EDICIC
Asociación de Educación e Investigación en Ciencia de la Información de Iberoamérica y el Caribe

UMA PROPOSTA DE PROCESSO DE SUBMISSÃO DE ARTIGOS CIENTÍFICOS ÀS PUBLICAÇÕES ELETRÔNICAS SEMÂNTICAS EM CIÊNCIAS BIOMÉDICAS

Leonardo Cruz da Costa

Departamento de Ciência da Computação – Universidade Federal Fluminense (UFF) – Brasil

RESUMO

Propõe-se um processo de publicação e registro de artigos científicos em Medicina que torne possível representar em formato "inteligível" por programas o conhecimento nele contido. Segundo o processo, artigos científicos seriam publicados não só em formato textual, legível por pessoas, mas também como instâncias de uma ontologia (OCCA), representando o conhecimento específico contido em cada artigo.

Palavras-Chave: Publicações Eletrônicas; Representação do Conhecimento; Ontologias; Comunicação Científica; Publicações Semânticas.

ABSTRACT

A process of publication and register of scientific articles in Medicine which can make it possible to represent it in an intelligible format for programs the knowledge contained in it. According to this process, scientific paper would be published not only in textual format, readable by people, but also as instances of an ontology (OCCA) representing the specific knowledge contained in each article.

Keywords: Electronic Publishing; Knowledge Representation; Ontology; Scientific Communication; Semantic Publishing.

1 INTRODUÇÃO

Bibliotecas digitais necessitam de padrões para descrever o documento eletrônico, tornando possível a sua organização e a sua recuperação. Para isso, lança-se mão dos chamados metadados, que tem por fim descrever um objeto de modo a permitir sua identificação, localização, recuperação, manipulação e uso.

Especificamente os metadados bibliográficos criados de acordo com as regras de catalogação e padrões estruturais, incluem elementos originados de padrões como *Anglo-American Cataloguing Rules* (AACR2) e o *Machine Readable Cataloging Format* (MARC), definidos desde os Anos 60 de forma cooperativa

(GILLILAND-SWETLAND, 2000). Atualmente, padrões como o Dublin Core e *Metadata Encoding & Transmission Standard* (METS) são utilizados para descrever os documentos eletrônicos de forma sucinta, possibilitando a interoperabilidade entre diversas bibliotecas.

Os vários padrões existentes de metadados apresentam campos específicos (assunto, descritor, *subject*) utilizados para descrever os termos que representam os assuntos correspondentes ao documento com o objetivo de recuperá-lo posteriormente. A representação do documento é chave para sua recuperação; é nesta área que reside à relação histórica e indissociável da Ciência da Informação com a Ciência da Computação (SARACEVIC, 1995).

Os campos específicos dos metadados podem, também, ser utilizados diversas vezes, delineando todo o conjunto de termos empregados para representar o conteúdo do documento. O conjunto pode ser oriundo do processo de Análise Documentária, que se baseia no conceito de *aboutness* e sofrem o impacto dos mesmos fatores negativos que influenciam a indexação manual. De maneira generalizada o registro bibliográfico em ambiente convencional tem a sua concepção refletida nos metadados em ambientes digitais, sendo algumas vezes chamados de metadados bibliográficos.

A internet cada vez mais se torna um repositório universal do conhecimento e cultura humana, permitindo o compartilhamento de ideias e informações em uma escala nunca vista anteriormente (BAEZA-YATES; RIBEIRO-NETO, 1999). A necessidade de recuperar informação em bibliotecas e centros de documentação é igualmente necessária na internet, mas tal como o registro bibliográfico que apoia a atividade em um ambiente convencional, os metadados são importantes para a recuperação da informação em bibliotecas digitais.

Contudo, não houve mudanças significativas ao longo dos anos com relação à estrutura do metadado e à descrição do conteúdo do documento. O conteúdo continua sendo descrito por meio do processo de análise dos documentos onde é determinado e representado sobre o que eles tratam. Essa descrição é posteriormente representada por uma sequência simples de palavras-chave incluídos nos metadados como nos campos de um registro bibliográfico. Além disso, descrições são independentes: não há um relacionamento entre as palavras-chave ali descritas.

Os artigos científicos eletrônicos são para serem lidos, discutidos, debatidos e validados por pessoas. Porém, são meras cópias digitais de sua versão em papel, eles não incorporam as potencialidades previstas na web semântica (MARCONDES, 2005), uma extensão da web atual, onde a informação possui um significado bem definido permitindo melhor interação entre os computadores e as pessoas (BERNERS-LEE, 2001); em outras palavras, os metadados não incorporam descrições que permitam um processamento mais inteligente dos documentos pelos sistemas de recuperação de informação.

Segundo Passin (2004) a literatura sobre web semântica apresenta diversas e diferentes visões, entre elas, a de que deve permitir que os dados disponíveis e "ligados" na Internet possam ser usados por máquinas não só para propósitos de exibição, mas para automatização, integração e uso por várias aplicações. E ainda, ela deve ter como função, a criação de uma web que possibilite o uso de agentes inteligentes para que se possa recuperar e manipular a informação de maneira automática e precisa utilizando o contexto no qual ela está inserida ao invés unicamente de palavras-chave.

O princípio básico da *web* semântica é tornar possível expressar e representar a semântica dos dados através de outros esquemas descritivos, objetivando a precisão dos significados, ou seja: um esquema que descreva quais os conceitos existentes em certo domínio e como eles se relacionam. Como um modelo conceitual, que serve como uma representação abstrata dos elementos de informação. Esse esquema é representado numa ontologia. Observa-se por meio dela como os metadados representam explicitamente a semântica das informações. Portanto, as ontologias têm um papel crucial para web semântica, no sentido de que permitem o acesso, a interoperabilidade e a comunicação baseados em conteúdo.

