IDENTIFICAÇÃO DE TRAÇOS DE DESCOBERTAS CIENTÍFICAS PELA COMPARAÇÃO DO CONTEÚDO DE ARTIGOS EM CIÊNCIAS BIOMÉDICAS COM ONTOLOGIAS NA WEB

Luciana Reis Malheiros[*]
Carlos Henrique Marcondes[*]

RESUMO

Este trabalho propõe um método para a identificação de indícios de descobertas (ID) significativas na área biomédica através da comparação do conteúdo dos artigos com o conteúdo de ontologias públicas na Web. Assim, seria possível reconhecer ID relatado no artigo antes mesmo dele ser citado pela literatura. Os resultados obtidos indicam que se os conteúdos da conclusão de um artigo estão pobremente representados nestas ontologias, isto pode ser um ID. A proposta metodológica descrita servirá de base, no futuro, para o desenvolvimento de um procedimento automatizado.

Palavras-chaves: conhecimento médico, representação do conhecimento, ontologia, descoberta científica, comunicação científica.

1 INTRODUÇÃO

A comunicação científica formal e, conseqüentemente, a criação dos periódicos científicos está diretamente ligada à criação das sociedades científicas. A *Royal Society*, por exemplo, começou a editar em 1665 as *Philosophical Transactions*, periódico que é editado até hoje. Na introdução do primeiro número Oldenburg (1665, p.1, grifo nosso) dizia: "Nada é mais importante para promover o progresso das Questões Filosóficas do que **comunicar** [...] os estudos e descobertas mais recentes na área"[1]. Desde seu início, havia um espaço no periódico para que fossem publicadas cartas de pesquisadores comentando algum artigo que havia sido publicado.

Com a introdução do periódico, o processo de formalização da comunicação científica teve início e o registro das pesquisas passou a ficar disponível por longo período de tempo para um público bem mais amplo do que aquele que discuta as questões científicas em cartas pessoais. (MEADOWS, 1999).

---

[*] Depto. de Fisiologia e Farmacologia, Universidade Federal Fluminense, doutoranda do Programa de Pós-Graduação em Ciência da Informação UFF/IBICT, malheiro@vm.uff.br
[*] Departamento de Ciência da Informação, Universidade Federal Fluminense, Professor do Programa de Ciência da Informação da UFF, marcon@vm.uff.br
[1] A *Royal Society* colocou disponível de novembro de 2005 a novembro de 2006 todas as revistas editadas pela sociedade, inclusive o primeiro número da *Philosophical Transactions*.

Para Ziman (1979, p.118) “a criação da revista científica teve uma importância muito maior do que qualquer outra iniciativa das Sociedades Reais e Academias Nacionais [...]”. O periódico é uma publicação que , como o próprio nome diz, deve ser regular e deve possibilitar a publicação rápida de um grande número de pesquisas. Ziman (1979, p.119) ressalta que “falar em rapidez como um atributo de uma atividade técnica tão primitiva quanto era a impressão por meio de prensa manual [...] cuja distribuição era feita por navios a vela [...] pode parecer um pouco de presunção”. Contudo, para os padrões da época, essa mudança foi suficiente para que um número muito maior de cientistas pudesse ler, discutir e escrever sobre as descobertas publicadas. De acordo com Kronick (apud HARMON, 1992) o número de periódicos passou de 4 , em 1670, para 118 em 1790.

É importante lembrar que o periódico foi, também, uma resposta ao método científico que dava seus primeiros passos. Os resultados das observações feitas aplicando-se o método eram bem representados no formato de artigos curtos (HARMON, 1992).

O artigo científico tem sua origem na troca de cartas entre cientistas de diversos países europeus. Este método de comunicação, chamado de *Republique des Lettres* , era tão eficaz que às cartas eram acrescentados comentários de outros autores o que dava origem a um texto que podia ser bem diverso do original (SABBATINI, 1999).

Inicialmente, os artigos eram escritos na forma de cartas enviadas para o editor, muitas delas, escritas em língua vernácula, e não em latim, com o propósito de alcançar um público maior de não cientistas. Muitos periódicos mantêm, até hoje, um espaço para a publicação de cartas cujo conteúdo pode fazer alusão a algum artigo publicado no periódico ou relatar o resultado de uma pesquisa.

Entretanto, com o passar dos anos, a profissionalização e a especialização da ciência fez com que os artigos e, conseqüentemente, os periódicos sofressem modificações (HARMON, 1992).

A partir do século XIX a estrutura do artigo científico mudou. O rigor científico esperado de um artigo não combinava com observações pessoais ou linguagem figurada: ao escrever o artigo o autor deveria limitar-se aos fatos.

