Um modelo semântico de publicações eletrônicas

Carlos Henrique Marcondes*

Resumo Publicações eletrônicas, apesar dos avanços das Tecnologias da Informação, são ainda calcados no modelo impresso. O formato textual impede que programas possam ser usados para o processamento "semântico" desses conteúdos. È porposto um modelo "semântico" de publicações cientificas eletrônicas, no qual as conclusões contidas no texto do artigo fornecidas por autores e representadas em formato "inteligível" por programas, permitindo recuperação semântica, identificação de indícios de novas descobertas científicas e de incoerências sobre este conhecimento. O modelo se baseia nos conceitos de estrutura profunda, ou semântica, da linguagem (CHOMSKY, 1975), de microestrutura, macroestrutura e superestrutura, (KINTSH, VAN DIJK, 1972), na estrutura retórica de artigos científicos (HUTCHINS, 1977), (GROSS, 1990) e nos elementos de metodologia cientifica, como problema, questão, objetivo, hipótese, experimento e conclusão. Resulta da análise de 89 artigos biomédicos. Foi desenvolvido um protótipo de sistema que implementa parcialmente o modelo. Questionários foram usados com autores para embasar o desenvolvimento do protótipo. O protótipo foi testando com pesquisadores-autores. Foram identificados quatro padrões de raciocínio e encadeamento dos elementos semânticos em artigos científicos. O modelo de conteúdo foi implementado como uma ontologia computacional. Foi desenvolvido e avaliado um protótipo de uma interface *web* de submissão artigos pelos autores a um sistema eletrônico de publicação de periódicos que implementa o modelo.

Palavras-chave publicações eletrônicas; metodológica científica; comunicação científica; representação do conhecimento; ontologias; processamento semântico de conteúdos; e-Ciência

A semantic model for electronic publishing

Abstract Electronic publishing, although Information Technologies advancements, are still based in the print text model. The textual format prevents programs to semantic process articles content. A semantic model of scientific electronic publishing is proposed, in which conclusion are prompted by author and recorded in machine-understandable format, enabling semantic retrieval, identification of traces of scientific discoveries and knowledge misunderstandings. The model is based on concepts as deep, or semantic, structure of human language (CHOMSKY, 1975), of microstructure, macrostructure and superstructure (KINTSH & VAN DIJK, 1972), of rhetoric structure of scientific articles (HUTCHINS, 1977), (GROSS, 1990) and on scientific methodology semantic elements, such as problem, question, objective, hypothesis, experiment and conclusion. It results from analysis of 89 biomedical articles. A prototype system was developed which partially implements the model. Questionnaires with authors were used to test the prototype development. The prototype was also tested with several researchers-authors. Four

* Doutor em Ciência da Informação pela Universidade Federal do Rio de Janeiro (UFRJ), pesquisador do CNPq e professor de Ciência da Informação na Universidade Federal Fluminese (UFF). Endereço postal: UFF, Depto de Ciência da Informação, Rua Lara Vilela, 126, São Domingos, Niterói, Rio de Janeiro, CEP. 24210-590. Tel. (21) 2629-9758 e e-mail marcon@vm.uff.br.

patterns of reasoning and sequencing of semantic elements were identified in articles analyzed. The content model is implemented as a computational ontology. A prototype of a web author's submission interface to a electronic journal system was developed and tested.

Keywords electronic publishing; scientific methodology; scientific communication; knowledge representation; ontologies; semantic content processing; e-Science

Introdução

Antes do surgimento da *web* o acervo de conhecimento científico validado da humanidade era difuso e armazenado de forma dispersa em coleções de periódicos em bibliotecas. Atualmente existem duas principais barreiras para a utilização em larga escala desse conhecimento: a grande quantidade de informação disponível em toda a *web* e o fato de que esse conhecimento estar incorporado no texto de artigos científicos, legível somente por seres humanos, não adequado para o processamento "semântico" por computadores.

Desde das Actas da Royal Society, no século XVII, artigos científicos são canais privilegiados de comunicação científica. Através deles autores fazem de suas descobertas, conhecimento público. Atualmente, a publicação eletrônica é uma atividade rotineira para acadêmicos e pesquisadores. A maioria das revistas científicas esta agora disponível na *web*. O potencial das tecnologias da informação (TI) tem sido aplicado a sistemas modernos de informação bibliográfica para melhorar a comunicação científica, proporcionando o acesso rápido e notificação imediata e acesso ao texto integral de artigos científicos. Mas esse potencial não tem sido usado para processar diretamente o conhecimento incorporado no texto de artigos científicos. Artigos publicados na *web* são "bases de conhecimento" (GARDIN, 2001). Apesar do formato digital desses textos, o que abriria muitas potencialidades, estas "bases de conhecimento" são apenas para leitura humana. Seu conteúdo necessita de leitura crítica, de avaliação, de ser citado, tudo isso através de um longo processo social, até que o conhecimento contido nesses textos possa ser finalmente incorporado aos estoques de conhecimento humano.

Estudos consistentes com dados recolhidos por extensos períodos (TENOPIR; KING, 2009) comprovam que pesquisadores precisam ler um número cada vez maior de artigos e dispensam cada vez menos tempos na leitura de cada artigo. Devido à "explosão informacional" tornada crítica coma *web* e as publicações científicas eletrônicas, os pesquisadores têm sempre lido estrategicamente (RENEAR; PALMER, 2009), trabalhando com diferentes artigos e fontes de informação simultaneamente, comparando-os, analisando os fragmentos de texto. Hoje há uma necessidade urgente de novas ferramentas para melhorar a leitura estratégica em ambiente *web*.

Uma quantidade crescente de registros da cultura humana, desde textos, imagens estáticas e de movimento, som, multimídia, etc, estão sendo criados diretamente em formato digital. Metadados são essenciais para a gestão desse amplo espectro de registros em um ambiente digital complexo. Desde que o registro *Machine Readable Cataloging (MARC)* foi criado na década de sessenta, modelos de registros bibliográficos evoluíram, incorporando, no entanto, apenas alterações menores. Na forma convencional, registros bibliográficos são conjuntos de campos de banco de dados, incluindo uma série de palavras-chave para a descrição do conteúdo, usadas para fins de recuperação, cada uma delas com igual peso e sem qualquer tipo de

relacionamento umas com as outras. Novos elementos de metadados foram adicionados a algumas propostas recentes de registros bibliográficos, do simples Dublin Core Metadata Element Set até o complexo *Metadata Enconding and Transmition Standard (METS)*[1], destinado a apoiar as tarefas de gerenciamento eletrônico de documentos, incluindo a preservação a longo prazo, gerenciamento de direitos, etc, além de recuperação de informação. Modelos de registros bibliográficos como *Functional Requirements for Bibliographic Records - FRBR* (1998) da IFLA são destinadas a recuperar diferentes versões da mesma obra e não especificamente o seu conteúdo. Acesso por conteúdo a documentos nos modernos sistemas de recuperação bibliográfica, incluindo bibliotecas digitais, repositórios, sistemas de publicação de periódicos, ainda é feito por comparação computacional de meros padrões de caracteres, das palavras-chave da consulta feita pelos usuários, unidas através dos 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.

Buscas por palavras-chave ligadas pelos 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*. Técnicas de mineração de dados e de textos, quando aplicadas à recuperação de informação, se mostram como técnicas de busca "cega", com base somente no poder de computacional, não conseguem identificar significados. São baseados somente em técnicas computacionais de correspondência entre padrões de caracteres no termos de busca, que remontam aos primórdios da era do computador.