A ontologia é similar a um dicionário, taxonomia ou glossário, porém com estrutura e formalismo que possibilita computadores processar seu conteúdo. Ela consiste de um conjunto de conceitos, axiomas e relações, e representa uma área de conhecimento. Diferente da taxonomia ou glossário permite estabelecer relações arbitrárias entre conceitos, propriedades lógicas e semânticas de relações tais como: transitividade, simetria e inferência lógica sobre as relações.

Do ponto de vista da informação a ser distribuída a respeito de um documento, metadados podem ser utilizados de maneira melhorar descrições do

conteúdo do mesmo (PASQUINELLI, 1997), (BERNERS-LEE, 1997) sem a limitação da descrição como um produto da catalogação. Assim, neste ponto, cabe uma questão: poderiam os campos dos metadados, responsáveis pela representação do conteúdo, incorporar relações, permitindo ligações entre eles e atribuindo um maior sentido à descrição?

A partir do questionamento toma-se como objetivo propor um processo de submissão de artigos para periódicos eletrônicos que implemente uma metodologia combinando o processamento linguístico do texto e as respostas obtidas através da interação com o usuário (autor) e que torne possível identificar, formalizar, representar e registrar as afirmações científicas como relações entre as partes constituintes da referida asserção científica.

2 REFERENCIAL TEÓRICO

Na busca pela resposta para a questão apresentada na introdução Marcondes (2005) propõe uma ontologia (OCCA) como um modelo para representar os elementos semânticos que constituem o conhecimento contido em um artigo científico, baseado em elementos do Método Científico (método hipotético-dedutivo) e na forma como eles aparecem em artigos científicos. O objetivo desse modelo é servir de base para publicações semânticas, uma nova forma de publicar artigos científicos. A publicação semântica fornece uma maneira dos computadores compreenderem a estrutura e até mesmo o significado das informações publicadas, tornando a pesquisa de informações e a integração de dados mais eficiente. Desta forma, os metadados que descrevem o conteúdo do artigo são representados como uma instância de OCCA, de modo a tornar seu conteúdo processável por programas.

Por meio de OCCA é possível representar quatro diferentes tipos de artigos (teóricos, experimentais-exploratórios, experimentais-indutivo e experimentais-dedutivos), cada um baseia-se em diferentes raciocínios, estratégias de argumentação e pressupostos observados em diferentes tipos de artigo científico (MARCONDES *et al*., 2009).

A proposta de OCCA se baseia na concepção de que o conhecimento científico consiste em propor e provar a existência de relações entre fenômenos e

relações entre fenômenos e a suas características (BUNGE, 1998) e (MILLER, 1947), até então desconhecidas. OCCA é constituído de elementos semânticos como a questão de pesquisa, a hipótese e a conclusão. Enquanto a questão de pesquisa se caracteriza como uma pergunta de caráter geral e não como relação, as hipóteses, ao propor uma relação entre fenômenos, têm importância decisiva quanto à manifestação do conhecimento novo em ciência.

Um exemplo disso é o seguinte enunciado: "*The aim of this study was to assess the cytototoxicity and the potential antiviral activity of violacein against the viruses: Herpes Simplex Virus type 1 (HSV-1) strains KOS, 29-R/acyclovir resistant and VR733/ATCC; Poliovirus type 2 (PV-2); Simian rotavirus SA11 strain, Hepatite A virus (HAV) strains HAF203 and HM175 and Adenovirus type 5 (AdV-5), a respiratory strain*"[i], a hipótese é operacionalizada através de um objetivo, em que o autor explicita e enfatiza o fenômeno ou fenômenos, sobre os quais o artigo vai tratar, no exemplo: *violacein* e vários tipos de vírus. Assim, a seguinte relação poderia ser descrita: *violacein* previne vários tipos de vírus, onde **previne** tem o sentido de parar, impedir ou eliminar uma ação ou condição (como uma atividade antiviral).

Uma RELAÇÃO, para OCCA, tem a forma de um **antecedente**, uma **relação** e um **consequente**, como uma 3-tupla (antecedente, relação, consequente). A relação pode ser um tipo de relação específica como "causa", "afeta", "indica", ou um tipo de relação ***tem_característica*** que define uma característica do antecedente, enquanto o antecedente e o consequente referem-se aos fenômenos estudados pelo autor. Por convenção utiliza-se a palavra relação em maiúscula para referenciar a 3-tupla e em minúscula quando se referir à relação específica contida na 3-tupla.

A importância de um esquema como de OCCA que utilize relações está no fato de que estas permitem que programas façam "inferências" sobre o conhecimento assim representado, como nos seguintes exemplos: *Violacein* (antecedente) previne (relação) que outras infecções (consequente?)? Que outras substâncias (antecedente?) previnem (relação) herpes (consequente?) além *violacein*?

Outro aspecto apontado por OCCA é que a RELAÇÃO pode, também, ser mapeada para ontologias públicas, estabelecendo uma ligação entre o artigo e o conhecimento público disponível e validado. A falta de mapeamento entre os elementos da RELAÇÃO e a ontologia pode indicar evidências de novas

descobertas por causa da falta de representação de tais elementos na ontologia (MALHEIROS, 2010).

Esse trabalho se insere em um grande contexto de pesquisa do grupo de pesquisa Informação, Conhecimento e Tecnologia da Informação (cadastrado no Diretório de Grupos de Pesquisa/CNPq). Porém, diferente daquilo que foi antevisto como um dos objetivos da pesquisa que era a construção de um editor de textos científicos, que permitisse a um pesquisador publicar eletronicamente seu artigo, tanto como texto como em formato legível por programas (MARCONDES, 2005), optou-se por um processo de submissão considerando-se que o seu desenvolvimento seria mais viável do que a construção de um editor de texto científico, pois, seria difícil convencer um pesquisador a abandonar o seu editor de texto preferido e adotar um novo.