A estrutura de artigo que se consolidou nesta mudança foi a do artigo composto por:

1 Introdução, onde o problema da pesquisa é delimitado

2 Método, descrição dos materiais e métodos utilizados na pesquisa.

3 Resultados, relato dos resultados obtidos com a aplicação dos métodos de investigação descritos.

4 Discussão dos resultados.

Somente após 1850, os artigos começaram a apresentar o que chamamos referência bibliográfica, isto é, "referências explícitas a trabalhos anteriores sobre os quais se baseia a nova contribuição [...]". (PRICE, 1976, p.42)

O primeiro livro de redação científica lançado nos EUA data de 1927. Desde então, livros de temática semelhante vêm sendo lançados e a estrutura de artigo mencionada acima continua sendo a mais divulgada e aceita (HARMON, 1992).

No final do século XX, com o advento e difusão da *World Wide Web*, tornou-se cada vez mais comum a publicação de artigos científicos em formato digital. Os periódicos científicos publicados na Web podem ser uma ferramenta cognitiva cujas potencialidades estão longe de terem sido totalmente avaliadas. Embora publicados na Web, periódicos eletrônicos são ainda calcados no modelo tradicional das publicações em papel e não utilizam todo o potencial do meio eletrônico. São para serem lidos, avaliados e criticados por pessoas; dependem de um longo processo de leitura, avaliação e citação pelos pares para que os novos conhecimentos possam, enfim, serem incorporados ao acervo de conhecimento público aceito num determinado campo.

Neste processo de comunicação do conhecimento científico fazer citações a outros artigos científicos não é só usual, mas necessário. Hamilton (1990) relata, contudo, que 55% dos artigos publicados em periódicos indexados pelo ISI, de 1981 a 1985, não receberam nenhuma citação cinco anos após serem publicados. E, mesmo os artigos que foram citados, não o foram com freqüência; somente 42% dos artigos citados receberam mais que uma citação.

Um artigo que demore a receber as devidas citações pode fazer parte de um grupo de artigos conhecido como de reconhecimento tardio[2] isto é, artigos que trazem contribuições importantes, mas que num primeiro momento não receberam a devida atenção por parte da comunidade científica. Com o tempo, o valor de um "artigo tardio" é (re)descoberto (CAMPANARIO, 1993).

Niiniluoto (2007, p.5) faz uma crítica severa ao uso dos indicadores cientométricos como instrumentos para a detecção do progresso científico dizendo "que eles não levam em consideração o *conteúdo semântico* das publicações cientificas".

Dentre os fatores que fariam que um artigo importante não recebesse a atenção necessária, destacam-se: o artigo apresentaria conclusões que não vão ao encontro com a teoria mais aceita por uma determinada área; o autor do artigo é um pesquisador iniciante e/ou

[2] Também chamado de descoberta prematura ou descoberta resistente.

trabalha em uma instituição de pesquisa de pouco prestígio; ou ainda, o grande número de artigos publicados impediria que artigos que trazem novas idéias tivessem destaque dentre aqueles que corroboram o conhecimento já estabelecido (GARFIELD, 1970).

O caso mais famoso de reconhecimento tardio é o artigo de Mendel sobre hibridização de plantas e publicado em 1865. O artigo foi citado poucas vezes até ser "redescoberto" em 1900 (GARFIELD, 1970). Garfield fornece o exemplo de outros cinco artigos que podem ser considerados de reconhecimento tardio e que foram identificados através da análise da freqüência de citação. Ele conclui seu trabalho dizendo que o fenômeno de reconhecimento tardio parece ser pouco usual.

Garfield retoma o tema e relata mais artigos que estariam nesta categoria, levantando algumas questões pertinentes como:

> O reconhecimento tardio é mais prevalente em artigos metodológicos ou conceituais? (...) Existe alguma diferença ao longo das últimas décadas, onde a existência de melhores métodos para a recuperação da informação tornou aparentemente mais difícil desconhecer artigos relevantes? Ou existe algum fator de atraso fundamental que deve inevitavelmente afetar a aceitação de novas idéias através do processo educação-pesquisa? (GARFIELD, 1990, p.73)

Em um último trabalho (GLÄNZEL; GARFIELD, 2004) os autores reafirmam que os casos de artigos que têm reconhecimento tardio são poucos e que a maioria dos artigos importantes é muito citada nos primeiros três a cinco anos de publicação. Dos 60 artigos de reconhecimento tardio encontrados por eles, 43% eram da área de ciências da vida.

Van Raan (2004, p.467) chamou de "Belas Adormecidas" os artigos que "passam despercebidos ('dormindo') por um longo tempo e, então, quase que de repente, atraem muita atenção ('são despertados pelo príncipe')". Ele estudou "as belas adormecidas" a partir de três variáveis. A primeira seria "a profundidade do sono", medida pelo número médio de citações recebidas num determinado período de tempo. Os artigos que receberam, no máximo, 1 citação em média por ano foram considerados "em sono profundo"; os que receberam entre 1 e 2 citações em média por ano foram considerados "em sono leve". A segunda variável a ser considerada foi o "tempo de sono", isto é a duração do período em que os artigos receberam, no máximo, 2 citações em média. Por último, considerou-se a "intensidade do despertar", ou seja, o número médio de citações quatro anos após o "despertar".