As tecnologias da *web* semântica (Berneers-Lee, 2001) propõem um passo adiante para a questão da recuperação e processamento semânticos de conteúdos em ambientes computacionais. Segundo esta proposta a descrição do conteúdo de um documento na *web* não é mais uma questão de combinar palavras-chave, como em ambientes computacionais convencionais desde os anos 60, mas consiste em conjuntos estruturados de conceitos ligados por relações de significado preciso, dado por padrões como em *Resource Description Framework - RDF* (2004) e RDF Schema (2000). Construídas com base no RDF Schema, ontologias computacionais, codificadas na *Web Ontology Languague - OWL* (2004), organizam o conhecimento em domínios específicos, registrando conceitos acordados por comunidades, organizados em hierarquia de classes e subclasses, em propriedades desses conceitos, em relações entre eles e em regras lógicas para aplicá-los a esse domínio. Esse rico esquema de representação "semântica" permite a agentes de software para executar "inferências" e tarefas sofisticadas com base no conteúdo de documentos.

Inspirando-se na proposta da *web* semântica, esta pesquisa propõe um modelo semântico de publicações cientificas eletrônicas, capaz de extrair o conhecimento do texto do artigo e representá-lo em formato "inteligível" por programas. O modelo, quando implementado, permitirá recuperação de informações de forma semanticamente muito mais rica, além de viabilizar a identificação de indícios de novas descobertas científicas e de incoerências no conhecimento veiculado em artigos científicos. O artigo esta estruturado da seguinte forma: após esta introdução, a segunda seção discute novos modelos e propostas de publicações científicas eletrônicas, em especial alguns utilizando as tecnologias da *web* semântica; a terceira seção discute matérias e métodos utilizados na pesquisa; a quarta seção apresenta os resultados, descrevendo o modelo proposto e a sua implementação parcial através do protótipo de uma interface *web* de submissão de artigos a sistemas de periódicos eletrônicos que processa lingüisticamente conclusões de artigos, registrando-as em formato "inteligível" por programas; finalmente a seção 5 apresenta conclusões e futuros desenvolvimentos da pesquisa.

[1]. Disponível em http://www.loc.gov/standards/mets/.

Novos modelos de publicações científicas eletrônicas

O cenário da comunicação científica com base em periódicos impressos e disponibilizados através de coleções em bibliotecas vem evoluindo rapidamente na atualidade, tendo como vetor principalmente a evolução das tecnologias da informação, para um modelo de acesso direto a textos completos de publicações eletrônicas. Com o surgimento da *web* diferentes comunidades científicas estão engajadas hoje no desenvolvimento e disponibilização públicas de ontologias computacionais em domínios específicos[2] como mecanismos de registro, reuso e intercâmbio de conhecimento. É previsível que este processo se acentue nos anos seguintes, modificando radicalmente as formas como a humanidade registra, armazena e usa o conhecimento científico.

Várias alternativas vêm sendo tentadas no sentido propor novos modelos de publicações que endereçam as questões levantadas anteriormente e tentam tirar partido das tecnologias da *web* Semântica a fim de otimizar a comunicação científica, a gestão, o compartilhamento e reuso do conhecimento e o acesso semântico aos conteúdos dos artigos científicos. Todas as alternativas identificadas têm de comum o fato de caminharem na direção de uma maior formalização dos textos de artigos científicos, com vistas a otimizar sua leitura e compreensão além de permitir identificar univocamente seu sentido. A seguir comentamos, em primeiro lugar suas bases conceituais e, a seguir, estas experiências.

A estrutura de textos tem sido objeto de intensa pesquisa pela lingüística em geral, em especial pela lingüística computacional. Noam Chomsky (1975) afirma que todo texto possui, subjacente à sua estrutura superficial ou lingüística, uma estrutura profunda ou semântica. Kintsh e Van Dijk (1972) propõe um modelo para a estrutura de textos formado por microestrutura, a sequência de proposições dentro do texto, e a macroestrutura, os elementos semânticos que formam um esquema, específico de um tipo de texto como uma estória, um registro médico ou um artigo científico.

Baseado em Kintsh e Van Dick, Hutchins (1977) aplicou esse esquema aos artigos científicos. Ele considerou que o texto de artigos tinha uma específica sequência de elementos semânticos de acordo com um esquema pré-definido, através do qual cientistas apresentam seus argumentos. Ele também enfatiza que o texto de um artigo tem uma função retórica. Propõe uma classificação para artigos científicos, composta de artigos que testam hipóteses ao lado de outros, de caráter exploratório, que usam a abdução para buscar hipóteses que expliquem um fenômeno. Gross (1990) também enfatiza a natureza retórica dos artigos científicos e os classifica em teóricos, equivalentes à classificação de Hutchins de artigos exploratórios, e experimentais, equivalentes à classificação de Hutchins de artigos que testam hipóteses. A estrutura retórica do texto é analisada e discutida Swales (1990) e Nwogu (1997).

Ao analisar a estrutura dos artigos científicos, Kando (1997, 1999), divide a estrutura tradicional *Introduction, Material and Methods, Results, Discussion and Conclusion (IMRAD)*, chamada por ele de primeiro nível, em dois níveis adicionais, a fim de identificar vários outros elementos, tais como dados e argumentação adicional do autor, etc. A identificação e marcação desses elementos são usados para facilitarem a recuperação do conteúdo do artigo. A proposta de Kando considera artigos científicos como tendo sempre a mesma estrutura.

[2] Ver em OBO, Open Biological and Biomedical Ontologies, http://www.obofoundry.org/.

Outras experiências relacionadas aos objetivos dessa pesquisa envolvem tentativas de usar a linguagem XML – Extensible Markup Language – para marcação e publicação de artigos científicos na *web*. Existem diferentes propostas, como a pioneira *Chemical Markup Language - CML* (MURRAY-RUST, RZEPA, 1999), a *System Biology Markup Language - SBML* (HUCKA, FINNEY, BOLORI, 2003), a *Mathematical Markup Language (MathML)* e também enfoques mais gerais como a *Scientific Technical and Medical Markup Language - STMML* (MURRAY-RUST, RZEPA, 2002). Outro importante experimento nesta direção é a *Text Encoding Initiative - TEI* (2005), que usa XML para marcação de textos acadêmicos em literatura e lingüística com o objetivo de facilitar a recuperação e a preservação de publicações eletrônicas. A *Data Documentation Initiative - DDI* (2004) – objetiva estabelecer um padrão internacional em baseado em XML para o conteúdo, preservação, transporte e preservação de documentação em bases de dados em ciências sociais.

Apesar dos paradigmas cogitivo/sociocognitivo considerarem conhecimento como um processo que ocorre na mente de indivíduos (ELLIS, 1992), (HJORLAND, 2002), a Ciência da Informação tem se preocupado sempre com a representação do conhecimento e este interesse tem evoluído na direção de representação do conhecimento em formatos legíveis por máquinas. Vickery (1986), numa revisão sobre representação do conhecimento, relaciona estruturas de registros em bancos de dados e arquivos, assim como estruturas de dados em programas de computadores como diferentes modelos e técnicas em representação do conhecimento. Buckland (1991) distingue "*information as knowledge*", um processo intangível que ocorre na mente de indivíduos, de "*information as a thing*", representações do conhecimento ou conhecimento registrado em textos, registros, imagens, etc. A Inteligência Artificial e as propostas semânticas de representação do conhecimento, tais como ontologias, têm como objetivos desenvolverem esquemas de representação do conhecimento não somente em formatos legíveis por pessoas, mas também em formatos inteligíveis por programas.