Assim, o objetivo desse trabalho é apresentar uma proposta de um processo de submissão de artigos científicos a sistemas de gestão de periódicos eletrônicos no qual artigos científicos sejam publicados simultaneamente na forma convencional, como texto, e em um formato “inteligível” por programas. Nesse processo de submissão os próprios autores, no momento de submeterem seus trabalhos, fornecem informações sobre os resultados da pesquisa e que vão além dos dados bibliográficos. De posse das respostas do autor e por meio do processamento automático do texto, o processo extrai a RELAÇÃO que constituem a contribuição do artigo. Busca-se, assim, estender o conjunto de metadados bibliográficos incluindo as afirmações científicas sob a forma de relações estruturadas.

3 A PROPOSTA DE UM NOVO PROCESSO DE SUBMISSÃO

3.1 Procedimentos Metodológicos

Como o processo de submissão tem por objetivo que os próprios autores, no momento de submeterem seus trabalhos, forneçam informações sobre os resultados de sua pesquisa, assume-se que o autor é o mais capaz de identificar os pontos importantes do seu trabalho e a sua principal contribuição.

Para apoiar o autor na tarefa de expressar por meio de uma RELAÇÃO sua principal contribuição, optou-se por direcionar o foco de atenção para a conclusão do artigo tornando-a o elemento chave para o processo de submissão. Enquanto as hipóteses são declarações sobre características de objetos, relações, associações e, até mesmo, declarações negativas (hipótese nula), as conclusões são as que expressam as afirmações resultantes do trabalho; além disso, elas comumente fazem parte dos resumos estruturados e os autores, possivelmente, estão mais acostumados a informar do que as hipóteses. Assim, assume-se que a principal afirmação científica contida no artigo científico se manifesta através da conclusão do autor.

Ao propor aqui um processo de submissão, implementado através de um protótipo de um sistema web, é necessário criar condições para que o autor através de um diálogo (com o sistema) consiga informar de maneira adequada a RELAÇÃO que melhor representa sua principal contribuição, isto é, a conclusão do trabalho sob a forma de uma RELAÇÃO. Porém, espera-se um nível de dificuldade elevado na construção dessa interface com o usuário, isso é explicado pelo fato de se exigir do autor o domínio de vários conhecimentos que impactam diretamente na usabilidade do sistema.

Expressar a RELAÇÃO possui alta exigência cognitiva do autor, pois vários fatores dependem de sua compreensão, como por exemplo: O que é o antecedente, consequente e a própria relação? O que é uma RELAÇÃO? Como mapear essa RELAÇÃO com uma ontologia pública? Todos esses aspectos devem ser compreendidos, identificados e indicados em um momento crítico que é o envio da versão final do artigo para um periódico.

Na tentativa de aliviar a pressão desses fatores sobre os autores, toma-se como princípio, que o sistema de submissão contará com uma tarefa de extração da RELAÇÃO de forma automática. Sua função é propor (sugerir) a RELAÇÃO ao autor, que poderá aceitá-la ou não. Acredita-se que a apresentação de uma RELAÇÃO ao autor, mesmo que imprecisa, facilite a condução do diálogo, criando meios de diminuir as exigências cognitivas e facilitando a compreensão daquilo que o sistema necessita para registrar.

A tarefa de extração da RELAÇÃO é apoiada em técnicas de processamento de linguagem natural (PLN) como as implementadas pelo programa MMTx -

http://mmtx.nlm.nih.gov/, programa esse que identifica os termos controlados do *Unified Medical Language System* (UMLS). A escolha do UMLS como suporte ao trabalho está apoiada no fato dele possuir uma infraestrutura pronta e disponível livremente para o processamento de texto, além disso, conhecimento é organizado como uma rede semântica, em que termos estão estruturados em grandes categorias como organismos, estruturas anatômicas, funções biológicas e químicas, eventos, objetos físicos e descritos em um metatesauro.

Outro aspecto que apoia essa escolha é que o UMLS possui um conjunto de relações, que coincide com a proposta de representação do conhecimento de OCCA, sendo, portanto, de especial interesse para esta pesquisa.

O foco em artigos científicos da área biomédica é motivado pelo fato de que os mesmos seguem um rígido padrão formal em seus textos. Esse padrão é comumente chamado de *Introduction, Method, Results Abstract and Discussion* (IMRAD), recomendado pelo *The International Committee of Medical Journals Editors* para artigos científicos em periódicos biomédicos, isso torna mais fácil o processamento automático do texto por contar com seções bem definidas e vocabulários médicos consagrados. Por outro lado, o processamento do texto demanda tempo e seu tamanho tem influência direta nesse aspecto, sendo comum a limitação do processamento a partes específicas. Várias propostas de mineração de textos, na área biomédica, atuam exclusivamente sobre os *abstracts*, obtidos no MEDLINE/PubMed, tornando-os as mais importantes fontes de informação para mineração de texto em bioinformática (DING *et al*., 2005).

Ora, a utilização apenas do *abstract* do artigo científico pode trazer diminuição do tempo de processamento, porém, a informação obtida pode ser restrita enquanto que a utilização de todo o artigo contará com uma informação mais rica, mas o tempo de processamento aumenta consideravelmente e a quantidade de informação obtida pode se tornar intratável. O princípio aqui estabelecido busca um “equilíbrio de forças”. O processo deverá utilizar o texto completo, porém minimizando as operações sobre ele e limitando-o a determinadas seções do artigo, quando for possível.