De um universo de cerca de um milhão de artigos, ele encontrou 41 artigos que depois de um "sono profundo" de 10 anos receberam, em média, 6 a 7 citações nos quatro anos

seguintes. Uma crítica que o próprio autor faz ao seu trabalho é que ele trabalhou com várias áreas do conhecimento e que o padrão de citação de cada área é muito particular.

A Web é um grande repositório e distribuidor de informações sejam textos, imagens ou sons. Graças à utilização de ferramentas próprias, qualquer pessoa pode encontrar essas informações, com diferentes graus de dificuldade, pois ele/a sabe reconhecer o seu significado. O desafio é fazer com que resultados e conclusões de pesquisa, como por exemplo, os encontrados nos artigos científicos, possam ser "interpretados" permitindo que computadores possam nos auxiliar em tarefas mais sofisticadas que demandem o processamento desses dados, diminuindo a intervenção humana e aumentando a precisão das informações obtidas. Particularmente na área biomédica, uma enorme quantidade de informação está disponível em formato digital como, por exemplo, dados sobre seqüenciamento genético (STEIN, 2008), mas que ainda não estão integrados a outras bases de dados, limitando a sua utilidade.

Berners-Lee et al (2001, p.35) propuseram o termo Web Semântica e a definiram como:

> A Web semântica não é uma Web separada, mas uma extensão da atual, na qual a informação é utilizada com significado bem definido, aumentando a capacidade dos computadores para trabalharem em cooperação com as pessoas.

Para que a Web Semântica se torne uma realidade é preciso que várias áreas do conhecimento trabalhem cooperativamente. Uma importante iniciativa nesse sentido é o *World Wide Web Consortium* (W3C). Criado em 1994, o W3C tem o objetivo de: "Fazer com que a Web alcance todo o seu potencial através do desenvolvimento de protocolos e diretrizes que garantam o crescimento em longo prazo da Web" [3].

As ontologias são fundamentais para que a Web Semântica se torne uma realidade. Segundo Ding e Foo (2002, p.375): "Ontologia é definida como uma especificação formal e explicita de uma conceituação compartilhada. Ela fornece um entendimento compartilhado e comum de um domínio que pode ser comunicado a pessoas e sistemas de aplicação."

O objetivo da construção de ontologias é registrar e armazenar conhecimento e permitir que múltiplos sistemas e agentes "entendam" o conteúdo de um recurso da Web e que possam "integrar esse conhecimento com o conteúdo de outros recursos; o sistema ou agente deve ser capaz de interpretar a semântica de cada recurso [...]." (JACOB, 2003, p.19).

De Roure, Jennings e Shadbolt (2001) enfatizam a importância da integração do

[3] http://www.w3.org/Consortium/

conhecimento de diferentes fontes, incluindo artigos científicos publicados na Web aos ambientes de e-Science. Para atingir esta meta o conhecimento precisa estar representado em um formato processável por máquina.

Um dos esforços de representação do conhecimento da área biomédica é o *Unified Medical Language System* (UMLS), um projeto da *National Library of Medicine* (NLM)[4] que combina diversas fontes terminológicas num único instrumento. Ele possui uma estrutura hierárquica, o *Metathesaurus* com cerca de 730.000 conceitos e mais de 1 milhão de nomes de conceitos. Ele é complementado por uma estrutura classificatória chamada de *Semantic Network*.

Desde sua criação existe uma preocupação de agregar profissionais de áreas distintas para pensar sobre o UMLS, assim, bibliotecários, cientistas da informação, lingüistas, cientistas da computação, médicos, biomédicos, dentre outros, sempre fizeram parte da equipe do UMLS (HUMPHREYS et al, 1998).

O objetivo do UMLS é "facilitar o desenvolvimento de sistemas de computadores que se comportem como se 'entendessem' o significado da linguagem biomédica e da saúde" (NATIONAL LIBRARY OF MEDICINE, 2008a, p.1). Para atingir este objetivo, a NLM produz e distribui bases de dados da UMLS, denominadas de UMLSKS (UMLS *Knowledge Sources*). Além da UMLSKS, a NLM produz e distribui, também, *softwares* de apoio que servem de ferramenta para que desenvolvedores de sistemas possam criar ou aperfeiçoar sistemas de informações que processem, criem, recuperem, integrem e/ou agreguem dados e/ou informações biomédicas e da saúde.

O UMLS constitui-se de três bases de conhecimento: um *Metathesaurus* formado pela agregação de mais de 100 vocabulários e classificações; uma *Semantic Network* que "provê uma consistente classificação de todos os conceitos representados no *Metathesaurus*" (NATIONAL LIBRARY OF MEDICINE, 2006, p.1) e um grupo de relações que existem entre os conceitos do *Metathesaurus*; e o *SPECIALIST Lexicon* que possui informações morfológicas, sintáticas e ortográficas para palavras da área biomédica e palavras freqüentemente usadas no inglês (NELSON; POWELL; HUMPHREYS, 2006). Neste trabalho não trataremos do *SPECIALIST Lexicon*.