A Lógica foi um dos primeiros formalismos usados para representação do conhecimento desde as primeiras experiências de inteligência artificial na década de 80 (NILSSON, 1980). Neste sistema, o conhecimento é representado como "regras de produção", consistindo de argumentos e relações entre eles. Recentemente *Knowledge Interchange Format (KIF)* – usou a Lógica como um formalismo para representação do conhecimento. Recentemente ocorre uma mudança do paradigma lógico para o ontológico: "AI researchers seem to have been much more interested in the nature of reasoning rather than in the nature of the real world. The potential value of task-independent bases (or "ontologies") suitable for large-scale integration has been underlined in many ways". (GUARINO, 1995). A pesquisa atual evolui na direção da integração de ambas as visões, conforme Sowa (2000): "Ontologies… supply the predicates of predicate calculus and the labels that fill the boxes of conceptual graphics".

Entre outros, os projetos mais relacionados com o aqui descrito são: *Communications in Physics* (2001); the *Scholarly Ontologies Project* (2004), desenvolvido na Open University, Reino Unido; *Research in Semantic Scholarly Publishing* (2005), da Biblioteca da Universidade Erasmus de Rotterdam; o projeto *Writing in the Context of Knowledge*" (CARR et al. 2004), do Laboratório de Inteligência, Agentes e Multimídia da Universidade de Southampton, Reino Unido, a Ontologia para a autopublicação de experimentos da *Scientific Publishing Task Force* (2006); o projeto *SWAN* (GAO et al. 2006); a *Basic Ontology for Scientific Study - BOSS* (2004); o projeto *ArkeoteK* (2002), (GARDIN, 2001); a *EXPO* – uma ontologia para experimentos científicos (SOLDATOVA e KING, 2006); HyBrow (RACUNAS et al. 2004), um sistema que objetiva auxiliar cientistas na formulação e avaliação de hipóteses com relação ao conhecimento prévio; o trabalho de Hunter e coautores (2008), que objetiva identificar conceitos para extração de relações de interação entre proteínas em textos biomédicos; propostas como as de Dinakarpandian et al. (2006), propõe formalizar as afirmações feitas em artigos científicos em

formato "inteligível" por programas; finalmente, sendo a comunicação científica uma fase decisiva em qualquer pesquisa, o projeto *Ontology for Biomedical Investigations - OBI* (2008) - uma ontologia mantida pela OBO Foundry (2010) que tem como objetivo ser uma referência de alto nível relativa à pesquisa biológica -, foi considerada no desenvolvimento da ontologia proposta nesta pesquisa.

Outros exemplos vêm de publicações científicas correntes, como o projeto *Prospect*[3], da Royal Society of Chemistry, no qual termos no texto de artigos que se referem a entidades químicas ou biológicas possuem "links" para ontologias ou dicionários que as definem; o grupo editorial Elsevier desenvolve o projeto *Article of the Future*[4] em cima do periódico biomédico Cell, com o objetivo de adicionar diversas funcionalidades aos artigos, incluindo mudança de forma de apresentação – apresentações hierárquicas -, resumos gráficos, uma seção *Highlights*, onde são destacadas de forma sucinta as conclusões do artigo, etc., facilidades estas que só são possíveis num ambiente *web* de artigos digitais. Na página do projeto estão disponíveis dois artigos experimentais, que ilustram as facilidades implementadas pelo projeto. Shotton et al. (2009) descrevem a experiência de uso de diferentes tecnologias semânticas na publicação[5] *Public Library of Science (PloS*) - incluindo ontologias biomédicas, comentários nos artigos e uma ontologia de tipos ou motivos para citações. Um número crescente de publicações científicas, em especial na área biomédica[6], como o *British*[7] *Medical Journal (BMJ), Journal of American Medical Association (JAMA),* entre outros, vem usando resumos estruturados (GUIMARÃES, 2006) como forma de otimizar a apreensão do conteúdos dos artigos.

Vemos assim que as experiências recentes em publicações eletrônicas caminham na direção de formalizar cada vez mais o texto dos seus artigos, quer estruturando-os, marcando-os e identificando partes significativas para facilitar uma leitura mais direta por humanos, quer relacionando esse texto a ontologias computacionais formais como meio de superar as possíveis ambiguidades dos textos e permitir seu processamento "semântico" por programas.

Material e métodos

Para a proposição do modelo foram buscados aportes teóricos de disciplinas como Ciência da Informação, em especial de Comunicação Científica, Metodologia da Ciência, Filosofia da Ciência e Ciência da Computação, referenciados anteriormente. Foram analisados 89 artigos em Medicina, subdivididos nos seguintes grupos: 20 artigos do periódico *Memórias do Instituto Oswaldo Cruz*, 20 artigos do periódico *Brazilian Journal of Medical em Biological Research,* ambos disponibilizados através do portal SciELO e escolhidos a partir da lista dos artigos mais consultados de ambas as publicações; foram analisados ainda 20 artigos sobre células-tronco, escolhidos a partir de três importantes artigos de revisão sobre o tema; outro grupo são artigos que relatam uma descoberta relevante em biomedicina, a descoberta da enzima telomerase. Fazem parte desse grupo 15 artigos entre as chamadas *key publications* do grupo de pesquisadores agraciados com o *Prêmio Albert Lasker de Medicina* do ano de 2006 e outras 14 publicações reportando os desenvolvimentos mais recentes da pesquisa sobre telomerase.

[3] Disponível em http://www.rsc.org/Publishing/Journals/ProjectProspect/.
[4] Disponível em http://beta.cell.com/.
[5] Disponível em http://www.plos.org/.
[6] Disponível em http://www.bmj.com/.
[7] Disponível em http://jama.ama-assn.org/.

A área de Medicina foi escolhida devido ao fato de que artigos científicos da área seguem um rígido padrão formal em seus textos, com seções definidas segundo o chamado padrão *Introduction, Method, Results and Discussion (IMRAD*) -, recomendados pelo *The International Committee of Medical Journals Editors*[8] para artigos científicos em periódicos biomédicos, facilitando assim a análise.

O modelo inclui o uso de uma base terminológica disponível na *web* para verificar até que ponto o conteúdo de cada artigo estava representado nesta. Para os artigos analisados, no domínio da biomedicina, foi usada a *Unifyed Medical Language System (UMLS),* uma grande e amplamente usada base terminológica no domínio da biomedicina. A UMLS vem caminhando na direção de se tornar uma ontologia formal, na qual termos biomédicos estão relacionados por relações formais, de semântica precisa e definida (BODENREIDER, 2008). Na documentação da UMLS pode-se encontrar a seguinte afirmação: *"*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*".*[9]

Foi desenvolvido e testado um protótipo de uma interface *web* de submissão de artigos a sistemas de gestão de periódicos eletrônicos que formata as conclusões dos artigos fornecidas pelos autores como relações semânticas; no seu desenvolvimento foi usado o programa *MetaMap* de processamento linguístico de textos biomédicos[10], que identifica em textos biomédicos termos controlados do UMLS Thesaurus.

Resultados

Trabalhamos há anos (MARCONDES, 2005) na uma 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 programas realizem "inferências" sobre este conhecimento, permitindo processar o conhecimento assim recuperado e processado de forma semanticamente mais rica que os atuais SRIs. Este modelo é descrito a seguir e pode ser subdividido em dois: um modelo semântico de conteúdo de artigos e a proposta de uma *interface web* de autosubmissão de artigos a sistema de gestão de periódicos eletrônicos.

