PROTÓTIPO DE UM SISTEMA DE SUBMISSÃO DE ARTIGO CIENTÍFICO PARA PERIÓDICO ELETRÔNICO SEMÂNTICO EM CIÊNCIAS BIOMÉDICAS

Leonardo Cruz da Costa

Resumo: Problema: publicações eletrônicas, apesar dos avanços das TIs, são ainda apoiadas no modelo impresso. O formato textual dificulta que programas possam ser usados para o processamento "semântico" do conteúdo do artigo científico. **Objetivos**: descreve a implementação de um protótipo de um sistema de submissão de artigo científico para periódico eletrônico, em ciências biomédicas, capaz de extrair o conhecimento do texto do artigo e representá-lo em formato "inteligível" por programas, tornado possível, futuramente, a recuperação semântica. **Fundamentação teórico metodológica:** o protótipo é a implementação parcial da proposta de OCCA (MARCONDES, 2005) um modelo para representar os elementos semânticos que constituem o conhecimento contido em um artigo científico, baseado em elementos do Método Científico (hipotético-dedutivo). O objetivo desse modelo é servir de base para publicações semânticas. **Resultados:** resultados dos testes com o protótipo envolvendo pesquisadores-autores são relatados.

Palavras-chave: publicações eletrônicas, representação do conhecimento, comunicação científica.

1 INTRODUÇÃO

Segundo Shera e Cleveland (1977), após a Segunda Guerra, uma enorme quantidade de informações foram produzidas devido ao aumento do número de cientistas, de pesquisas e do desenvolvimento de tecnologia, gerando expressões como, "explosão bibliográfica" ou "caos documentário" (BRADFORD, 1961). Esse fenômeno conduziu a uma grande preocupação quanto ao registro, acesso, recuperação e divulgação da informação e de documentos. Uso das tecnologias de informação, no sentido de intervir neste processo, procurando aperfeiçoar o acesso aos pesquisadores à informação relevante, não é novo. *As We May Thin*k, texto publicado em 1945 por Vannevar Bush, por exemplo, argumentava que o esforço científico deveria ser concentrado em deixar o conhecimento humano mais acessível e propunha uma máquina denominada Memex.

Porém, se, antes, para a obtenção de informações, era necessário se dirigir a uma biblioteca, atualmente tais informes podem estar armazenados em um *pen drive*. Contudo, o conhecimento continua sendo codificado em um formato textual, para ser lido por pessoas, e não em um formato legível por programas, permitindo que "inferências", "decisões" e "argumentações" automáticas pudessem ser realizadas sobre o conteúdo do texto. Vislumbra-se, então, investir na *web* semântica,

como forma de contribuir positivamente na recuperação de documento relevantes para ciência.

A *web* semântica é uma extensão da *web* atual, onde a informação possui um significado bem definido (BERNERS-LEE, 2001). Os documentos na *web* passam a ser representados através de metadados semânticos que descrevem o que significa a informação contida no documento. O significado é expresso de acordo com um conjunto de conceitos e relações de um domínio específico e expresso através de uma ontologia. A ontologia representa um sistema conceitual que normaliza o discurso em um dado domínio. Assim, as descrições sobre o documento estão, portanto, de acordo com um sistema conceitual, estabelecendo um significado claro e não ambíguo para aquilo que se descreve.

Podemos proceder a uma analogia das descrições baseadas em uma ontologia com a indexação por meio de um vocabulário controlado. Ambas apoiam a recuperação de documentos, porém as descrições com base em ontologias passam a contar com recursos semânticos que ultrapassam os limites das palavras chaves e dos operadores lógicos comumente usados nas máquinas de buscas. Passa, em consequência, a ser possível a construção de agentes de *software* aptos a "compreender" e a processar de forma "inteligente" (inferir) o conteúdo descrito pelo metadados semânticos.

Desta maneira, podemos entender que as metas da *web* semântica mostram-se mais ambiciosas do que simplesmente permitir indexar melhor a informação. O documento digital, dentro do contexto da *web* semântica, processado por agentes inteligentes pode-se constituir em nova e poderosa ferramenta cognitiva.

Esse cenário do documento científico digital - com seu conteúdo representado em formato legível por programas - trará novas perspectivas para a validação, registro, recuperação, disseminação e aprendizagem de novos conhecimentos: a publicação de artigos científicos na *web*, com essa característica, pode vir a ser uma ferramenta cognitiva onde suas potencialidades estão longe de serem avaliadas (MARCONDES, 2005).