O UMLS *Metathesaurus* é "uma vasta base de dados de vocabulário, de múltiplos propósitos, multilíngüe e que contém informações sobre conceitos biomédicos e relacionados

[4] A *National Library of Medicine* é um dos Institutos que fazem parte dos *National Institutes of Health* americano.

à saúde, seus vários nomes e suas relações entre eles." (NATIONAL LIBRARY OF MEDICINE, 2008b, p.1). Um desses vocabulários é o MeSH..

Um dos aspectos que mais geraram polêmica na construção do UMLS foi a definição de como o *Metathesaurus* deveria ser elaborado. Não havia consenso sobre a decisão da NLM de construir o *Metathesaurus* a partir da combinação dos conceitos de vocabulários fonte. Entretanto, a NLM argumentava não dispor de recursos para empreender a construção de um vocabulário controlado tão extenso que pudesse atender a demanda do UMLS (HUMPHREYS et al, 1998). A forma utilizada para a construção do *Metathesaurus* implicou em que todos os conceitos, nomes e relações presentes nos diferentes vocabulários básicos estejam presentes no *Metathesaurus*, sendo assim:

> Quando dois diferentes vocabulários fonte usam o mesmo nome para diferentes conceitos, o Metathesaurus representa ambos os significados e indica qual significado está presente em qual vocabulário fonte. Quando o mesmo conceito aparece em contextos hierárquicos diferentes, em diferentes vocabulários fonte, o Metathesaurus inclui todas as hierarquias. Quando relações divergentes entre dois conceitos aparecem em diferentes vocabulários fonte, ambas as visões são incluídas no Metathesaurus. [...] **o Metathesaurus não representa uma abrangente ontologia biomédica de autoria da NLM ou uma única visão consistente do mundo (exceto no mais alto nível dos tipos semânticos atribuídos a todos os seus conceitos**). (NATIONAL LIBRARY OF MEDICINE, 2008b, p. 1, grifo nosso)

Para Bodenreider (2001, p.4) "o *Metathesaurus* pode prover as bases para uma ontologia no domínio biomédico".

A organização do *Metathesaurus* é feita por conceitos e tem por objetivo relacionar diferentes nomes para o mesmo conceito de vocabulários diferentes. O *Metathesaurus* retém todos os identificadores presentes nos vocabulários fonte, além de atribuir diversos tipos de identificadores permanentes e únicos para conceitos e nomes de conceitos que ele possui. Este identificador, atribuído a cada conceito ou significado, é chamado de *Concept Unique Identifier* (CUI) e ele não possui um significado por si só, isto é, não é possível fazer nenhum tipo de inferência somente olhando-os.

No caso de se descobrir que dois CUI referem-se ao mesmo conceito, um CUI é removido do *Metathesaurus* e toda a informação relacionada ao CUI retirado é transferida para o CUI remanescente. Um CUI que foi retirado nunca é reutilizado e a NLM mantém um arquivo que rastreia todas as mudanças realizadas nos CUIs desde 1991.

Como já mencionado anteriormente, a publicação de artigos científico na Web é uma atividade comum no meio científico e a maioria dos periódicos científicos possui uma versão

acessível na Web. Entretanto, os recursos da tecnologia da informação (TI) não são usados diretamente para processar co conhecimento contido no texto de artigos científicos. Artigos publicados eletronicamente são "bases de conhecimento", mas, somente, para a leitura humana. Existem duas barreiras para o uso em larga escala desse conhecimento: a quantidade de informação disponível através da Web e o fato de que o conhecimento está em um formato textual, de forma não estruturada, inadequado para o processamento por programas de computador. Ainda hoje, os periódicos eletrônicos são baseados no modelo do periódico em papel.

Kuhn (2005, p.149) discute a importância das categorias para a percepção de novos fenômenos, no contexto das mudanças de paradigma e diz "Todavia, depois que a experiência em curso forneceu as categorias adicionais indispensáveis, foram capazes de perceber as cartas anômalas (...)". Estabelecer novas categorias e cunhar termos que as representam seria, portanto, algumas das características das mudanças de paradigmas científicos. Contudo, deve-se considerar que uma mudança de paradigma pode acontecer sem a inclusão de novas categorias ou fenômenos, mas, por exemplo, pelo estabelecimento de um novo sistema de relações entre eles. Assim, sempre existirá um intervalo de tempo entre a conceituação de uma nova descoberta e a sua representação como conceito em uma terminologia.

De que forma indícios de descobertas importantes (ID) podem ser identificadas? Artigos que trazem ID têm o conteúdo de suas conclusões bem representado em ontologias públicas do mesmo domínio de conhecimento do artigo? Novos conceitos ou fenômenos recém cunhados são imediatamente representados nessas ontologias?