Um modelo semântico de publicações eletrônicas

Relações são o elemento essencial do esquema de representação do conhecimento proposto. Relações são expressas por três elementos: dois relata e um tipo de relação. Os dois relata – Antecedente e Consequente - 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

[8] Disponível em http://www.icmje.org.
[9] Disponível em http://www.nlm.nih.gov/pubs/factsheets/umls.html.
[10] Disponível em http://mmtx.nlm.nih.gov/.

afirmações feitas pelo autor no artigo são representadas através de triplas como Antecedente-Tipo de Relação-Consequente. 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 (Consequente, 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 científico: 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).

O modelo proposto pode ser dividido em dois componentes, interligados e complementares. Primeiro, o modelo Ontologia do Conteúdo de Conhecimento em Artigos Científicos (OCA) – uma ontologia para representar o conhecimento contido em artigos biomédicos, que considera que este conhecimento consiste em afirmações feitas pelo autor ao longo do texto do artigo, expressando relações entre fenômenos ou relações entre um fenômeno e suas características. Segundo, a proposta de uma *interface* web a um Sistema de Submissão de artigos a publicações eletrônicas, na qual, além do autor fazer o "upload" do arquivo com o texto do seu artigo e entrar com os metadados bibliográficos convencionais (como Autor, Título, Palavras-chave, etc.), informará também as principais afirmações feitas no seu artigo.

A figura a seguir mostra um esquema do modelo semântico de publicações eletrônicas, identificando seus componentes. São eles: *interface web* a um Sistema de Submissão de artigos a publicações eletrônicas, a Base de Dados, a Ontologia pública na *web* e a Ferramenta de Identificação de Novidades.

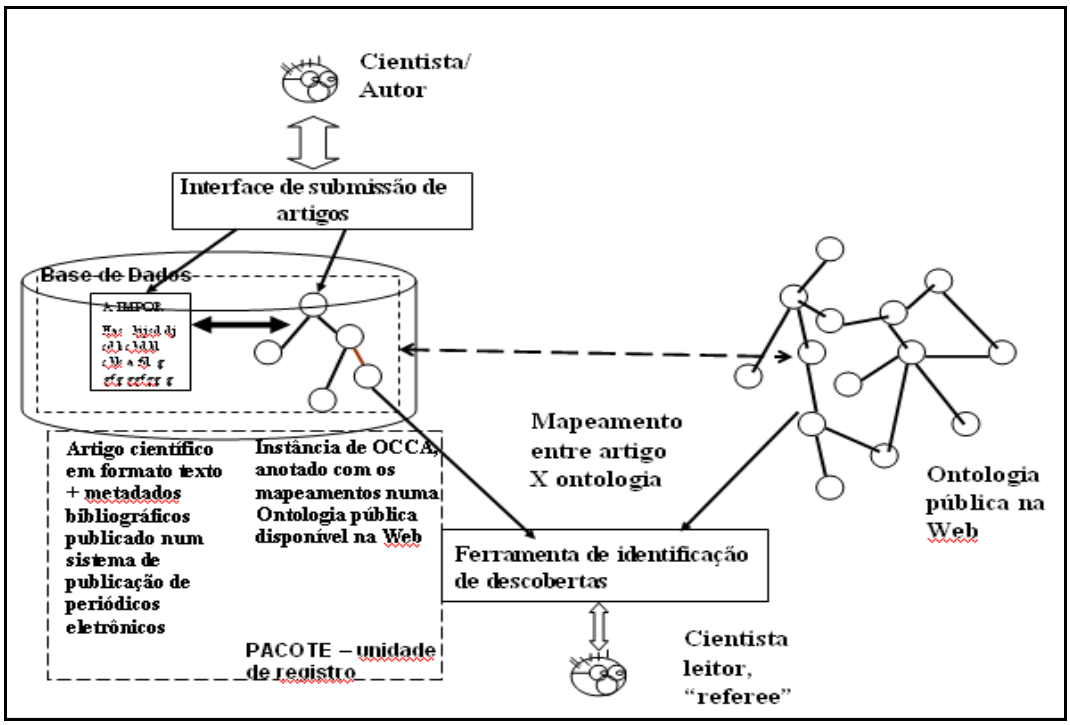

Figura 1: Visão geral dos componentes do modelo semântico de publicações eletrônicas

Modelo semântico de conteúdo de publicações eletrônicas

Com foi explicado na seção Material e Métodos, o modelo semântico proposto é o resultado da análise de um total de 89 artigos de periódicos em biomedicina, com a finalidade de identificar elementos semânticos de metodologia científica e padrões de raciocínio e encadeamento que combinassem estes elementos.

Quatro tipos de artigos foram identificados – Teóricos, Experimentais-exploratórios, Experimentais-indutivos e Experimentais-dedutivos – que expressam diferentes raciocínios, estratégias de argumentação e pressupostos de um artigo. Maiores detalhes podem ser encontrados em Marcondes (2009). Como pode ser visto dos exemplos anteriores, dos componentes semânticos que formam o modelo OCA, excetuando o Experimento, todos os outros – Questões, Hipóteses e Conclusões - podem ser expressos sob a forma de relações.

A análise permitiu identificar os seguintes elementos semânticos, que compõe o modelo de conteúdo.

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(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; isso em um artigo 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.

Esses elementos semânticos se organizam no texto de artigos segundo os seguintes tipos de artigo e padrões de encadeamento de raciocínio:

- 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.

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 (LAURA, 2004) ao desvendar e buscar caracterizar um fenômeno, trabalhando na direção proposta por Dahlberg (1995) de formular e provar proposições que caracterizam um fenômeno.

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 veem 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 OCA podem ser representados como uma ontologia como pode ser visto através da seguinte Figura, ilustrando sua hierarquia de classes e propriedades.