Assim, a tarefa de extração da RELAÇÃO é desenvolvida como parte da estratégia do processo de submissão de artigos. A RELAÇÃO é obtida automaticamente, por meio do processamento do texto e servirá para iniciar o

diálogo com o autor para que ele possa interagir com o sistema e representar a conclusão do trabalho como uma RELAÇÃO.

O processamento do texto, para a extração da RELAÇÃO, apoia-se em uma abordagem híbrida que utiliza tanto técnicas oriundas do PLN e técnicas estatísticas. As técnicas do PLN utilizadas se desenvolvem de acordo com a perspectiva da teórica da gramática gerativa de Chomsky (1957) e as estatísticas sob a perspectiva da representação do texto através de seus termos, que são valorados através de sua frequência e peso, mais especificamente sob a inspiração do trabalho de Edmundson (1969) que propõe modificações no método de Luhn (1958), ao estabelecer que as palavras, além de sua frequência, passassem a ter pesos de acordo com o local e características dentro do texto.

3.2 O Processo de submissão

A seguir são descritas cada atividade do processo de submissão.

Atividade 1: envio do artigo científico (*upload* do artigo)

A tarefa de envio do arquivo assemelha-se a outras já conhecidas em sistema computacionais, como anexar um arquivo em uma mensagem eletrônica.

Atividade 2: obtenção de dados gerais do artigo

Essa tarefa tem por objetivo obter informações do autor, com relação ao contexto do trabalho, a importância do estudo e como o artigo pode ser classificado (experimental, teórico, etc.).

Atividade 3: extração do objetivo

A extração do objetivo baseia-se na identificação de frases indicativas (por ex. *The aim of our study is*...) em pontos específicos do texto, por causa de sua alta concentração nessas partes (*abstract e* introdução) (SWALES, 1990) (NWOGU, 1997). A partir de uma análise realizada e a identificação de vários tipos de frases indicativas (COSTA, 2008) foi possível elaborar um modelo de regras para a extração automática. O objetivo é posteriormente utilizado na extração da RELAÇÃO.

Atividade 4: validação do objetivo

O conjunto de frases selecionadas como indicativas do objetivo é apresentado ao autor para que ele tome a decisão e informe qual representa melhor o objetivo do

trabalho, já que o texto pode conter várias frases indicando o objetivo (COSTA, 2008).

Atividade 5: descrição da conclusão

Cabe ao autor descrever a principal conclusão do trabalho. O texto da conclusão será processado com o intuito de sintetizá-lo por meio de uma RELAÇÃO.

Atividade 6: extração automática da RELAÇÃO

Essa atividade tem por objetivo transformar a conclusão de sua forma textual para: **antecedente relação consequente**. Com este fim, duas hipóteses são estabelecidas: 1) As relações são expressas por típicos marcadores léxicos, restringindo-se, nessa primeira versão, aos verbos; e 2) As relações interessantes acontecerão entre as macroproposições mais relevantes.

Dois passos são empregados: a) Identificação das Macroproposições Relevantes e b) Identificação da Relação Propriamente Dita.

Subatividade 6.1: Identificação das macroproposições relevantes

Tem por finalidade limitar o espaço de pesquisa na busca da relação, restringindo a quantidade de termos a serem verificados. A identificação das macroproposições se baseia em três etapas: pré-processamento da frase da conclusão, determinação dos sintagmas da frase e o cálculo da importância do sintagma.

- **pré-processamento da frase de conclusão**

  A conclusão é submetida a um pré-processamento automático que tem por objetivo eliminar apostos, vírgulas e textos entre parênteses.

- **determinação dos sintagmas da frase**

  Após o pré-processamento é realizada a identificação dos sintagmas que formam a oração. Utiliza-se o programa MMTx, que além de fornecer os sintagmas, apresenta, também, as unidades em sua forma “condensada”, isto é, o programa retira os artigos, preposições que iniciam os sintagmas.

- **cálculo da importância do sintagma – pesagem do sintagma**

  A estratégia usada é identificar os dois principais sintagmas da frase de conclusão, usando-os como limites (inferior e superior) e estabelecendo um intervalo para a busca da relação. Os sintagmas mais importantes são aqueles que possuem os termos mais importantes. A atribuição da relevância é baseada na pontuação obtida por meio de uma função linear,

utilizando a frequência dos termos que compõe a conclusão e os pesos estabelecidos de acordo com cada “local”. Se o termo da oração da conclusão ocorre, também, no objetivo - obtido na atividade de extração do objetivo - no título e nas palavras-chave, a frequência é multiplicada por pesos atribuindo uma maior importância.

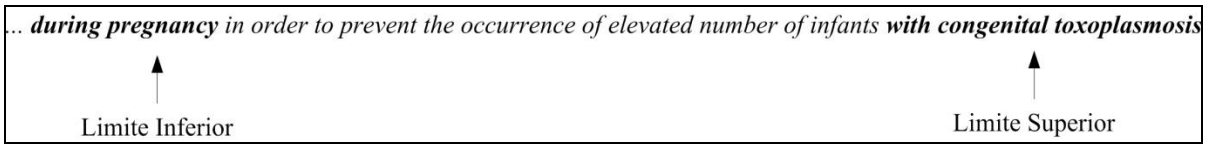

Figura 1: Identificando os Sintagmas mais Importantes.

Subatividade 6.2: identificação da relação propriamente dita

A identificação da relação se baseia na hipótese de que elas são expressas por típicos marcadores léxicos, aqui restritos, até o momento, aos verbos. Buscam-se entre os limites os sintagmas verbais para candidatos à relação.