Acredita-se que um avanço pode ser feito na área de publicação eletrônica. Trabalhamos há anos (MARCONDES et al, 2009) na proposta de um modelo de publicação de artigos científicos cuja proposta é permitir que suas conclusões sejam "inteligíveis" por programas. Artigos, além de serem publicados no formato textual, também teriam as suas conclusões identificadas, extraídas, gravadas e publicadas como instâncias de uma ontologia em um formato processável por máquina. Isto seria um subproduto do processo de autopublicação onde os próprios autores descreveriam as suas conclusões ao submeterem o artigo a um sistema de publicação eletrônica de um periódico.

A nossa abordagem à representação do conhecimento das conclusões de artigos científicos é baseada no fato do conhecimento científico ser constituído por asserções feitas pelos cientistas no texto dos artigos, expressando relações entre fenômenos ou entre um fenômeno e suas características. Considerou-se as relações como a unidade básica do conhecimento científico e que sintetizariam as conclusões do artigo. A partir do momento em

que as conclusões puderem ser extraídas, marcadas como relações e gravadas em um formato processável por máquina, será possível o seu processamento por agentes de software, fornecendo aos cientistas novos meios de recuperar, comparar e avaliar este conhecimento.

Uma vez representada em um formato processável por máquina as conclusões dos artigos poderão ser comparadas pelos programas com o conhecimento registrado em ontologias públicas na Web revelando, então, inconsistências, erros e possíveis índicos de descobertas. Dessa maneira é possível que um artigo científico, no momento de sua publicação em um periódico eletrônico e sem ainda ter sido referenciado ou citado, revelar indícios que podem indicar que ele traz uma descoberta importante.

A nossa hipótese é que existe uma correlação entre um artigo cuja conclusão é fracamente representada ou representada somente de modo genérico em bancos de dados terminológicos, como o UMLS (o *Unified Medical Language System*), e o fato desses artigos reportarem descobertas científicas importantes.

Isso é facilmente percebido quando se compara a defasagem entre as palavras-chaves do autor em artigos biomédicos com os descritores do *Medical Subject Headings* (MeSH) atribuídos aos ao artigo quando este é depositado em bibliotecas digitais como o PubMed. Um indício a favor dessa hipótese é o fato de que, entre os artigos analisados do grupo que reporta a descoberta da enzima telomerase, o artigo que marca a descoberta da enzima é de 1985 (GREIDER; BLACKBURN 1985), mas o termo telomerase só foi incluído no MeSH 10 anos depois.

O objetivo desde trabalho é demonstrar a viabilidade de um método que compare as conclusões de artigos científicos com o conhecimento expresso em ontologias públicas na Web a fim de identificar possíveis descobertas importantes.

2 MATERIAL E MÉTODOS

Artigos da área Biomédica foram escolhidos como material empírico já que eles costumam apresentar uma estrutura mais rígida, contendo: Introdução, Métodos[5], Resultados e Discussão (IMRD[6]). Segundo Burrough-Boenisch (1999, p.296) "os cientistas escrevem neste formato, não somente para cumprir os requerimentos dos periódicos, mas também para atender as expectativas da comunidade científica." Ele também comenta que a maioria dos manuais de redação científica encoraja o uso da estrutura IMRD por considerá-la a mais adequada para a organização do artigo científico. O ICMJE - *International Committee of*

[5] Também pode ser denominado de Material e Métodos.
[6] Alguns autores abreviam como IMRAD sendo o "A" de "AND".

Medical Journals Editors (2008, p.11) diz que a estrutura IMRD "não é um formato arbitrário para a publicação, mas um reflexo direto do processo de descoberta científica".

No total, foram analisados manualmente 75 artigos da área biomédica. Vinte artigos do periódico Memórias do Instituto Oswaldo Cruz (MIOC), 20 do *Brazilian Journal of Medical and Biological Research* (BJMBR), 20 artigos que tratavam da pesquisa com células tronco e 15 artigos dos ganhadores do Lasker de 2006. O prêmio Lasker é um importante prêmio, outorgado anualmente e considerado tão importante quanto o Nobel, apesar de menos conhecido. Ele é tido como uma premiação que, muitas vezes, antecipa o Nobel e, segundo Garfield (1984, p.405) "40 ganhadores do prêmio Lasker receberam o Nobel, 39 deles antes e 1 depois de receber o Nobel."

Os artigos do MIOC e do BJMBR foram escolhidos através do portal Scielo utilizando a lista de artigos mais visitados de cada um deles. Ambos os periódicos publicam artigos em inglês e possuem um corpo editorial qualificado, com revisores nacionais e internacionais.

O primeiro grupo de artigos analisados consistiu-se de artigos do MIOC que é editado desde 1909 e tem uma excelente reputação nacional e internacional. Posteriormente, analisamos os artigos do BJMBR que é editado desde 1981 e substituiu a Revista Brasileira de Pesquisas Médicas e Biológicas. Tanto o MIOC como o BJMBR são indexados pelo Scielo, LILACS, Medline e ISI/Thompson. O fator de impacto (2006) para o BJMBR foi de 1,075 e de 1,208 para o MIOC.