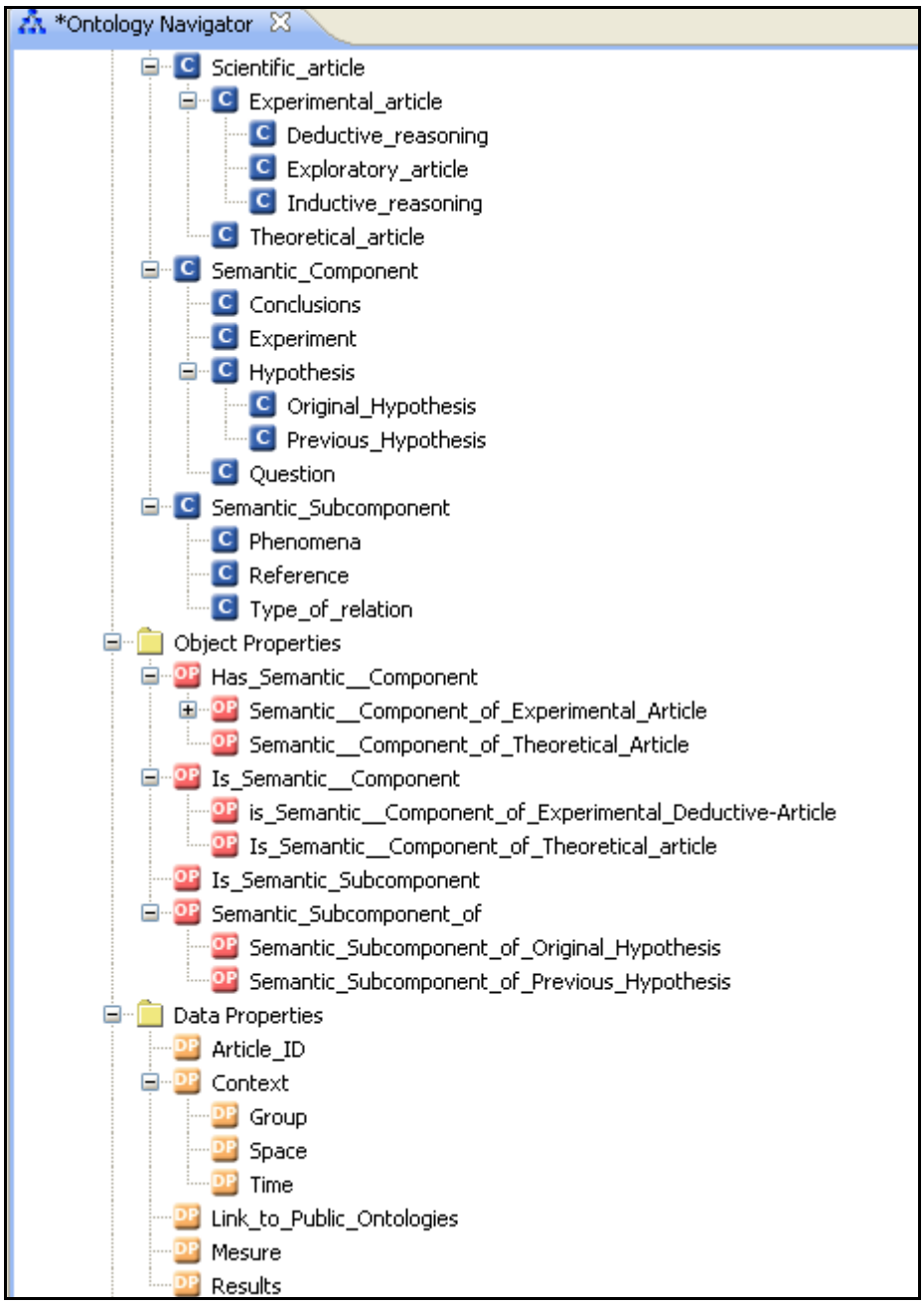

Figura 2: Modelo OCA de Representação do Conhecimento implementado como uma ontologia, visualizada através do programa NeOnToolkit

A importância de um esquema de representação de conhecimento contido em artigos que utilize relações esta em que estas permitem que programas façam "inferências" sobre o conhecimento assim representado, como nos seguintes exemplos:

- O papiloma vírus (Antecedente) causa (Relação) que outros tipos de câncer (Consequente?)?

- Que outras (Antecedente?) causas (Relação) pode ter o câncer de colo de útero (Consequente?) além do papiloma vírus?

Interface web de autosubmissão de artigos

A *interface web* a um sistema de submissão de artigos a publicações eletrônicas, como nas diversas interfaces de submissão hoje tão comuns, permite a um autor submeter o arquivo com o texto do seu artigo juntamente com os metadados bibliográficos usuais utilizados para descrevê-lo. A diferença desta proposta das interfaces de submissão convencionais é que o autor é solicitado a entrar também com as afirmações que constituem a síntese do conhecimento contido no artigo, em especial, com suas conclusões.

O projeto de pesquisa de Costa (2006) desenvolve o modelo de conteúdo proposto nessa pesquisa e tem como objetivo implementar (parcialmente) o protótipo de uma interface *web* a um sistema de submissão, pelos autores, de artigos a um periódico eletrônico. Além dos metadados bibliográficos convencionais, a interface solicita que o autor entre também com o texto das conclusões do artigo. Através dessa interface os próprios autores entram num diálogo interativo, respondendo a questões colocadas pela interface. A interface realiza processamento de linguagem natural, tanto em trechos do texto do artigo quanto das respostas do autor. O objetivo da interface *web* é formatar as afirmações feitas pelo autor – as conclusões do artigo - através do processamento lingüístico do texto informado por este, formatando-as segundo relações semânticas (COSTA 2006, 2008), compostas por um Antecedente, um Tipo-de-relação e um Consequente. As relações assim formatadas são registradas em formato "inteligível" por programas, segundo o modelo de conteúdo OCA, sendo instâncias desta ontologia e fazendo parte de um pacote, juntamente com os metadados bibliográficos convencionais e o texto digital do artigo. Esses pacotes podem ser tratados por SRIs semânticos, possibilitando buscas semanticamente mais ricas.

Para atingir este objetivo a interface solicita ao autor que entre com o texto do artigo, os metadados bibliográficos convencionais e as conclusões do artigo. A interface processa o texto destes elementos em quatro etapas: Extração do Objetivo, Formatação da Relação, Mapeamento dos elementos da Relação nos termos da UMLS e Representação da Relação em formato "inteligível" por computador.

- Extração do Objetivo – a extração do objetivo baseia-se na identificação de frases indicativas (por ex. *The aim of our study is*) em partes específicas do texto (*abstract* e *introduction*), por causa da concentração do objetivo nessas partes (SWALES, 1990), (NWOGU, 1997).

- Formatação da Relação – após a conclusão ser informada pelo autor, esta é processada pelo programa MMTX que identifica tanto componentes sintáticos da frase que expressa a conclusão (NOUN_PHRASE, VERB_PHRASE, ADVERB_PHRASE, PREPOSITIONAL_PHRASE) quanto se estes componentes são termos biomédicos existentes na UMLS. Os candidatos a serem os fenômenos que formam Antecedente e Consequente são os sintagmas nominais e os sintagmas preposicionais; os candidatos a formarem o Tipo-de-Relação são os sintagmas verbais. É feita então um calculo de peso dos sintagmas nominais e preposicionais, com base na sua frequência em trechos específicos do texto do artigo, como Abstract, Introdução, Método, Resultados, Título, Palavras-chaves e Objetivo. Identificados os dois sintagmas não contíguos com maior peso como os candidatos a Antecedente e Consequente, busca-se a existência de um sintagma verbal entre eles, como candidato ao Tipo-de-Relação. Uma vez identificada a Relação completa pela interface, esta é apresentada ao autor para que ele a valide.

- Mapeamento dos elementos da Relação – a interface consulta uma ontologia ou base terminológica pública no mesmo domínio que os artigos submetidos para tentar identificar termos nos Antecedente, Consequente e Tipo-de-Relação que estejam presentes nessa ontologia pública. No caso dos artigos analisados, cujo domínio é a biomedicina, a ontologia ou terminologia utilizada foi a UMLS. O resultado da tentativa de mapeamento dos termos nos Antecedente e Consequente, e o Tipo-de-Relação é mostrado para que o autor os valide, dizendo se concorda com os termos apresentados pela interface ou não. O resultado do mapeamento é registrado e servirá posteriormente como insumo para a identificação de novas descobertas, que é outro componente dessa pesquisa (MALHEIROS, 2010).

- Representação da Relação em formato "inteligível" por computador – depois de validada pelo autor, a relação obtida, o resultado do mapeamento, juntamente com o texto completo do artigo e os metadados bibliográficos convencionais, são registrados numa base de dados. A relação é representada como uma instância da ontologia OCA. Esta etapa ainda não esta implementada.

A sequência de telas seguinte mostra a interação autor-interface.