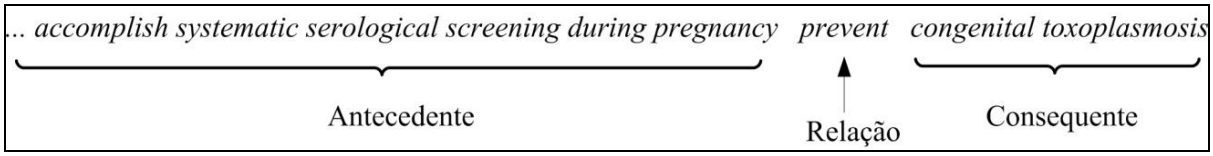

Figura 2: A RELAÇÃO Gerada.

Atividade 7: Validação da RELAÇÃO

Essa atividade tem por finalidade a confirmação, ou não, da RELAÇÃO produzida automaticamente. O autor é exposto às informações geradas (antecedente, relação, consequente) sendo possível a alteração de todos os elementos se necessário (Figura 3).

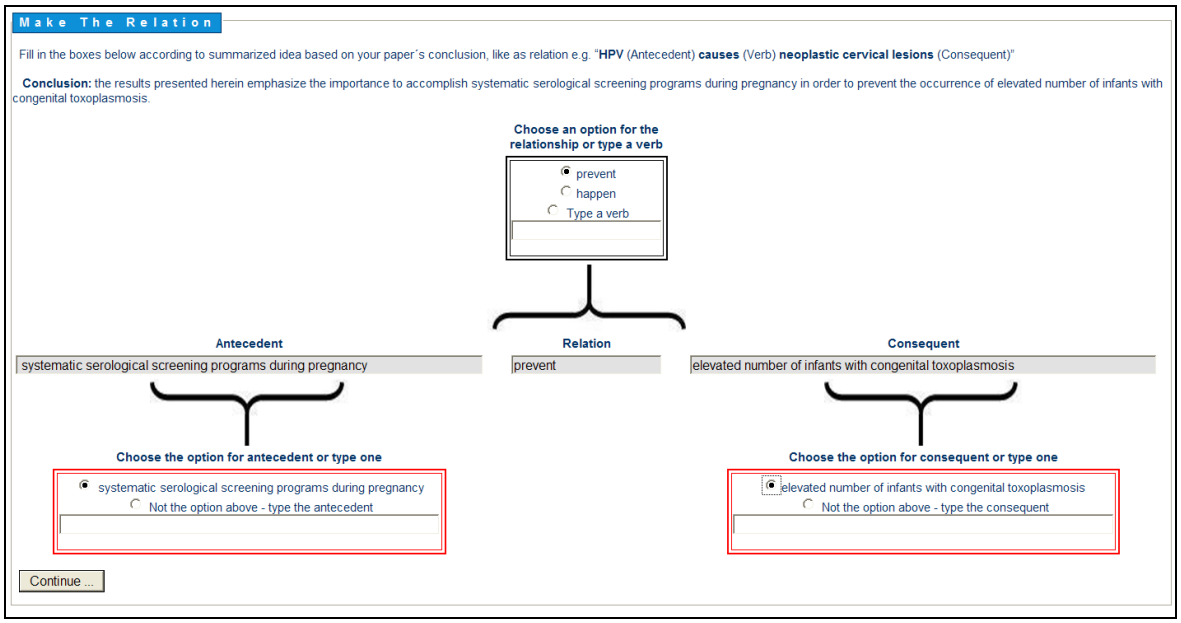

Figura 3: Validação da RELAÇÃO.

Atividade 8: Mapeamento da RELAÇÃO para a Ontologia Pública

A função do mapeamento é prover um sentido (semântico) para os elementos da RELAÇÃO por meio de ligações deles com uma ontologia pública, neste caso o UMLS. Assim, essa atividade objetiva identificar se os elementos que compõem a RELAÇÃO podem ser mapeados ou não para uma base de conhecimento público, isto é, se cada elemento já está conceitualmente descrito na ontologia. Para o mapeamento do antecedente e do consequente é utilizado o programa MMTx que relaciona termos a conceitos do metatesauro do UMLS e para tipos semânticos da rede semântica. A tabela 1 apresenta os seguintes mapeamentos para o antecedente *systematic serological screening during pregnancy*.

Tabela 1 – Candidatos para mapeamento do antecedente.

Antecedente	Cód. Conceito	Nome	Tipo Semântico
systematic serological screening during pregnancy	C0220922	Systematic	Functional Concept
	C0205473	Serologic	Functional Concept
	C0220909	Aspects of disease screening	Functional Concept
	C0220908	Screening procedure	Health Care Activity
	C1409616	Special screening finding	Finding
	C0032961	Pregnancy	Organism Function

Embora o MMTx seja um instrumento valiosíssimo, ele não possui suporte ao mapeamento de relações, no sentido desejado aqui. Para buscar esse mapeamento foi criado um dicionário que relaciona as 54 relações do UMLS a um conjunto de verbos. A construção do dicionário foi guiada pela análise manual da definição de

cada uma das relações do UMLS, utilizando-se de dois critérios: 1) Interpretação do texto da definição de acordo com o nome atribuído a relação; e 2) Observação dos verbos utilizados na definição.

Para cada verbo contido na definição foram obtidos verbos sinônimos, através de consultas ao *Wordnet* (um dicionário semântico para a língua inglesa, desenvolvido na Universidade de *Princeton*), que mantinham o mesmo sentido obtido pela interpretação da definição. Os verbos sinônimos, juntamente com os verbos contidos na definição, passam a compor o dicionário e são associados à relação.

