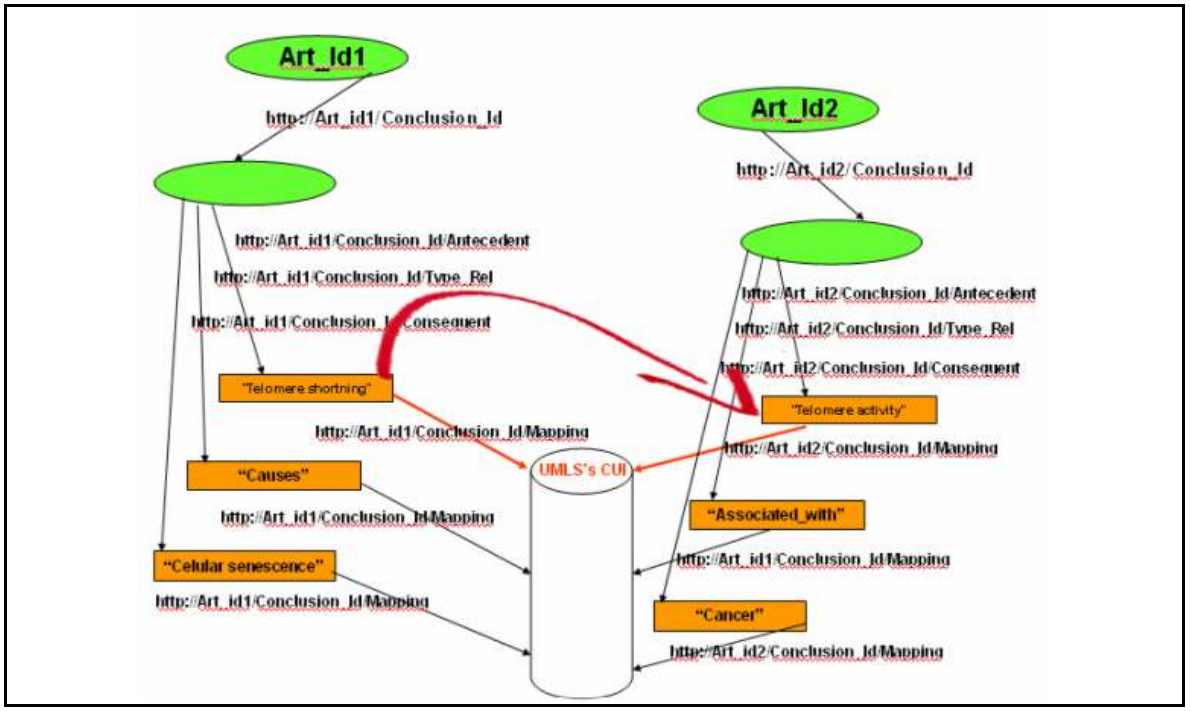

Figura 4. Dois artigos com conclusões relacionadas, conectados por um tipo semântico do UMLS.

Uma questão a ser testada é verificar se os elementos semânticos do modelo proposto apresentam outras relações além daquelas constitutivas dos diferentes raciocínios/tipos de artigos descritas anteriormente. Que tipos de relações existiriam entre elementos semânticos de um artigo como Problema, Dados, Questão de pesquisa, Objetivos, Hipóteses, Método, Procedimentos metodológicos, Experimento, Variáveis, Resultados e Conclusões? Pode-se supor que um PROBLEMA suscite uma QUESTÃO e esta, se desdobre em OBJETIVOS de pesquisa; uma QUESTÃO seria uma relação em que ou um dos relatas ou um dos relatas mais o TIPO-DE-RELAÇÃO são *desconhecidos*; no primeiro caso teríamos um artigos experimental, indutivo ou dedutivo; no segundo, um

experimental-exploratório, com uma questão do tipo “o que é?”. Ao ser proposto um relata, ou um relata mais o TIPO-DE-RELAÇÃO, chega-se a uma HIPÓTESE. A HIPÓTESE se desdobra em variáveis para as quais se propõe experimentos para medi-las. Os experimentos chegam a RESULTADOS, que confirmam, refutam ou reformulam a HIPÓTESE proposta.