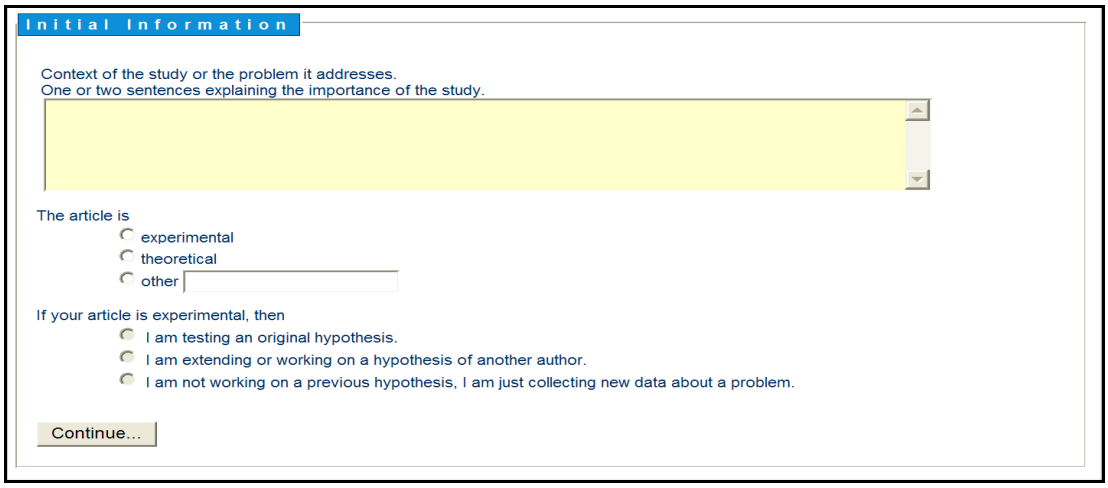

Figura 3: Autor, após entrar com os metadados bibliográficos, informa o tipo de raciocínio usado no artigo