Uma vez construído o dicionário, a tarefa de mapeamento da relação passa a ser uma simples consulta ao dicionário, que busca identificar as relações do UMLS às quais o verbo está associado.

Atividade 9: Validação dos Mapeamentos

A etapa anterior descreveu a geração de possíveis mapeamentos da RELAÇÃO para conceitos do UMLS. Contudo, na busca de tais mapeamentos vários conceitos, vários tipos semânticos (Tabela 1) e relações podem ser gerados. O autor necessita escolher, ou não, os mapeamentos mais adequados, para estabelecer o sentido desejado para a RELAÇÂO. A etapa de validação dos mapeamentos se dá por meio da interface do sistema de submissão no qual o usuário é exposto a esse conjunto de informação e opta, atribuindo o sentido desejado (Figura 4).

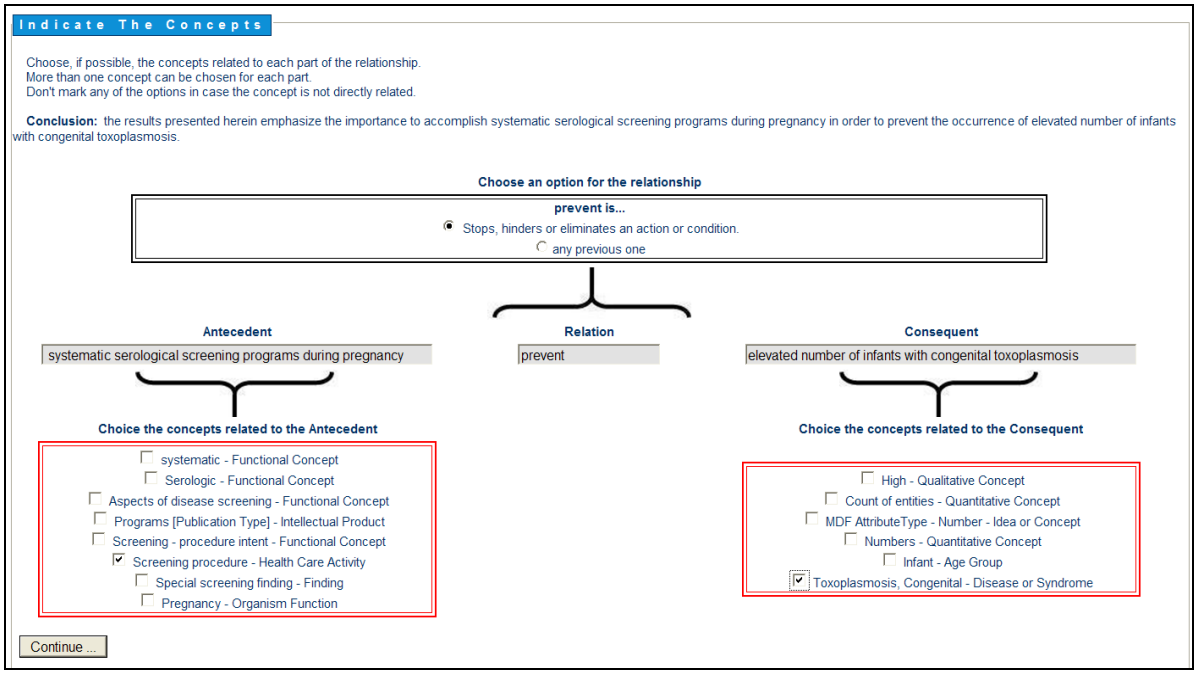

Figura 4: Validando os Mapeamentos.

Atividade 10: Geração da Instância da Ontologia OCCA

Na última atividade os elementos da RELAÇÃO são representados como instâncias de OCCA e um arquivo no formato XML é gerado (Figura 5).

```
<Relation>
    <Antecedent> systematic serological screening programs during
pregnancy</Antecedent>
    <Relation>prevent</Relation>
    <Consequent>elevated number of infants with congenital toxoplasmosis</Consequent>
    <Map_Antecedent>
      <string>Screening procedure - Health Care Activity</string>
    </Map_Antecedent>
    <Map_Relation>Stops, hinders or eliminates an action or condition.    </Map_Relation>
    <Map_Consequent>
      <string>Toxoplasmosis, Congenital - Disease or Syndrome</string>
    </Map_Consequent>
  </Relation>
```

Figura 5: Trecho da Instância de OCCA.

4 RESULTADOS FINAIS

Para testar e validar o processo de submissão foi necessário a construção de um software (protótipo) que implementasse as atividades definidas anteriormente. Assim, foram convidados seis professores pesquisadores da Universidade Federal Fluminense (Instituto Biomédico), que submeteram ao protótipo seus trabalhos já publicados em diferentes revistas internacionais em língua inglesa. Todas as interações dos autores foram gravadas e posteriormente analisadas. Ao terminar o uso do protótipo, os pesquisadores responderam o *System Usability Scale* (SUS), um questionário simples, padronizado e validado, desenvolvido pela *Digital Equipment Corporation*, e utilizado como uma ferramenta para medir o nível de usabilidade de um sistema. Composto de 10 quesitos, utilizando a escala Likert, permite obter uma visão global das avaliações subjetivas com respeito à usabilidade, por meio de uma pontuação geral.

A usabilidade de um sistema é definida pela norma ISO 9241(11) e só pode ser aferida tendo em conta o contexto de uso do sistema, isto é, o sistema só poderá ser avaliado por usuários reais e em situações reais (ambiente real), por isso foram convidados pesquisadores cujos artigos já foram publicados. Tipicamente, a usabilidade é medida por três aspectos diferentes: **eficácia** do sistema (os usuários podem alcançar com êxito os seus objetivos); **eficiência** (quanto esforço e recurso

foram gastos na consecução desses objetivos); e **satisfação do usuário** (a experiência com o sistema foi satisfatória).