Novas aplicações das tecnologias da Web Semântica permitem supor que em breve artigos científicos publicados na Web formarão uma rede, incluindo textos completos, metadados semânticos, bases de dados, citações semânticas, terminologias/ontologias biomédicas (MARCONDES, 2012). Esta rede poderá ser percorrida por programas inteligentes, que terão acesso aos elementos semânticos do conteúdo do artigo, realizando com muito mais eficiência tarefas relacionadas este conteúdo, viabilizando aplicações de “literature-based discovery” ou mineração de textos, que permitirão a identificação de inconsistências, “gaps” no conhecimento existente, ou indícios de novas descobertas.

6. CONSIDERAÇÕES FINAIS

Juntamente com as tecnologias da Web Semântica, a proposta de dados abertos interligados (BIZER et al., 2007) aponta na direção de existência independente e permanente dos registros de artigos científicos (identificando-os através de URIs), sua integração com outros recursos disponíveis na Web (através dos “links” semânticos usando RDF) e ampliação da sua semântica (usando diferentes vocabulários). Vê-se assim que a proposta de dados abertos interligados tem grande potencial de descrever, identificar permanentemente, estruturar e interligar recursos científicos na Web, agregando semântica a esta descrição ao lançar mão dos inúmeros vocabulários que vêm sendo desenvolvidos com esta finalidade, resultando que *“all statements provided about a particular uniquely identified resource can be aggregated into a global graph*” (LIBRARY LINKED DATA INCUBATOR GROUP FINAL REPORT, 2011).

A OC sempre valorizou as relações como portadores de significado (PERREAULT, 1994), (DAHLBERG, 1995), (TILLET, 2001), (VELTMANN, 2004), as sistematizou, conceituou e organizou-as em taxonomias. As únicas relações entre artigos científicos tratadas até agora na gestão do conhecimento científico eram as relações de citação. Agora estão disponíveis tecnologias baseadas, como dados abertos interligados, que se baseiam exatamente em relações, com potencial de ampliarem a semântica computacional à disposição dos SOCs. Os SOCs atuais que armazenam e recuperam registros de artigos científicos – catálogos de arquivos, bibliotecas e repositórios digitais - são sistemas fechados, com tecnologias de armazenamento e recuperação de registros de conhecimento que remotam à década de 1980. Estes SOCs encerram e condicionam o significado dos registros neles armazenados em verdadeiros"silos" (BERMES, 2011) impedindo que estes os mesmos tenham existência independente fora deste ambiente computacional e que possam ser integrados – terem "links" para e receber "links" de -, aos fluxos do conhecimento científico e aos fluxos gerais da Web (MARCONDES, 2012b).

Pesquisas recentes sobre bibliotecas digitais semânticas apontam na superação destas limitações dos SOCs atuais (LYTRAS et al., 2005), (KRUK et al., 2008), (LI et al., 2010). As tecnologias semânticas permitem que os SOCs armazenem e recuperem conhecimento, não mais num sentido metafórico – simples registros de conhecimento -, mas sim proposições que representam o conhecimento em si, em especial na área científica - afirmações representadas como triplas RDF. A figura a seguir mostra uma consulta por "doenças associadas a uma proteína" no Semantic System Biology[7] – BioGateway, base de conhecimento com mais de 2 bilhões de triplas RDF de diversas fontes.

[7] Disponível em http://www.semantic-systems-biology.org/home.