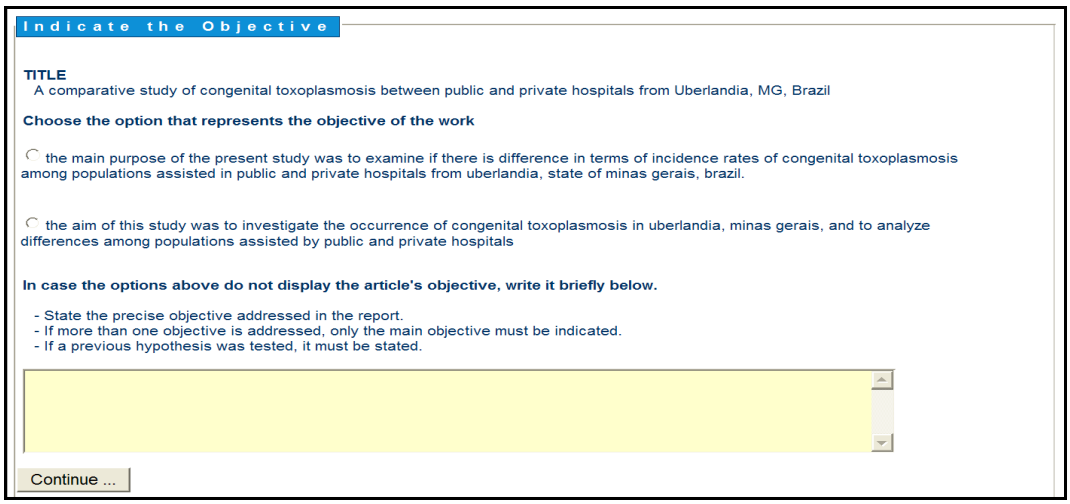

Figura 4: Autor valida o objetivo do artigo, extraído pela Interface

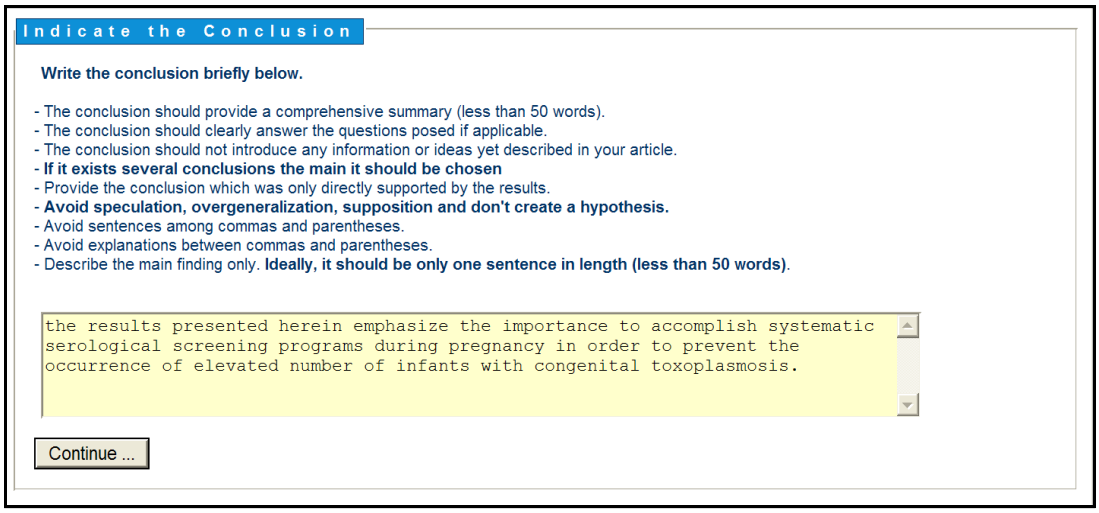

Figura 5: Autor entra com a conclusão do artigo

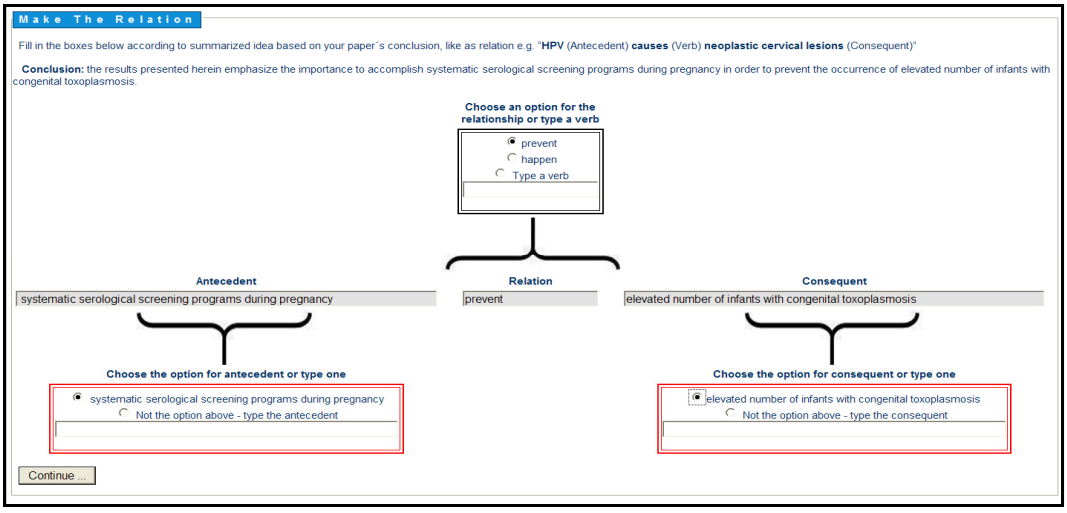

Figura 6: A conclusão do artigo é formatada segundo uma relação

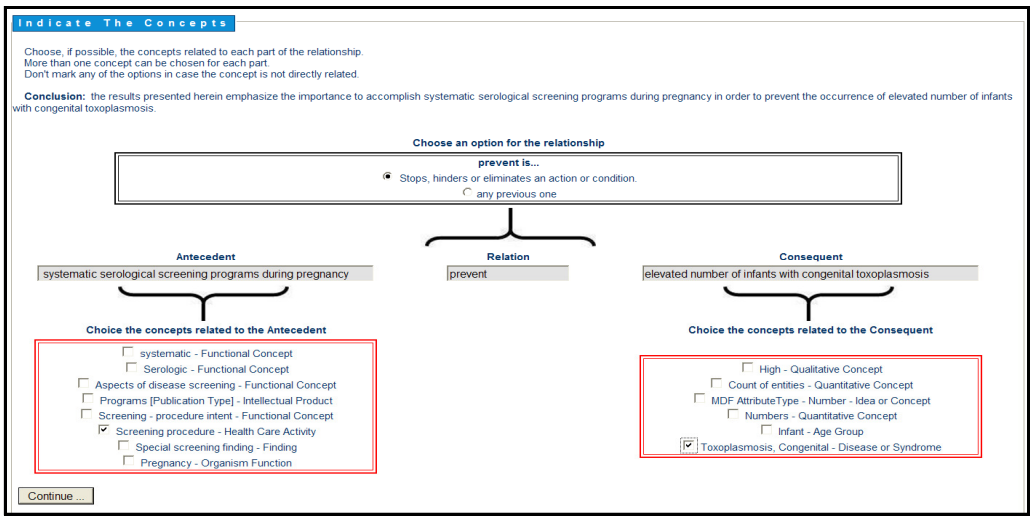

Figura 7: O autor é solicitado a mapear os conceitos contidos na conclusão em termos da UMLS

O protótipo da interface esta em sua fase inicial de desenvolvimento. Além das 10 entrevistas, o protótipo foi testado com 5 desses 10 autores e, em todos os casos, conseguiu formatar a conclusão do artigo segundo uma relação.

Conclusões

Consideramos que o momento da submissão de um artigo pelo seu autor a um sistema de gestão de publicações eletrônicas, como o Sistema Eletrônico de Edição de Revistas (SEER) por exemplo, é um momento especial, em que o autor esta especialmente motivado a prestar informações sobre seu artigo e suas conclusões. O processamento lingüístico do texto de uma conclusão, por ser um texto curto, é consideravelmente mais fácil e menos complexo que processar o texto completo do artigo, como em outras propostas; além disso, próprio autor esta presente para tirar as ambiguidades e validar o processamento lingüístico feito pelo sistema. Outras partes do texto do artigo como título, resumo, palavras-chave, introdução, objetivos, são utilizadas para pesagem dos termos e permitirão escolher quais deles irão compor o Antecedente e o Consequente da relação.

Está previsto também que a *interface* interaja com o autor, permitindo que este navegue pela ontologia pública ou base terminológica (como mostrado na figura 7), no mesmo domínio que os artigos submetidos à interface, e verifique, no momento da submissão/publicação do artigo, se o conteúdo dessas relações pode ser mapeado com conceitos existentes nessa ontologia pública. Esta é a etapa denominada "Mapeamento dos elementos da Relação". O resultado da avaliação do autor em relação aos termos da ontologia pública apresentados pela interface como possíveis equivalentes aos termos contidos no Antecedente, no Tipo-de-Relação e no Consequente, isto é, se o autor concorda ou não com o termos sugeridos como equivalentes pela interface e os códigos ou *links* para os termos que o autor concorda, são também registrados juntamente com a relação propriamente dita, como parte do pacote de dados. Também é registrado para cada artigo

assim representado um "link" para a ontologia pública usada nos mapeamentos (uma espécie de "meaning spaces" da relação, equivalente aos *name spaces*[11] da linguagem XML) e a data em que o mapeamento foi feito.

Desta forma o resultado desse mapeamento e avaliação do autor poderá ser consultado posteriormente pela Ferramenta de Identificação de Novidades, permitindo verificar até que ponto conteúdos veiculados no artigo já foram reconhecidos naquele domínio cientifico e incorporados na ontologia ou terminologia que representa os conceitos neste domínio. A incorporação de novos conceitos científicos em ontologias/terminologias como a UMLS costuma ser um longo processo social de discussão dentro de um domínio científico específico, que demanda tempo até que um consenso conceitual e terminológico possa ser atingido; portanto, o não mapeamento ou o mapeamento parcial podem ser usados como indicadores de novidades científicas (MALHEIROS, 2010).

Exemplos de consultas que poderiam ser feitas numa futura Ferramenta de Identificação de Novidades seriam as seguintes:

- Que artigos tratam (tem como Antecedente ou Consequente, mais o Tipo-de-Relação) de infecções na medula óssea e que foram parcialmente mapeados na UMLS?

- Que artigos que tratam de causas (Relação) de diskeratosis congênita não foram mapeadas na UMLS? (Antecedente não mapeado).

O uso da UMLS, que possui uma estrutura classificatória como o *Semantic Network*[12] na qual estão presentes como categorias, além dos 134 tipos semânticos, 54 tipos de relações, facilitou a proposta de representação do conhecimento como relações e ajudou o processamento linguístico para identificá-las nas conclusões. No entanto outras ontologias biomédicas vêem incorporando uma tipologia de relações[13], o que permite supor que o modelo proposto possa trabalhar com outras ontologias e bases terminológicas.

O conhecimento contido no artigo sob a forma de relações, bem como o resultado do mapeamento validado pelo autor, são registrados como instâncias da ontologia OCA. Estas instâncias são então gravadas como um "pacote" na *Base de Dados*, juntamente como o texto do artigo e seus metadados bibliográficos convencionais. Assim, todas as características e funcionalidades de um SRI convencional são mantidas no modelo, acrescidas das possibilidades de recuperação semântica e identificação de descobertas exemplificadas anteriormente.

Esta *Base de Dados* pode ser acessada por usuários através de Sistemas de Recuperação de Informações, provendo mecanismos para busca semântica, identificação de indícios de novidade, inconsistências no conhecimento veiculado em artigos e diversos outros tipos de aplicações.

Os resultados do protótipo da interface são ainda iniciais. Serão necessários mais testes reais, com pesquisadores-autores usando o protótipo da interface para simularem a submissão de seus artigos. A importância do protótipo é que ele materializa o modelo proposto e pode permitir a avaliação dos seus pressupostos num conjunto mais amplo de usuários. Pretende-se também incluir o protótipo, com a rotina de extração de relações, em interfaces de sistemas de gestão de publicações eletrônicas como o SEER[14].

[11] Ver em http://www.w3.org/TR/REC-xml-names/.
[12] Disponível em http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html.
[13] OBO Relationship Types. Disponível em http://www.obofoundry.org/ro/.
[14] Disponível em http://www.ibict.br/secao.php?cat=seer.

Agradecimentos

Esta pesquisa foi apoiada, em diferentes momentos, pelo CNPq, Capes, Faperj. e Propp/UFF.

Artigo recebido em 01/02/2011 e aprovado em 10/02/2011.

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.

The ArkeoteK Project. 2002. Disponível em: <http://www.arkeotek.org/>. Acesso em10 Jun. 2006.

BODENREIDER, O. Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support. In: *IMIA Yearbook of Medical Informatics*, :p. 67-79, 2008.

BUCKLAND, Michel. *Information and Information Systems* (Westport (CT). Praeger/Greenwood, 1991.

CARR, L.; MILES_BOARD, T.; WOUKEU, A.; WILL, G.; HALL, W. The case for explicit knowledge in documents**.** In: *Proceedings of the 2004 ACM Symposium on Document Engineering*, Milwaukee, Wisconsin, 2004 (ACM, 2004) 90-98. Disponível em : <http://www.eprints.ecs.soton.ac.uk/9360/>. Acesso em 6 maio 2006.