De uma maneira geral, a avaliação da usabilidade foi considerada adequada, com boas perspectivas de uso do sistema pelos autores, como pode ser observada (tabela 2) pela média final alcançada (83,75).

TABELA 2 – Resultado do SUS.

Valores	Aval. 1	Aval. 2	Aval. 3	Aval. 4	Aval. 5	Aval. 6
Escore SUS	87,50	70,00	90,00	95,00	77,5	82,5
Media Final	83,75					

Além da avaliação por meio do SUS, foram observados alguns aspectos ligados à exatidão e à integralidade com que o usuário consegue alcançar os objetivos iniciais. Um aspecto bastante positivo está relacionado com o fato de que todos os avaliadores completaram as tarefas dentro do tempo esperado (Tabela 3).

TABELA 3 – Aspectos Analisados.

Aspectos Analisados	Aval. 1	Aval. 2	Aval. 3	Aval. 4	Aval. 5	Aval. 6
Nº de vezes que o usuário relatou desorientação	0	0	1	0	0	0
Nº de vezes que o usuário solicitou informações	3	1	2	1	1	0
Nº de vezes que o usuário relatou discordância com relação a algum aspecto do *software*	0	2	1	1	1	1
Nº de tarefas não completadas	0	0	0	0	0	0

Com relação ao número de solicitações de informações (8), a maioria dos usuários pediu informações, mesmo sendo dito na fase de instruções que não haveria suporte (uso de manuais, consultas pessoais, entre outros). Porém, ao analisar o protocolo de voz, percebe-se que uma parte dos pedidos (5, 62,5%) foi em função de informações já prestadas (na fase de instruções), como, por exemplo: “É para escrever em inglês?”, “Posso começar?”. Esses pedidos foram interpretados como ansiedade ou nervosismo por parte dos avaliadores, por estarem sendo monitorados e gravados. Essa interpretação pode ser confrontada com os quesitos 4 (“*I think that I would need the support of a technical person to be able to use this system*”) e 10 (“*I needed to learn a lot of things before I could get going with this system*”) do questionário SUS que indaga, respectivamente, sobre a necessidade de

aprendizagem antes de usar o sistema e de suporte técnico; apenas um avaliador indica a necessidade de treinamento para uso do sistema e nenhum avaliador indica a necessidade de suporte técnico.

Outros dois pedidos de informação (25%) estão em função do fato de que, ao escrever a conclusão, existe uma limitação do número de palavras (máximo 49); dois avaliadores questionaram se o programa indicaria o limite. Embora o protótipo exiba uma mensagem de erro, indicando que o número de palavras permitido fora extrapolado, pode-se pensar em alterar a interface, com o objetivo de indicar dinamicamente a quantidade de palavras digitadas, provendo dessa forma um melhor *feedback* para o usuário.

Em apenas um momento o pedido de informação foi considerado indicativo de problema. Ao executar a tarefa “Validação da RELAÇÃO” no protótipo percebeu-se a desorientação do terceiro avaliador com respeito às informações apresentadas, sendo ele o único avaliador a indicar a necessidade de aprendizagem no questionário SUS. Ao apresentar o conjunto de candidatos à relação, a interface exibe esse conjunto na porção superior da tela, enquanto os candidatos para o antecedente e consequente são apresentados na parte inferior. Acredita-se que essa organização espacial tenha levado o avaliador a interpretar que o antecedente e consequente estavam subordinados à relação.

Com relação ao software, os avaliadores foram quase unânimes em um aspecto: os conceitos apresentados para um possível mapeamento (Atividade 9) eram muito gerais e não representavam adequadamente o antecedente e o consequente da RELAÇÃO. Isso ocorreu em função do instrumento utilizado (UMLS) para o mapeamento. Embora o UMLS (na versão 2010) abarque 2.201.360 conceitos e 9.976.458 nomes de conceitos (termos), os artigos apresentados pelos avaliadores e utilizados para a avaliação do protótipo abordavam assuntos bem específicos, o que dificultou a adoção de conceitos mais apropriados para a representação.

4.1 Limitações

A primeira limitação imposta desde o início do trabalho foi a de que apenas a principal conclusão do artigo científico seria tratada. Essa limitação objetivava

manter a atenção voltada para o problema de representar a conclusão como uma RELAÇÃO; ao considerar a possibilidade de tratar várias conclusões, outros problemas seriam somados, o que poderia tornar a discussão da obtenção da relação ofuscada. Tal limitação deve ser eliminada com a continuação do estudo, visto que se trata da primeira tentativa de implementação do modelo OCCA. A segunda limitação foi o número de autores disponíveis para efetuar os testes. A dificuldade de acesso a autores que se dispusessem a testar o protótipo foi bastante grande. Tentativas de contato, através de editores de periódicos nacionais, não foram frutíferas, até o momento, o que acabou limitando o número de autores-avaliadores para teste. Os artigos utilizados foram da área de biomedicina, todos na língua inglesa.

5 CONSIDERAÇÕES FINAIS

O processo proposto estabelece uma exigência fundamental para o autor, a de que ele crie uma RELAÇÃO para representar sua principal conclusão, de maneira que ela possa contribuir, não só para a recuperação do trabalho, como também para ser utilizada nos relacionamentos com outros trabalhos, através de associações automáticas, aumentado assim às possibilidades de recuperação de informação. Tal exigência obriga o autor a ser conciso, reduzindo o seu principal resultado a umas poucas palavras, a uma RELAÇÃO.