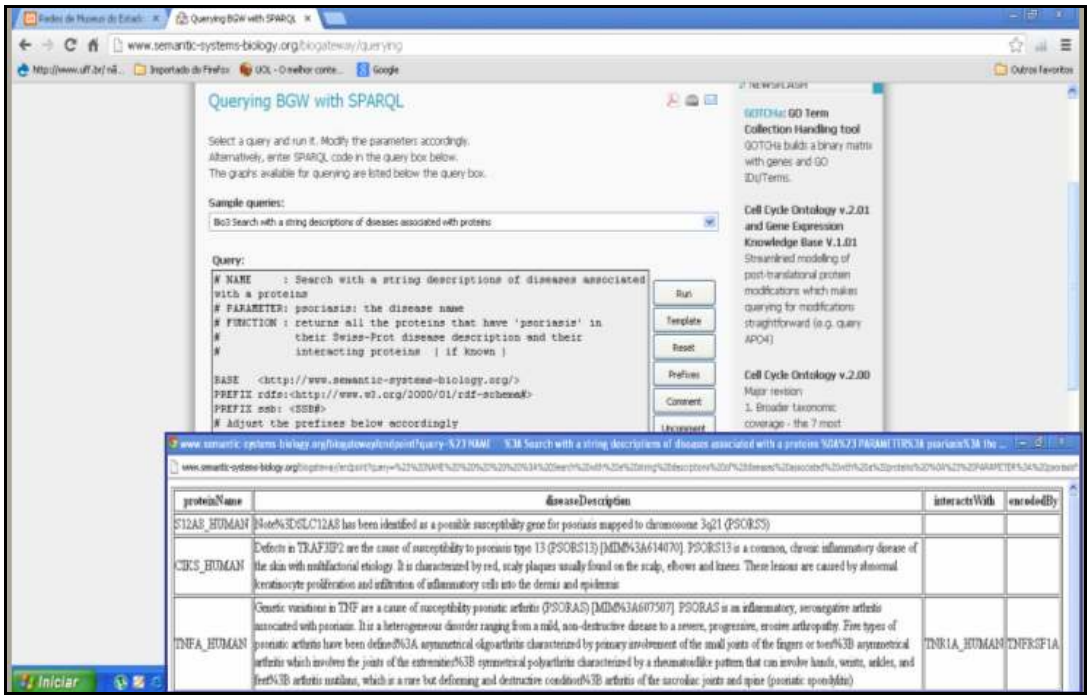

Figura 5 – registros formatados como triplas RDF são recuperados usando a linguagem SPARQL, da base BioGateway.

Formatos semânticos de registros de conhecimento científico, juntando metadados convencionais, texto completo, dados de pesquisa, elementos semânticos como questões, objetivos, hipóteses e conclusões, podem tirar partido destas tecnologias e ampliar as possibilidades de gestão do conhecimento científico armazenado e disponibilizado na Web.

REFERÊNCIAS

ALLSOPP, R. C.; VAZIRI, H.; PETTRSON, C.; GOLDSTEIN, S.;YOUGLAI, E. V.; FUTCHER, C. W.; GREIDER, C. W.; HARLEY, C. B. Telomere length predicts the replicative capacity of human fibroblasts, *Proc. Nat. Acad. Sci. USA*, v. 89, p. 10114-10118, 1992.

ATTWOOD, T. K.; KELL, D. B.; MCDERMOTT, P.; MARSH, J.; PETTIFER, S. R.; THORNE, D. Calling international rescue: knowledge lost in literature and data landslide! *Biochemical Journal*, [S.l.], dec. 2009.

BATH, P. Data Mining in Health and Medical Information, *Anual Review of Information*

Science and Technology, v. 38, p. 331–369, 2002.

BERMES, Emmanuelle. Convergence and Interoperability: a Linked Data perspective. In: IFLA World Library and Information Congress, 77th. Puerto Rico, 2011. *Proceedings...* 2011. Disponível em: <conference.ifla.org/past/ifla77/149-bermes-en.pdf>. Acesso em: 3 fev. 2012.

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

BEZERMAN, Charles. *Shaping written knowledge*: *Rhetoric of the human sciences*. Madison: The University of Wisconsin Press, 1988.

BIZER, C.; HEALTH, T.; BERNERS-LEE, T. Linked data – the story so far, In: T. Heath, M. Hepp, C. Bizer (eds.), Special Issue on Linked Data, *International Journal on Semantic Web and Information Systems* (IJSWIS).

BIZER, C.; CYGANIAK, R.; HEATH, T. How to publish Linked Data on the Web. [2007]. Disponível em: < http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/>. Acesso em: 03 nov. 2011.

BJÖRK, B. C., ROOS, A.; LAURI, M. Scientific journal publishing: yearly volume and open access availability. *Information Research*, v**.** 14, n**.** 1 paper 391, 2009. Disponível em: <http://InformationR.net/ir/14-1/paper391.html>. Acesso em 22 fev. 2013.