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

CHOMSKY, Noan. *Aspectos da teoria da sintaxe.* In: Textos selecionados. São Paulo: Abril Cultural, 1975. (Os Pensadores, 44).

COMMUNICATIONS IN PHYSICS. 2001. Disponível em: <http://www.science.uva.nl/projects/commmphys>. Acesso em 15 Mar. 2005.

COSTA, Leonardo Cruz da. *Uma ferramenta para edição, extração e representação do conhecimento contido em artigos científicos publicados na web*. Projeto de Tese de Doutorado para ingresso no PPGCI UFF/IBICT. Niterói, 2006.

COSTA, Leonardo Cruz da; MARCONDES, Carlos Henrique. Um ambiente para edição, extração e representação do conhecimento contido em artigos científicos publicados na *web*. In: ENANCIB - ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, São Paulo, set. 2008, 9, *Anais...* (Poster).

DAHLBERG, Ingetraut. Conceptual structures and systematization. International Forum on Information and Documentation, v. 20, n. 3, July, 1995.

Data Documentation Initiative, 2004. Disponível em: <www.icpsr.umich.edu/DDI/codebook/index.html>. Acesso 29 fev. 2006.

DINAKARPADIAN, Deendayal; LEE, Yugyung; VISHWANATH, Kartik; LINGAMBHOTLA, ROHINI. MachineProse: An Ontological Framework for Scientific Assertions. Journal of the American Medical Informatics Association, v. 13, n. 2, Mar/Apr, p. 220-232, 2006. DOI 10.1197/jamia.M1910.

ELLIS, D. Paradigms and proto-paradigms in information retrieval research. In: P.Vakkari and B. Cronin (eds.), *Conceptions of Library and Information Science: historical, empirical and theoretical perspectives*. London: Graham Books, 1992. p. 165-186.

FRBR – FUNCTIONAL REQUIREMENTS FOR BIBLIOGRAPHIC RECORDS : final report / IFLA Study Group on the Functional Requirements for Bibliographic Records. München: K . G. Saur, 1998. (UBCIM Publications New Series).

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, v.* 4, n. 3, 2006. Disponível em: <http://www.websemanticsjournal.org/ps/pub/2006-17>. Acesso em 12 Dez.

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.

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.

GUARINO, Nicola. Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human Computer* Studies, v. 43, n. 5/6, p. 625-640, 1995.

GUIMARÃES, Carlos Alberto. Structured Abstracts. Narrative Review. *Acta Cirúrgica Brasileira* v. 21, n. 4, 2006. Disponível em < http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-86502006000400014 >. Acesso em 20 abril de 2009.

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.

HJORLAND, B. Epistemology and the sociocognitive perspective in information science, *Journal of the American Society for Information Science and Technology, v.* 53, n. 4, p. 257-270, 2002.

HUCKA, M.; FINNEY, A.; BOLORI, H. *System Biology Markup Language* (SBML) Level 1: structures and facilities for basic model definitions (2003). Available at: http://www.sbml.org/specifications/sbml-level-1/version-2/sbml-level-1-v2.pdf (access 2 Nov. 2005).

HUNTER, L.; BAUMGARTNER Jr, A.; LU, Z.; JOHNSON, H. L.; CAPORASO, J. G.; PAQUETTE, J.; LINDERMANN, A.; WHITE, E. K.; MEDVEDEVA, O.; 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 20 nov.20

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. Retrieved 14 Jul. 2005 from at: www.icmje.org.

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

KINTSH, W.; VAN DIJK, T. A. Towards a model of text comprehension and production, *Psycological Review, v.* 84, n. 5, p. 363-393, 1972.

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

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.

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.

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.

MARCONDES, Carlos Henrique; MENDONÇA, Marília Alvarenga Rocha; MALHEIROS, Luciana Reis; COSTA, Leonardo Cruz da; SANTOS, Tatiana. Cristina Paredes. Bases ontológicas e conceituais para um modelo do conhecimento científico em artigos biomédicos. Reciis, v. 3, n. 1, p. 19-30, 2009. Disponível em <http://www.reciis.cict.fiocruz.br/index.php/reciis/article/view/240/251>. Acesso em 8 abr. 2009.

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

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. Available at: http://journals.eecs.qub.ac.uk/codata/journal/contents/1_2/1_2pdfs/ds121.pdf (accessed 18 Sept. 2005).

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.

NILSSON, N.J. *Principles of Artificial Intelligence.* California: Tioga Publishing Co., 1980.

NWOGU, Kevin Ngozi. The Medical Research Paper: Structure and Functions. *English for Specific Purposes*, v. 16, n. 2, p. 119-138, 1997.

OBI – Ontology for Biomedical Investigations. 2008. Disponível em http://obi-ontology.org. Accesso 20 nov. 2008.

THE OPEN BIOLOGICAL AND BIOMEDICAL ONTOLOGIES. 2010. Disponível em < http://www.obofoundry.org/>. Acesso em 29 out. 2010.

OWL Ontology Web Language Overview. 2004. Disponível em: <http://www.w3.org/TR/owl-features/>. Acesso em 15 maio 2007.

RACUNAS, S. A.; SHAH, N. H.; I. ALBERT, I; FEDOROV, N. V. HyBrow: a prototype system for computer-aided hypothesis evaluation. *Bioinformatics, v.* 20, n.1, p. 257–264, 2004.

RDF Resource Description Framework. 2004. Disponível em: Retrieved January 7, 2007, from http://www.w3.org/RDF/. Acesso em 7 jan. 2007.

RDF Schema Specification. 2000. Disponível em: <http://www.w3.org/TR/2000/CR-rdf-schema-20000327/>. Acesso em 18 nov. 2010

RENEAR, Allen H.; PALMER, Carole. Strategic reading, ontologies and the future of scientific publishing. *Science*, v. 325, p. 828-832, 2009.

RESEARCH IN SEMANTIC SCHOLARLY PUBLISHING. 2005. Disponível em: htt://rssp.net/. Acesso em 13 Mar. 2006.

SCHOLARLY ONTOLOGIES PROJECT. 2004. Disponível em: <http://kmi.open.ac.uk/projects/scholarly>. Acesso em 12 Jun. 2005.

Scientific Publishing Task Force – Ontology for Experiment Self-Publishing, 2006. Disponível em <http://esw.w3.org/topic/HCLS/SciPubSPERequirements>. Acesso em 15 Maio 2006.

SOLDATOVA, L. D; KING, R. D. An ontology of scientific experiments. Journal of the Royal Society Interface v. 3 n. 11, p. 795-803, 2006. Disponível em <http://journals.royalsociety.org/content/u552845783800t73/fulltext.pdf>. Acesso em 5 Fev 2008.

SHOTTON, David; PORTWIN, Katie; KLYNE, Graham; MILES, Alistair. Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article. PLoS Comput. Biol. v. 5, n. 4, April, 2009. Disponível em <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2663789/>. Acesso em 27 jul. 2010.

SOWA, J. *Knowledge Representation: logical, philosophical and computational foundations.* Pacific Grove: Brooks/Cole, 2000.

SWALES, J. M. *Genre analysis: english in academic and research settings*. Nova Iorque: Cambridge University Press,1990.

TENOPIR, Carol; KING, Donald W. Electronic journals and changes in scholarly article seeking and reading patterns. *Aslib Proceedings: New Information Perspectives*, v. 61, n. 1, 2009. p. 5-32 Disponível em <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.156.2701&rep=rep1&type=pdf>. Acesso em 28 jun. 2010.

TEI: Text Encoding Initiative. 2005. Disponível em: http://www.tei-c.org. Acesso em 29 fev. 2006.

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