Convidar o autor a ser conciso não é uma ideia nova, diversos mecanismos foram propostos com este fim para auxiliar na recuperação do artigo científico, tais como, as palavras-chave, as terminologias padronizadas, os resumos estruturados, entre outros. A novidade é que, diferente das palavras-chave, que são normalmente “soltas”, a construção de sentido é realizada através de uma RELAÇÃO. Essa forma de reescrever seu principal resultado pode ser comparada com o texto para a divulgação científica, em que o autor deve reformular seu discurso, a fim de torná-lo inteligível a uma sociedade constituída de leigos. No caso da RELAÇÃO, ela é a forma proposta para que ele reformule seu discurso de maneira que se torne inteligível a um programa de computador.

A metodologia proposta na forma de um processo de submissão combinando o processamento linguístico do texto e as respostas obtidas através da interação

com o usuário (autor) teve como base, técnicas e práticas encontradas na Ciência da Informação. Como exemplo, cita-se, o uso de frequência de termos, palavras chaves, muito utilizadas em indexação automática e estudos bibliométricos.

Uma possibilidade que se vislumbra é a de “embutir” o processo aqui definido no Sistema Eletrônico de Editoração de Revistas, software desenvolvido para a construção e gestão de uma publicação periódica eletrônica, com a finalidade de tornar o processo viável e aceito para os editores. Estamos em um momento de transição, das publicações eletrônicas (ainda concebidas segundo o modelo impresso) para as publicações eletrônicas semânticas, as que utilizam, como a nossa proposta, as tecnologias da web semântica. Este parece ser cada vez mais o caminho para endereçar a questão fundadora da Ciência da Informação, a crescente “explosão informacional”.

A discussão dos resultados, apresentados anteriormente, permite afirmar que o objetivo do trabalho foi alcançado.

REFERÊNCIAS

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval***.* Boston: Addison Wesley, 1999. 544p.

BERNERS-LEE. **W3C Data Formats**. The World Wide Web Consortium Note. Disponível em: <http://www.w3.org/TR/NOTE-rdfarch>. Acesso em: 23 out. 2008.

BUNGE, M. **Philosophy of science**: From Problem to Theory. New Brunswick, London: Transaction Publishers, 1998. 2v.; 607p.

CHOMSKY, N**. Syntactic structures**. The Hague: Mouton,1957. 117p.

COSTA, L. C.; MARCONDES, C. H. Padrões linguísticos para identificar elementos de ontologia. In: SEMINÁRIO DE PESQUISA EM ONTOLOGIA NO BRASIL, Niterói, 2008. (Poster)

DING, J.; VISWANATHAN, K.; BERLEANT, D.; HUGHES, L.; WURTELE, E. S.; ASHLOCK, D.; DICKERSON J. A.; FULMER, A.; SCHNABLE, P. S. Using the Biological Taxonomy to Access Biological Literature with PathBinderH. **Bioinformatics**, v.21, n.10, p.2560-2562, maio 2005.

EDMUNDSON, H. P. New methods in automatic extracting. **Journal of the ACM***,* v.16, p.264-285, 1969.

GILLILAND-SWETLAND, A. J. Setting the Stage: Defining Metadata. In: MURTHA, B. (Ed.). **Introduction to metadata**: Pathways to digital information. Los Angeles: Getty Information Institute, 2000. 48p.; p.12

LUHN, H. P. The automatic creation of literature abstracts. **IBM Journal of Research and Development**, v.2, n.2, p.159-165, 1958.

MALHEIROS, L. R. A identificação de traços de descobertas científicas pela comparação do conteúdo de artigos em Ciências Biomédicas com uma ontologia pública. Niterói: UFF, 2010.

Tese (Doutorado) - Programa de Pós-Graduação em Ciência da Informação / UFF/IBICT.

MARCONDES, C. H. Da comunicação científica ao conhecimento público: artigos científicos digitais como bases de conhecimento. In: ENCONTRO DA ASSOCIAÇÃO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO, 6., 2005, Florianópolis, **Anais...** Florianopolis, 2005. (CD-ROM)

_____ *et al*. A publishing system to extract and represent the knowledge content of scientific articles on health science in machine-processable format. In: INTERNATIONAL CONFERENCE ON ELECTRONIC PUBLISHING, 13., 2009, Milão. **Proceedings...** Milão: Edizione Nuova Cultura, 2009. (CD-ROM). Disponível em: <elpub.scix.net>. Acesso em: 1 ago. 2009.

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

NWOGU, K. N. The medical research paper: Structure are functions. **English for Specific Purposes**, v.16, n.2, p.28-32, 1997.

PASSIN, T. B. **Explorer's guide to the semantic web**. United States of America: Manning Publications, 2000. 300p.

PASQUINELLI, A. Information technology directions in libraries: a sun microsystems white paper. Disponível em: <http://www.infoperpus.8m.com/artikel/00005.htm>. Acesso em: 20 mar. 2010.

SARACEVIC, T. Interdisciplinary nature of information science. **Ciência da Informação**, Brasília, v.24, n.1, p.35-40, jan./abr. 1995. Disponível em:<http://www.ibict.br/cionline>. Acesso em: 22 ago. 2006.

NOTAS

[i] Conclusão retidada de: ANDRIGHETTI-FROHNER, CR, ANTONIO, RV, CRECZYNSKI-PASA, TB et al. **Cytotoxicity and potential antiviral evaluation of violacein produced by Chromobacterium violaceum**. Mem. Inst. Oswaldo Cruz., v.98, n.6, p.843-848, set. 2003.Disponível em:<http://www.scielo.br/scielo.php? script =sci_arttext&pid=S0074-02762003000600023 &lng=en&nrm=iso>. Acesso em: 30 Abr. 2010.