CHEN, J.; BLASCO, M. A.; GREIDER, C. W. Secondary structure of vertebrate telomerase RNA., *Cell***,** v. 100, p. 503–514, 2000.

COSTA, Leonardo Cruz. Da. Um proposta de processo de submissão de artigos científicos à publicações eletrônicas semânticas em Ciências Biomédicas, Tese (doutorado), Programa de Pós-graduação em Ciência da Informação UFF-IBICT. Niterói, (2010).

DINAKARPADIAN, Deendayal et al. MachineProse: an ontological framework for scientific assertions**.** *Journal of the American Medical Informatics Association*, [S.l.], v. 13, n. 2, p. 220-232, mar./apr. 2006. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1447552/>. Acesso em

THE FOURTH PARADIGM: data intensive scientific discovery. HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin. (Eds.). Redmond, Washington: MicroSoft Research, 2009.

FRAKLIN, Laura R. Exploratory Experiments. In *Philosophy of Science Assoc. 19th*

Biennial Meeting - PSA2004: Contributed Papers, 2004, Proceedings…. Austin, Texas; 2004. Disponível em: <http://philsci-archive.pitt.edu/archive/00002070/01/UploadedPSA2004.doc>. Acesso em 13 jun. 2008.

GARDIN, J-C. Vers un remodelage des publications savantes: ses rapports avec sciences de l'information. In: Chaudrion & Fluhr (Eds). Filtrage et Résumé Automatique de l'Information sur les Reseaux - Actes du 3ème Colloque du Chapitre Français de l'ISKO, 2001.

GAO, Y; KINOSHITA, J.; WU, E.; MILLER, E.; LEE, R; SEABORNE, A.; CAYZER, S.; CLARK, T. SWAM: a distributed knowledge infrastructure for Alzeimer disease research. *Journal of Web Semantic*, [S.l.], v. 4, n. 3, 2006. Disponível em: <http://www.websemanticsjournal.org/ps/pub/2006-17>. Acesso em: 12 dez. 2010.

GREIDER, C. W.; BLACKBURN, E. H. Identification of a specific telomere terminal transferase activity in Tetrahymena extracts, *Cell***,** v**.** 43, p. 405-413, 1985.

GREIDER, C. W.; BLACKBURN, E. H. The telomere terminal transferase of Tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity. *Cell, v.* 51, p. 887-898, 1987.

GROSS, A. G. The Rhetoric of Science. Cambridge, Massachusetts; London: Harvard University Press, 1990. ISBN 0-674-76873-6.

GUO-LIANG, Y.; BRADLEY, J. D.; ARTTARDI, L. D.; BLACKBURN, E. In vivo alteration of telomere sequences and senescence caused by mutated Tetrahymena telomerase RNAs. *Nature,* v. 344, p. 126-132, 1990.

HOFFMANN, M. Is there a "logic" of abduction? In: A. Gimate-Welshe (ed), *Selected paper- 6th Congress of the International Association for Semiotics Studies*, Guadalajara, Mexico 1997 (Grupo Editorial Miguel Angel Porrua, Mexico City, 2000. Available at: http://www.unibielefeld.de/idm/personen/mhoffman/papers/abduction-logic.html (accessed 14 Dez. 2005).

HUNTER, Jane. Scientific Publication Packages – A Selective Approach to the Communication and Archival of Scientific Output. The International Journal of Digital Curation, v. 1, n. 1, 2006. Disponível em: <http://www.ijdc.net/index.php/ijdc/article/view/8>. Acesso em: 1 fev. 2013.

HUNTER, L.; BAUMGARTNER JR, W. A.; LU, Z.; JOHNSON, H. L.; CAPORASO, J. G.; PAQUETTE, J.; LINDEMANN, E. K.; WHITE, O. Medvedeva; COHEN, K. B. Concept recognition for extracting protein interaction relations from biomedical text. *Genome Biology*, v. 9, 2008, Suppl. Disponível em: <http://genomebiology.com/2008/9/S2/S9>. Acesso em: Nov.20 2008.