A publicação de resultados, na *web* mesmo apoiada em uma infraestrutura como a que é projetada para a *web* semântica, torna-se um desafio. Entretanto, definir como a comunidade científica publicará os resultados em um formato apropriado, para que a sua recuperação ultrapasse os limites impostos pelo modelo impresso, configura um problema ainda sem solução.

A construção de um protótipo de um sistema de submissão, objetivo desse trabalho, encontra-se no âmbito desse contexto, permitindo criar condições para a descrição da semântica, e a futura recuperação e validação do novo conhecimento e de seu processamento por agentes de *software*, apoiando, de forma muito mais significativa, o trabalho do pesquisador. As tarefas a serem realizadas pelos autores, no sistema, são similares àquelas que eles já desempenham quando do envio de seu trabalho, para publicação, em um periódico científico, repositórios ou bibliotecas digitais.

2 REFENCIAL TEÓRICO

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. 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) 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?

Na busca pela resposta para a questão acima Marcondes et al. (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), (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.

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.

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).

3 O PROTÓTIPO DE UM PROCESSO DE SUBMISSÃO

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? O que é 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 UMLS – *Unified Medical Language System*. 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 IMRAD (*Introduction, Method, Results Abstract and Discussion*), isso torna mais fácil o processamento automático do texto por contar com seções bem definidas e vocabulários médicos consagrados.

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.

3.1 O Processo de submissão

Uma visão geral do processo de submissão é representada por meio da figura 1. A seguir são descritas cada atividade que compõe o 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.).

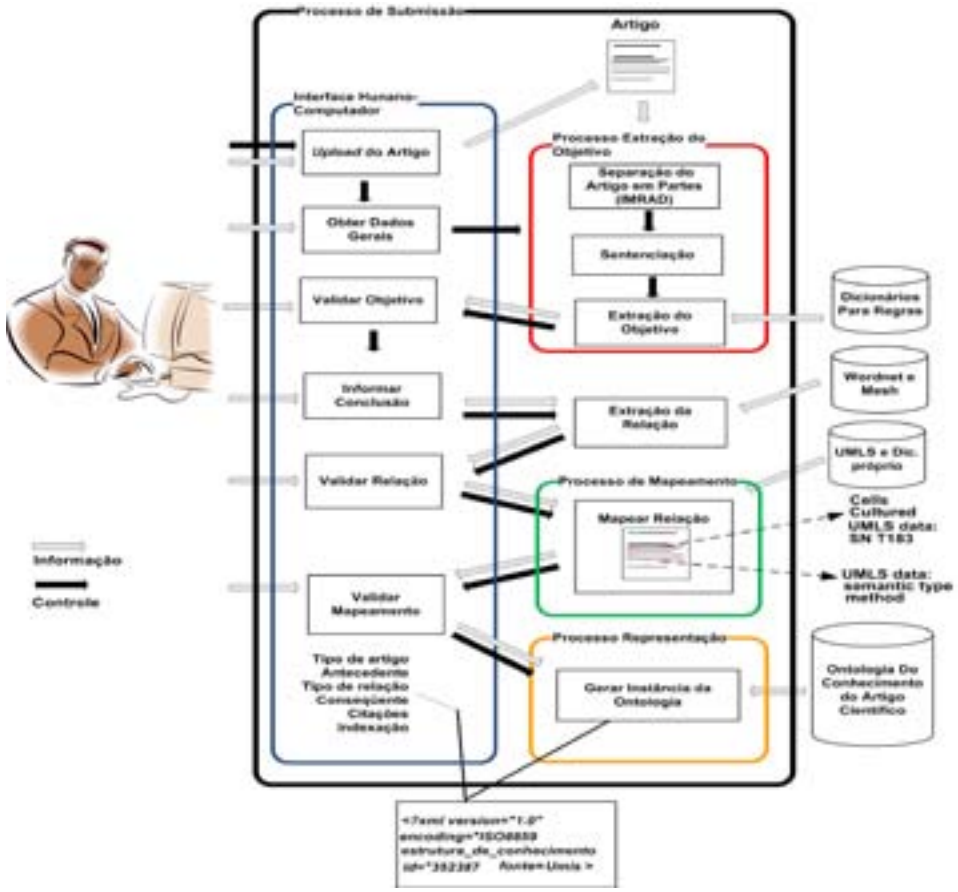

Figura 1- Visão geral processo de submissão

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). 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 do autor é posteriormente utilizado na tarefa de 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 seu 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 RELAÇÃO (antecedente, relação, consequente). Com este fim, duas hipóteses são estabelecidas: 1) As relações são expressas por típicos marcadores léxicos, restringindo-se, aos verbos, verbos substantivados (nominalização) e termos ou conceitos médicos específicos; 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) (figura 2) 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" no artigo. 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.