Na busca de artigos que trouxessem indícios de descobertas importantes, o terceiro grupo de artigos analisados tratava de pesquisas sobre células tronco. A seleção dos artigos desse grupo foi feita através da leitura de três artigos de revisão recentes da área (NATIONAL INSTITUTES OF HEALTH, 2006; FRIEL; SAR; MEE, 2005; BONGSO; RICHARDS, 2004) que apresentavam uma visão história da pesquisa em células tronco, destacando os avanços mais importantes, informação extremamente relevante para essa pesquisa.

Ainda buscando artigos que reportassem descobertas importantes, um último grupo de artigos foi escolhido. Esse grupo foi formado por artigos que constavam da bibliografia selecionada de três pesquisadores - Elizabeth H. Blackburn, Carol W. Greider e Jack W. Szostak - ganhadores, em 2006, do prêmio Albert Lasker de Pesquisa Médica Básica pelos trabalhos que levaram a descoberta da telomerase[7]. Cada autor laureado forneceu uma lista dos seus trabalhos que julgavam mais importantes e, da junção das três listas, obtiveram-se os

[7] Na página da Fundação Lasker a justificativa para o prêmio é colocada da seguinte forma: "por terem predito e descoberto a telomerase, uma importante enzima que contém uma porção RNA e que sintetiza a porção final dos cromossomas, protegendo-os e mantendo a integridade do genoma."
http://www.laskerfoundation.org/awards/2006basic.htm

15 artigos analisados. Deste último grupo constam artigos de vários periódicos científicos como *Cell* e *Nature*, periódicos com alto fator de impacto, 29,887 e 28,751, respectivamente.

A análise do conteúdo dos artigos do Lasker foi facilitada pelos comentários feitos pelos próprios autores sobre a maioria dos artigos selecionados. Esses comentários fazem parte da revisão que os autores escreveram para a *Nature Medicine* (BLACKBURN, GREIDER e SZOSTAK, 2006) por ocasião da premiação com o Lasker. Nela, os autores apresentam a trajetória das pesquisas destacando os artigos que julgam mais importantes e especificando a contribuição dada por cada um deles.

Selecionaram-se periódicos da área Biomédica devido à estrutura textual altamente formalizada de seus artigos. A maior parte dos artigos do MIOC era da área de microbiologia; os do BJMBR eram mais heterogêneos havendo uma predominância de artigos das áreas de fisiologia e neurociências; por fim, os artigos de células tronco e telomerase que tratavam de questões relacionadas à genética. É importante enfatizar que a escolha desses periódicos não foi feita em um único momento, mas, gradualmente, ao longo do desenvolvimento da pesquisa, sempre procurando artigos que trouxessem descobertas científicas importantes, foco desse trabalho.

O processo de análise dos artigos deu-se em duas etapas. Em um primeiro momento, o grupo deveria tentar identificar no texto qual era a principal conclusão apresentada pelos autores. Para esta tarefa lançou-se mão, também, de artigos de revisão que faziam referência ao trabalho analisado. Identificada a principal conclusão, discutia-se a melhor maneira de expressá-la sinteticamente na forma de antecedente (um conceito que se refere a um fenômeno), uma relação semântica e um conseqüente (outro conceito que se refere a um fenômeno ou uma característica do fenômeno expresso no antecedente). Como, por exemplo, a análise do artigo "*A mutant with a defect in telomere elongation leads to senescence in yeast*" (LUNDBLAD; SZOSTAK, 1989) a conclusão foi sintetizada na seguinte afirmação: O encurtamento do telômero causa senescência celular. Ou esquematicamente:

Antecedente: encurtamento do telômero

Relação: causa

Conseqüente: senescência celular

Os descritores MesH desse artigo são: aging/physiology*, alleles, amino acid sequence, base sequence, cell survival, chromossome aberrations, chromossome disorders, chromossome/physiology*, cloning/molecular, DNA/analysis, molecular sequence data, mutation*, phenotype, Saccharomyces cerevisiae/genetics*.

O artigo ao ser publicado é quase que imediatamente indexado. Considerando que a indexação foi feita com os melhores termos disponíveis na época de publicação e estabelecidos o antecedente, a relação e o conseqüente, verificou-se em que grau estes elementos estavam representados na indexação MeSH do artigo. A isso, foi dado o nome de mapeamento. Se todos os elementos (o antecedente, relação e conseqüente) fossem mapeados no UMLS, o artigo era considerado completamente mapeado. Se um ou dois elementos não fosse mapeado no UMLS, o artigo era considerado parcialmente mapeado. Finalmente, se nenhum dos elementos fosse mapeado, o artigo era considerado não mapeado.

Embora "novidade científica" seja uma noção sem uma definição precisa e a literatura mostre tentativas de definição mais qualitativas, para efeitos desta pesquisa vamos considerar indício de novidade científica o mapeamento parcial ou o não mapeamento de conceitos da conclusão do artigo em ontologias públicas no mesmo domínio científico do artigo.