HUTCHINS, J. On the structure of scientific texts. In: UEA Papers in Linguistics, Norwich. Norwich, UK: University of East Anglia, 1977, 5, Proceedings… p. 18-39. 1977. Disponível em: <http://ourworld.compuserve.com/homepages/wjhutchins/UEAP/L-1977.pdf>. Acesso em: 20 Mar 2006.

International Committee of Medical Journals Editors. 2003. Disponível em: <www.icmje.org>. Acesso em 14 jul. 2005.

KANDO, N. Text-level structure of research papers: implications for text-based information processing systems. In: J. Furner and D. J. Harper (eds.), *Information Retrieval Research: Proceedings of the 19th BCS-IRSG Colloquium on IR Research*, Aberdeen, 1997 (Springer-Verlag, Aberdeen, Scotland, 1997).

KANDO, N. Text structure analysis as a tool to make retrieved documents usable**.** In: *Proceedings of the 4th International Workshop on Information Retrieval with Asian Language*, Taipei, 1999 (Academia Sinica, Taipei, Taiwan, 1999).

KLAHR, D.; SIMON, H. A. Studies of scientific discovery: complementary approaches and convergent findings, *Psychological Bulletin* 125(5) (1999) 524-543.

KINTSH, W.; VAN DIJK, T. A. Towards a model of text comprehension and production, *Psycological Review, v.* 84, n5, p. 363-393, 1972.

KRUK, Sebastian Ryszard; MCDANIEL, Bill (Ed.). *Semantic digital libraries*. Springer, 2008.

KOSTOFF, R. N.; BRIGGS, M. B.; SOLKA, J. L.; RUSHENBERG, R. L. (2008). Literature-related discovery (LRD): Methodology. *Technological Forecasting and Social Change*, v. 75, n. 2, p.186–202. Disponível em: <doi:10.1016/j.techfore.2007.11.010>. Acesso em: 20 jul. 2010.

LATOUR, Bruno. *Ciência em ação*: como seguir cientistas e engenheiros sociedade afora. São Paulo: Ed. UNESP; 2000.

LI, Na; ZHU, L.; MITRA, P.; MUELLER, K.; POWELEIT, E.. OreChem ChemXSeer: a semantic digital library for chemistry. In: The Annual joint conference on Digital libraries, 10 th., *Proceedings*... ACM, 2010. p. 245-254.

LIBRARY LINKED DATA INCUBATOR GROUP FINAL REPORT. W3C, 2011. Disponível em: <http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/>. Acesso em: 3 fev. 2012.

LYTRAS, Miltiadis; SICILIA, Miguel-Angel; DAVIES, John; KASHYAP, Vipul. Digital libraries in the knowledge era: Knowledge management and Semantic Web technologies, *Library Management*, v. 26, n. 4/5, p.170 – 175, 2005.

MAGNANI, L. *Abduction, Reason, and Science: processes of discovery and explanation.* New York: Kluwer Academic, Plenun Publishers, 2001.

MALHEIROS, Luciana Reis. *A identificação de traços de descobertas científicas pela comparação do conteúdo de artigos em Ciências Biomédicas com uma ontologia pública*. Tese (Doutorado em Ciência da Informação)-Programa de Pós-Graduação em Ciência da Informação convênio UFF/IBICT, Niterói, 2010.

MALHEIROS, Lucia Reis; MARCONDES, Carlos Henrique. Identificación de los rasgos de descubiertas cientificas en artículos biomedicos. *Revista EDICIC*, v. 1, n. 4, 2011, ISSN: 2236-5753. Disponível em: <http://www.edicic.org/revista/index.php?journal=RevistaEDICIC&page=article&op=view&path[]=74>. Acesso em 28 nov. 2011.

MARCONDES, Carlos Henrique. A semantic model for scholarly electronic publishing. In: International Workshop on Semantic Publication - SePublica2011-, 1st, at the Extended Semantic Web Conference (ESWC), 8th, in Hersonissos, Crete, Greece, Proceedings... *CEUR Workshop Proceedings*, v. 721, 2011. ISSN: 1613-0073. Disponível em: <http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-721/>. Acesso em 30 maio 2011.