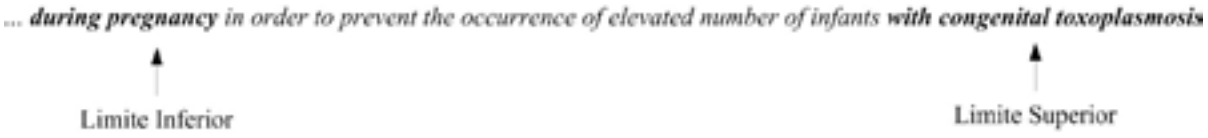

Figura 2 - Identificando os sintagmas mais importantes

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, aos verbos, aos verbos substantivados (nominalização) e aos termos ou conceitos médicos específicos. No exemplo anterior, entre *congenital toxoplasmosis* e *pregnancy* (os limites) ocorrem dois sintagmas, um nominal (*the occurrence*) e outro verbal (*prevent*) tomado imediatamente como candidato à relação (figura 3).

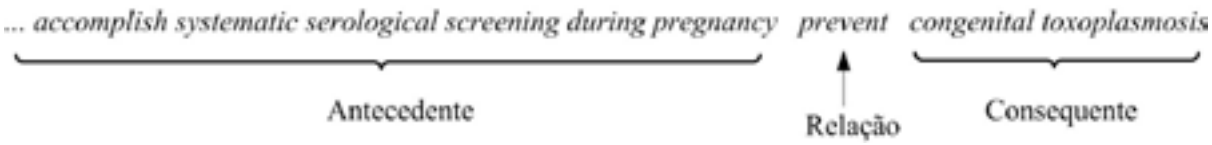

Figura 3 - Um candidato a RELAÇÃO

O nominal (*the occurrence*) é submetido a um processo de verificação de nominalização. Esta verificação toma como base uma consulta ao *Wordnet*, obtendo as possíveis nominalização do termo. Caso o substantivo esteja relacionado diretamente a um verbo, esse passa a ser candidato à relação. Para identificar os conceitos e termos médicos utiliza-se o programa MMTx. Na frase a seguir o conceito *antibiotics* é definido no vocabulário MeSH como "s*ubstances produced by microorganisms that can inhibit or suppress the growth of other microorganism"*.A ideia básica é utilizar os verbos (*produce, inhibit, suppress*), contidos na definição do conceito como marcadores (candidato) de uma possível relação (figura 4).

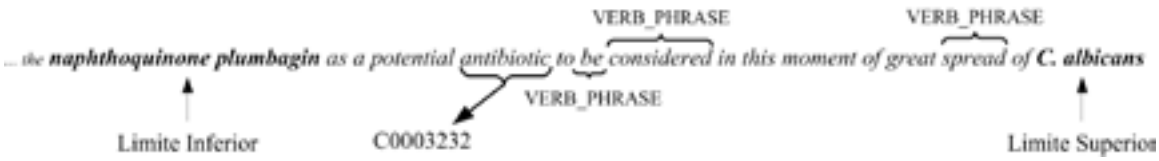

Figura 4 - Identificando um conceito médico associado aos verbos *produce, inhibit, suppress*

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 5).