3 RESULTADOS

Dentre os 75 artigos analisados, o grupo que reportava indícios de descobertas importantes, isto é, os artigos dos ganhadores do prêmio Lasker de 2006, seguidos dos artigos de células-tronco – obteve o pior percentual de mapeamento Nesse grupo, todos os artigos não foram mapeados ou foram parcialmente mapeados. Dentre os artigos parcialmente mapeados, o mapeamento só foi possível por conta do tipo de relação. Não foi possível mapear nem o antecedente nem o conseqüente.

No grupo de artigos de células-tronco, 80% foram parcialmente mapeados e 20% foram não mapeados.

Somando os artigos do MIOC e BJMBR e os comparado com a soma dos artigos do Lasker e de células-tronco (L+CT), os artigos MIOC+BJMBR receberam mapeamento completo em percentual maior (25%) que os L+CT (0%).

Tabela I. Análise da representação da conclusão dos artigos dos periódicos Memórias do Instituto Oswaldo Cruz (MIOC), Brazilian Journal of Medical and Biological Reserach (BJMBR), artigos relevantes sobre células-tronco e artigos sobre a pesquisa da telomerase de ganhadores do prêmio Lasker.

Artigos analisados	**MIOC**	**BJMBR**	**Células-Tronco**	**Telomerase**	**TOTAL**
Totalmente mapeados	7 (35%)	3 (15%)	0 (0%)	0 (0%)	10
Parcialmente mapeados	13 (65%)	11 (55%)	16 (80%)	6 (40%)	44
Não mapeados	0 (0%)	6 (30%)	4 (20%)	9 (60%)	21
Total de artigos	20	20	20	15	75

4 DISCUSSÃO

Observa-se, na literatura, que o termo "ontologia biomédica" pode referir-se tanto a terminologias usadas para indexar a literatura científica como a ontologias computacionais formais de alto nível. Seu desenvolvimento, evolução e integração é uma empreitada científica complexa. Enquanto, por exemplo, a Gene Ontology foi desenvolvida recentemente de forma a permitir o compartilhamento de uma terminologia comum para a anotação de produtos genéticos, outras, com o MesH, que é parte do UMLS, tem que lidar com o legado de milhões de registros indexados em bases de dados bibliográficas como o PubMed e o Medline (Bodenreider, 2008).

Ao se tornarem mais formais, as terminologias biomédicas estão evoluindo para serem bases do conhecimento. As ontologias são classificadas de acordo com o seu grau de formalismo, podendo variar desde uma simples taxonomia usada por pessoas até uma ontologia altamente formalizada codificada em uma linguagem como o OWL.

O objetivo desse trabalho foi demonstrar a viabilidade de um método que permitisse comparar as conclusões de artigos científicos com o conhecimento registrado em ontologias públicas na Web a fim de identificar possíveis descobertas importantes. No momento, não dispomos de uma ontologia altamente formal na área biomédica, porém, acredita-se que com o desenvolvimento desta área, o método aqui proposto poderá apontar indícios de possíveis descobertas científicas de forma muito mais precisa. O MesH foi usado aqui como uma ferramenta na falta, no momento, de outra melhor.

Os resultados indicam que o grau de sucesso/insucesso obtido pelo mapeamento da representação da conclusão usando o MeSH pode estar associado ao fato de que os artigos relatam descobertas científicas importantes. Parece metodologicamente possível propor um procedimento em que os autores expressem sua principal conclusão de maneira sintética e que a mesma seja automaticamente processada e comparada ao conhecimento científico já previamente estabelecido e representado nas ontologias públicas.

A crescente quantidade de artigos que vêm sendo publicada constantemente, em especial na área biomédica, torna muito difícil e lento o processo de identificação por pesquisadores de possíveis artigos relevantes, sua leitura, avaliação, crítica e eventual citação. Um método automático que possa apontar indícios de novidades pode otimizar este processo, fazendo com que a atenção do pesquisador ou do gestor de C&T possa se concentrar em artigos que sejam potencialmente relevantes.

Tal procedimento pode chamar a atenção dos cientistas para indícios de descobertas científicas importantes. Na amostra analisada, artigos com baixo percentual de mapeamento total eram aqueles que relatavam descobertas científicas importantes

Deve-se considerar que a indexação dos artigos não é feita pelos autores que conhecem melhor o que está sendo relatado e a contribuição que estão dando para a ciência. A indexação é feita, posteriormente e logo após a publicação, quando os artigos são incluídos em bases de dados ou repositórios como o Medline ou Pubmed.

Uma nova descoberta científica pode criar novos conceitos para os quais um termo ainda não foi cunhado nas bases de dados terminológicas como o UMLS. Existe um atraso entre a descoberta de um fenômeno, ou conceito, e a atualização do UMLS. Como já citado anteriormente, o termo telomerase foi relatado em 1985 (GREIDER; BLACKBURN, 1985), mas só foi incorporado ao MeSH em 1995, dez anos depois.

É importante ressaltar que, em alguns casos a "novidade científica" não é acompanhada da criação de novos termos, mas, por exemplo, pela maneira como dois fenômenos se relacionam.