MARCONDES, Carlos Henrique. Um modelo semântico de publicações eletrônicas. *Liinc em revista*, v. 7, n. 1, 2011b. Disponível em <http://revista.ibict.br/liinc/index.php/liinc/article/viewFile/404/262>. Acesso em 30 maio 2011.

MARCONDES, Carlos H. Em busca de uma semântica do digital, ou “as they may think”. Ponto de Acesso, v. 6, n. 12, 2012. Disponível em: <http://www.portalseer.ufba.br/index.php/revistaici/article/view/6103>. Acesso em 2 dez. 2012.

MARCONDES, Carlos H. From scientific communication to public knowledge: the scientific article Web published as a knowledge base. In: Egelen, Jan, Dobreva, Milena, ed. ICCC ElPub - INTERNATIONAL CONFERENCE ON ELECTRONIC PUBLISHING, Leuven, Bélgica, 2005, 9, *Proceedings*... Leuven, Bélgica, 2005. p.119-127. Disponível em <http://elpub.scix.net>.

MARCONDES, Carlos Henrique; MALHEIROS, Luciana Reis. Identifying traces scientific discoveries by comparing the content of articles in biomedical sciences with web ontologies. In: ISSI - International Conference on Informetrics and Scientometrics, 2009, Rio de Janeiro. 12, Proceedings. São Paulo: BIREME/PAHO/WHO, UFRJ, 2009. v. 1. p. 173-177.

MCEACHERN, M. J.; BLACKBURN, E. H. Runaway telomere elongation cause by telomerase RNA mutations. *Nature,* n. 376, p. 403-409, 1995.

METS – Metadata Enconding & Transmission Standard, http://www.loc.gov/standards/mets/.

MILLER, D. L. Explanation Versus Description, Philosophical Review 56(3) (1947) 306-312.

MULLER, Hans Michael, KENNY, Eimear , STERNBERG, Paul W. Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. *PLoS Biology*, v.2, n.11. 2004.

MURRAY-RUST, P.; RZEPA, H. S. Chemical Markup, XML and the worldwide web. I: basic principles, *Journal of Chemical Information and Computer Science* v. 39, p. 928-942, 1999.

MURRAY-RUST, P.; RZEPA, H. S. STMML. A markup language for scientific, technical and medical publishing, *Data Science Journal* v. 1, n. 2, p.128-193. 2002. Disponível em: <http://journals.eecs.qub.ac.uk/codata/journal/contents/1_2/1_2pdfs/ds121.pdf>. Acesso em: 18 set. 2005.

NATIONAL LIBRARY OF MEDICINE. Structured abstract. Disponível em:

<http://www.nlm.nih.gov/bsd/policy/structured_abstracts.html>. Acesso em: 1 fev. 2013.

NIINILUOTO, I. Scientific progress. In: *Stanford Encyclopedia of Philosophy*. 2002.

OWL Ontology Web Language Overview. W3C, 2004. Disponível em: http://www.w3.org/TR/owl-features/. Acesso em 28 fev. 2013.

PERREAULT, Jean. Catagories and relators: a new schema. *Knowledge Organization*, v. 21, n. 4, p. 189198, 1994.

POPPER, K. *A Lógica da Pesquisa Científica*. (Ed. Cultrix, Ed. USP, São Paulo, 2001).

RACUNAS, S. A. et al. HyBrow: a prototype system for computer-aided hypothesis evaluation. *Bioinformatics*, [S.l.], v. 20, n. 1, p. 257-264, 2004.

RDF Primer, W3C, 2004. Disponível em: http://www.w3.org/TR/rdf-primer/. Acesso em 28 fev. 2013.

SEGUNDO, G. R. S. et al. A comparative study of congenital toxoplasmosis between public and private hospitals from Uberlândia, MG, Brazil. *Mem. Inst. Oswaldo Cruz* [online], v. 99, n. 1, p. 13-17, 2004.

SHOTTON, David; PORTWIN, Graham Klyne; MILES, Alistair. Adventures in semantic publishing: Exemplar semantic enhancements of a research article. PLoS Comput. Biol., v. 5, n. 4, 2009.