Figura 5 - 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	C0205473	Serologic	Functional Concept
serological	C0220909	Aspects of disease screening	Functional Concept
screening	C0220908	Screening procedure	Health Care Activity
during	C1409616	Special screening finding	Finding
pregnancy	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

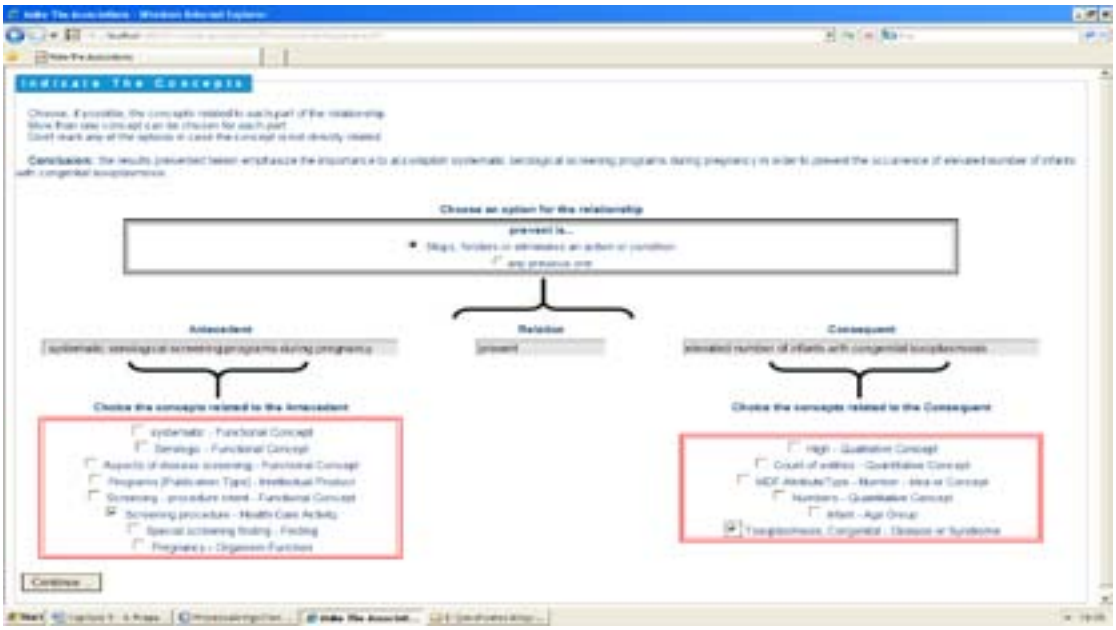

Figura 6 - Validando os mapeamentos para a UMLS

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 (como apresentado na 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 6).