SHUM, Simon Buckingham et al. Visualizing internetworked argumentation.In: *Visualizing Argumentation: Software Tools for Collaborative and Educational Sensemaking***.** Springer-Verlag, 2003. p. 185-204.

SOLDATOVA, L. D; KING, R. D. An ontology of scientific experiments. *Journal of the RoySociety Interface*, [S.l.], v. 3, n. 11, p. 795-803, 2006. Disponível em: <http://journals.royalsociety.org/content/u552845783800t73/fulltext.pdf>. Acesso em: 5 fev. 2011.

SPARQL Query Language for RDF, 2008. http://www.w3.org/TR/rdf-sparql-query/.

SPASIC, I.; ANANIADOU, S.; MCNAUGHT, J.; KUMAR, A. Text mining and ontologies in biomedicine: making sense of raw text. Briefings in bioinformatics, v. 6, n. 3, p. 239–51, 2005. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/16212772>. Acesso em 27 jul. 2010.

STRUCTURED ABSTRACT LABELS RESEARCH DATASET. Disponível em: <http://structuredabstracts.nlm.nih.gov/Downloads/Cohort_Study_Appendix.pdf>. Acesso

em: 1 fev. 2013.

SWANSON, D.R.; SMALHEISER, N. R.; TORVICK V. I. Ranking indirect connections in literature based discovery. The role of Medical Subject Headings, *Journal of the American Society for Information Science and Technology,* v. 57, n.11, p. 1427–1439, 2006.

TENOPIR, Carol; KING, Donald W.; EDWARDS, Sheri; WU, Lei, Electronic Journals and Changes in Scholarly Article Seeking and Reading Patterns. School of Information Sciences Publications and Other Works. 2009. Disponível em: http://trace.tennessee.edu/utk_infosciepubs/7>. Acesso em: 15 fev. 2013.

TANABE, L.; SCHERF, U.; SMITH, L. H.; LEE, J. K.; HUNTER, L. ; WEINSTEIN, J. N. MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. BioTechniques, v. 27, n. 6, p. 1210–4, 1216–7, 2009. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/10631500>. Acesso em 4 abr. 2009.

THAGARD, P. Computational Philosophy of Science. Cambridge, MA: The MIT Press, 1993.

TILLET, Barbara. Bibliographic relationships. In: C. A. Bean & R. Green (Eds.): *Relationships in the organization of knowledge.* Dordrecht: Kluwer Academic Publishers, 2001. p. 19-35.

UMLS - Unified Medical Language System Fact Sheet. Disponível em: <http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html>. Acesso em: 2 dez. 2012.

RFC 2396 - Uniform Resource Identifiers (URI): Generic Syntax, Berners-Lee T., Fielding R., Masinter L., IETF, August 1998. Disponível em: <http://www.isi.edu/in-notes/rfc2396.txt>. Acesso em: 15 dez. 2011.

DE WAARD. From Proteins to Fairytales: Directions in Semantic Publishing. IEEE Intelligent Systems v. 25, n.2, p. 83-88, 2010.

DE WAARD; BUCKINGHAM, Simon SHUM; CARUSI, Carus; PARK, Jack; SAMWALD, Matthias; SANDOR, Ágnes Hypotheses, evidence and relationships: The HypER approach for representing scientific knowledge claims. In: Proceedings 8th International Semantic Web Conference, Workshop on Semantic Web Applications in Scientific Discourse. Lecture Notes in Computer Science. Washington DC, Berlin: Springer Verlag, 2009.

VAN HAAN, Anthony F. Sleeping beauties in science. *Scientometrics,* v. 59, n.3, p. 467-472, 2004.

XML Extensible Markup Language. W3C. Disponível em: http://www.w3.org/XML/. Acesso em 28 fev. 2013.

WATSON, J. D.; CRICK, F. H. C. The molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, n. 4356, April, 1953.

ZHANG, Y. et al. Protein-protein interaction extraction based on improved all-paths kernel. *Journal of Computational and Theoretical Nanoscience*, v. 8, n. 10, p. 1925-1932, 2011.