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 7).

```
<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 7 - Trecho da instância de OCCA

4 AVALIAÇÃO E RESULTADOS OBTIDOS

A avaliação teve como objetivo testar se o protótipo que implementa o processo de submissão proposto é apropriado para a tarefa de representar a principal conclusão do artigo, em forma de RELAÇÃO. Trata-se de um teste de usabilidade do protótipo do sistema de submissão, isto é, testar em condições reais de funcionamento e com usuários reais (autores) o software.

O material a ser utilizado é o artigo científico fornecido pelo próprio autor. Para realizar o teste, foi definido um protocolo com o objetivo de normalizar as informações fornecidas para as pessoas envolvidas no teste. O protocolo estabeleceu com clareza os procedimentos a serem executados durante o teste, especificando-os e padronizando-os, bem como a definição das variáveis quantitativas e as tabulações obtidas e sua posterior organização e compilação para análise. Para a avaliação da satisfação do usuário (avaliação qualitativa) foi aplicado um questionário de pós-teste (*System Usability Scale*).

Para testar o software (protótipo) 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.

Os testes realizados com os usuários revelaram algumas informações importantes. O tempo esperado, para realizar todas as tarefas, foi de 15 minutos. O maior tempo consumido foi de 17 min. 17 s. (avaliador 3), enquanto o menor tempo 9 min. 24 s. (avaliador 4), sendo tempo médio obtido de 12min 23s dentro das margens projetadas (figura 8).

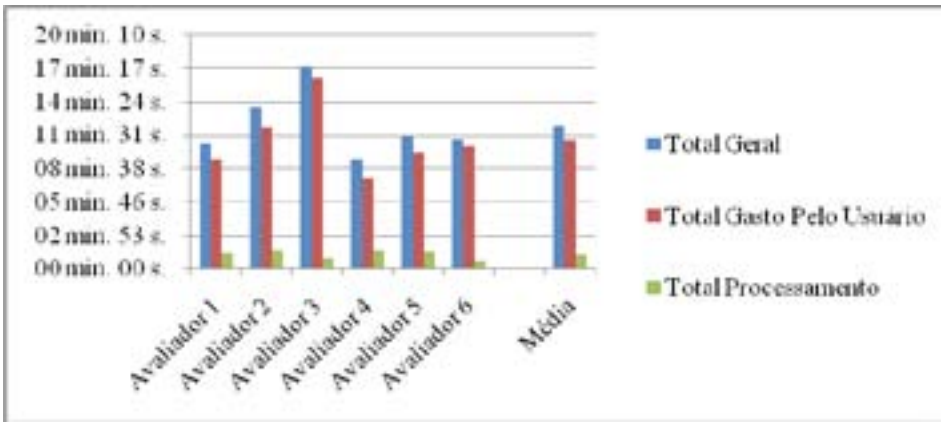

Figura 8 - Tempos gasto em cada tarefa pelos avaliadores e pelo protótipo (automática).

Figura 9 - Percentual sobre média de tempo despendido em cada tarefa.

Porém, ao analisar as médias de tempos despendidos (figura 9), em cada tarefa individualmente, identifica-se a existência de dois "gargalos", na obtenção dos dados gerais (Obter Dados Gerais) e na definição da conclusão (Informar a Conclusão). Em média, as duas tarefas consumiram 55% do tempo gasto. Na primeira tarefa, é exigido do autor o fornecimento de informações, com relação ao contexto do estudo e a sua importância, e, na segunda, a descrição da principal conclusão. O tempo gasto está em função do fato de que o autor é convidado a argumentar, no momento da submissão, o porquê da importância de sua pesquisa e qual a sua principal conclusão. Tais informações podem

estar "diluídas" no texto do artigo, exigindo atenção e precisão na descrição; isso faz com que as tarefas despendam um tempo elevado para sua execução. Ora, tais tarefas poderiam ser eliminadas do processo, adotando como pré-requisito a utilização do resumo estruturado, isto é, o texto do artigo científico contaria com o abstract no formato de resumo estruturado, onde rótulos indicam claramente o contexto, a principal conclusão do trabalho, entre outros.

Com relação aos tempos obtidos nas tarefas de extração automática, tanto para o objetivo quanto para relação, se observa um elevado tempo para o primeiro (9% em média do tempo total gasto) se comparado ao segundo (1% em média do tempo total gasto).

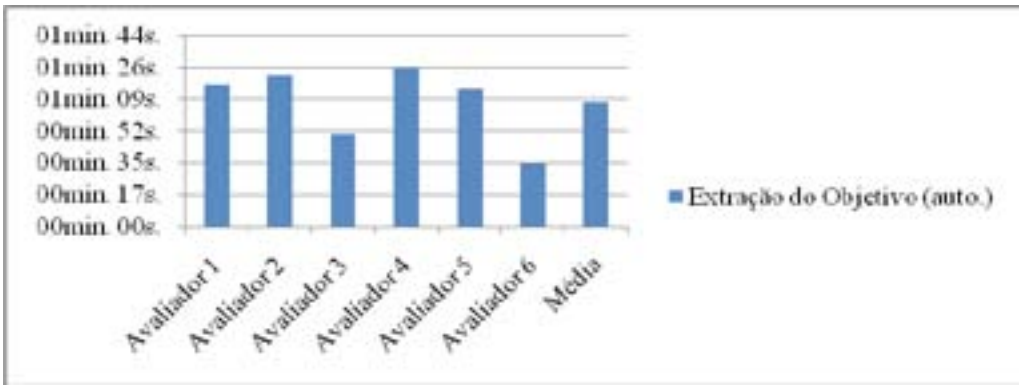

Figura 10 - Tempos obtidos pela extração automática do objetivo.

Constatou-se, também, para a extração do objetivo, uma considerável variação nos tempos (34s., no melhor caso, e 2min. 27s. no pior, com uma média de 1 min. 8s.) (figura 10). Isso pode ser explicado em função do número de sentenças, que ocorrem em cada artigo, a serem analisadas para que se extraia o objetivo.

Contudo, a adoção de um resumo estruturado, como sugerido anteriormente, conduz, além da eliminação das tarefas de Obtenção dos Dados Gerais do Artigo e a Obtenção da Conclusão, a eliminação, também, da tarefa de Extração do Objetivo, pois, como é comum no resumo estruturado (BAYLEY; ELDREDGE, 2003), a indicação do objetivo, a tarefa de extração de tal elemento torna-se desnecessária, gerando uma diminuição não só no número de etapas do processo, como também, no tempo total.

Com relação à extração da relação, o tempo médio obtido foi de 9s, representando apenas 1% do tempo despendido com o conjunto total de tarefas (figura 11).

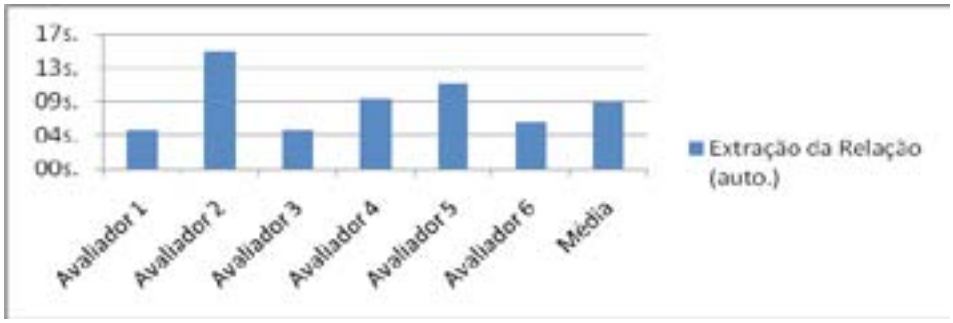

Figura 11 - Tempos obtidos para a extração automática da relação.

Um aspecto identificado, porém ainda não explicado, é a variação nos tempos obtidos. Em alguns casos (avaliação 1 e 3), o tempo despendido (5 segundos) para a extração da relação foi três vezes menor que para outro caso (15 segundos na avaliação 2), embora o número de palavras empregadas na conclusão (40 palavras no primeiro e 39 no terceiro) seja maior que a utilizada na segunda avaliação (27 palavras). Possivelmente, o tempo não está em função do número de palavras e, sim, no tipo construção sintática e o uso de determinadas palavras na oração.

As discordâncias apresentadas pelos avaliadores, com relação ao *software*, estão relacionadas a um único aspecto: os conceitos apresentados eram muito gerais e não representavam adequadamente o antecedente e o consequente da relação. Isso se deve ao instrumento utilizado – a UMLS – para o mapeamento da relação. Embora a UMLS (na versão 2010) abarque 2.201.360 conceitos e 9.976.458 nomes de conceitos (termos), os artigos escolhidos abordavam assuntos bem específicos, o que dificultou a adoção de conceitos mais apropriados para a representação.

Ao terminar o uso do protótipo, os pesquisadores responderam o *System Usability Scale* (SUS- vide em: <*hell.meiert.org/core/pdf/sus.pdf*>), um questionário simples, padronizado, validado 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. De uma maneira geral, a avaliação da usabilidade foi considerada adequada (83,75), com boas perspectivas de uso, pelos autores, de um futuro um sistema a ser construído com essa proposta.

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.

5 CONSIDERAÇÕES FINAIS

O processo submissão proposto, implementado por meio de um protótipo, 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.

O sistema de submissão, com as características propostas aqui, deverá ampliar as funcionalidades dos atuais programas de autopublicação de modo a permitir a um autor publicar seu artigo simultaneamente, em formato textual legível por pessoas, e em formato "inteligível" por programas.

A exigência de acesso, tempo e motivação criam um "gargalo" para a busca de resultados. Ao encontrar um artigo científico é necessário localizar e extrair a informação desejada e isso só pode ser realizado percorrendo todo o corpo do artigo para identificar seus resultados fundamentais que devem ser avaliados dependendo da habilidade e experiência do leitor, do tamanho do próprio artigo e se bem organizado, bem escrito e completo.

Vários mecanismos foram sendo criados ao longo dos anos como forma que diminuir essas restrições: vocabulários controlados, métodos de indexação, indexação automática, resumos estruturados.

Modificações na forma de publicar um artigo científico, como proposta, deve ser considerada nesta tese como natural, dada à contínua evolução tecnológica e da própria comunidade científica, tal ponto é salientado por Meadows (1974) ao afirmar que muitas das mudanças por que tem passado o artigo científico estiveram relacionadas com o crescente aumento e complexidade da comunidade científica - com a consequente necessidade de melhorar a eficiência de suas atividades de comunicação.

Ainda segundo o mencionado autor, o volume sempre crescente de informações expande-se na mesma velocidade, tanto para as antigas gerações de pesquisadores quanto para nós e que a sensação de "estar afogado em informações" já era lugar-comum há muitos anos. Mas, se antes

os pesquisadores restringiram sua atenção a determinados temas - criando desta forma uma maior especialização na ciência, de tal modo que a informação que precisam absorver continuava a situar-se dentro de limites aceitáveis - hoje esses limites já alcançam dimensões gigantescas.

Não podemos negar que os atuais estudos na CI devem muito às práticas documentárias desenvolvidas a partir do final do século XIX, pois foram as primeiras a contemplar o tratamento da informação (FROHMANN, 2004). Porém, apesar dos aportes das TIs para facilitar a publicação, a busca e o acesso ao texto final de documentos eletrônicos, o processo sociocultural em que se baseia e que apoia a comunicação científica continua o mesmo, dentro do paradigma da Biblioteconomia e da Ciência da Informação. Informar-se sobre os artigos científicos é tarefa fundamental no exercício da prática científica. Entretanto, mesmo com o advento dos computadores e dos instrumentos de busca o que se consegue obter, normalmente, é muito mais "informação" ou "dados" do que "conhecimento" (ZIMAN, 1979).

O documento científico digital, publicado na *web*, que hoje já é construído e estruturado dentro de um formalismo muito grande poderia, então, transformar-se, no sentido de incluir, além de suas partes textuais tradicionais o conhecimento declarado pelo autor. Isso faria com que as práticas documentárias sofressem alterações por um lado e, por outro, viabilizaria a recuperação e a disseminação não só do documento, mas, também, do conhecimento nele contido materializando, assim, a visão de Ziman (ZIMAN, 1979), do conhecimento científico como um empreendimento coletivo, em que a contribuição de um pesquisador se soma, como "mais um tijolo na construção do edifício da ciência".

ABSTRACT

Problem: Electronic publications, in spite of the progress of TIs, are still based on the printed text model. The textual format hinders programs from being used for the "semantic" processing of the contents of scientific articles. **Objective**: this work describes the implementation of a prototype of a system of submission of scientific articles for scholarly electronic publishing, in biomedical sciences capable of extracting knowledge from the text of the article and to represent it in "intelligible" format for programs. **Theoretical-methodological bases**: the prototype is the partial implementation of the proposal of OCCA (MARCONDES, 2005), a model to represent the semantic elements that form the knowledge contained in a scientific article, based on elements of the Scientific Method (hypothetical-deductive). The objective of that model is to serve a basis for semantic publications. **Results**: results are related to testing the prototype involving researcher-authors.

Keywords: electronic publishing, knowledge representation, scientific communication, *Semantic publishing*.

5 REFERÊNCIAS

BAYLEY, L.; ELDREDGE, J. The structured abstract: an essential tool for researchers. Hypothesis, v.17, n.1, p.11-13, **Spring**. 2003.

BERNERS-LEE. W3C Data Formats. **The World Wide Web Consortium Note**. Disponível em:<http://www.w3.org/TR/NOTE-rdfarch>. Acesso em: 23/10/2008.

BRADFORD, S. C. **Documentação.** Rio de Janeiro: Fundo de Cultura, 1961.

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

BUSH, V. As we may think. **The Atlantic Monthly**, July, 1945. Disponível em: <http://www.theatlantic.com/unbound/flashbacks/computer/bushf.html> Acesso em: 1 jun. 2004.

CHOMSKY, N. **Syntactic structures**. The Hague: Mouton & Co.,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.

FROHMANN, B. **Documentation redux: prolegomenon to philosophy of information.** United States of Americ**a:** Library Trends, Winter, 2004.

GILLILAND-SWETLAND, A. J. Setting the Stage: Defining Metadata. In: MURTHA BACA (Ed.). **Introduction to Metadata**: Pathways to Digital Information. Los Angeles: Getty Information Institute, 2000. 48p.; pp.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, 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, 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).

_____; MENDONÇA, R. M. A; MALHEIROS, L. R.; COSTA, L. C.; MORAES, G. V. 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.

MEADOWS, A. J. **Communication in science**. London: Butterworths, 1974. 248 p.

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

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

SHERA, J. H.; CLEVELAND, D. B. History and foundations of information science. **Annual Review of Information Science and Technology**, v.12, p.249-275, 1977.

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

ZIMAN, J. **Conhecimento público**. Belo Horizonte: Itatiaia, São Paulo: Editora da Universidade de São Paulo, 1979.