Com o crescimento das ontologias como novos artefatos científicos (SMITH, 2008), provavelmente, haverá novos processos de validação/ratificação científicos. As ontologias também estão evoluindo para uma maior formalização e necessitarão de novos métodos de curadoria (WILLIAMS; ANDERSON, 2003).

O mesmo pode ser dito dos artigos científicos publicados em formato digital: assim que eles puderem ser publicados em um formato mais rico e formal, isto possibilitará o processamento de sua conclusão e a comparação com ontologias públicas da Web, conforme proposto aqui.

Acredita-se que o método proposto, depois de totalmente automatizado e implementado, possa vir a ser mais uma ferramenta de avaliação da produção científica, complementar aos já tradicionais métodos bibliográficos e cientométricos.

IDENTIFICATION OF LINES OF SCIENTIFIC DISCOVERY BY COMPARISON OF THE CONTENT OF ARTICLES IN BIOMEDICAL SCIENCES WITH THE WEB ONTOLOGY

ABSTRACT

This paper reports a methodological proposal consisting of comparing the content of scientific articles with the content of Web public ontologies in order to identify traces of scientific discoveries (SD) reported by the article. Articles which its content is poorly represented in

those ontologies are strong candidates to report traces of SD. The methodological proposal described will be the basis for the future development of an automatic procedure. New scientific indicators might be derived from the findings reported.

Keywords: medical knowledge, knowledge representation, ontology, scientific discovery, scientific communication.

REFERÊNCIAS

BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. The semantic web. **Scientific American**, v. 284, n.5, p. 34-43, 2001. Disponível em: <http://www.sciam.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>. Acesso em 16 jun. 2006.

BLACKBURN, Elizabeth H.; GREIDER, Carol W.; SZOSTAK, Jack W. Telomeres and telomerase: the path from maize, *Tetrahymena* and yeast to human cancer and aging. **Nature Medicine**, v. 12, n.10, p. vii-xii, 2006.

BONDEREIDER, Olivier. Medical ontology research. **Report to the board of scientific counselors of the Lister Hill National Center for Biomedical Communications**, 2001. Disponível em: < http://mor.nlm.nih.gov/> Acesso em: 06. jan. 2008

_____.Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support. **Yearb. Med. Inform.**, p. 67-79, 2008. Disponível em: <http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=2592252&blobtype=pdf>.Acesso em: 31.ago.09

BONGSO, A.; RICHARDS, M.History and perspective of stem cell research. **Best Practice & Research Clinical Obstetrics & Gynaecology**, v.18, n.6, p. 827-842, 2004.

BURROUGH-BOENISH, Joy. International reading strategies for IMRD articles. **Written Communication** , v.16, n.3, p.296-316, 1999.

CAMPANARIO, Juan Miguel. Consolation for scientists: sometimes it is hard to publish papers that are later highly-cited. **Social Studies of Science**, v.23, p. 342-362, 1993.

DE ROURE, David; JENNINGS, Nicholas; SHADBOLT, Nigel. Research agenda for the Semantic Grid: a future s-Science infrastructure. **Report Commissioned for EPSRC/DTI Core e-Science Programme**. 2001, 78p.

DING, Ying; FOO, Schubert. Ontology research and development. Part 2 – a review of ontology mapping and evolving. **Journal of Information Science**, v. 28, n. 5, p. 375-88, 2002.

FRIEL, R.; SAR, S.; MEE, P.Embryonic stem cells: understanding their history, cell biology and signalling. **Advanced Drug Delivery Reviews**, v.57, n.13, p. 1894-1903, 2005.

GARFIELD, Eugene. Would Mendel´s work have been ignored if the Science Citation Index® was available 100 years ago? **Essays of an Information Scientist**, v.1, p. 69-70, 1970.

_____. The awards of Science: beyond the Nobel prize. Part 2. The winners and their most-cited papers. _____, v.7, p. 405-419, 1984.

_____. More delayed recognition. Part 2. From inhibin to Scanning Electron Microscopy. _____, v.13, p. 68-74, 1990.

GLÄNZEL, Wolfgang; GARFIELD, Eugene. The myth of delayed recognition. **The Scientist**, v.18, n. 11, p.8, 2004.

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

HAMILTON, David P. Publishing by – and for? – the numbers. **Science**, v. 250, n. 4986, p. 1331-32, 1990.

HARMON, Joseph E. Evolution of the Scientific Paper. In: INTERNATIONAL PROFESIONAL COMMUNICATION CONFERENCE (IPCC), 1992, Santa Fé. **Proceedings…**[S.I.:s.n.], 1992. p. 468-475.

HUMPHREYS, Betsy L ; LINDENBERG, Donald A. B.; SCHOOLMAN, Harold M.; BARNETT, G. OCTO. The Unified Medical Language System : an informatics research collaboration. **Journal of the American medical Informatics Association**, v.5, n.1, p.1-11, 1998.

INTERNATIONAL COMMITTEE OF MEDICAL JOURNAL EDITOR (ICMJE). Uniform requirements for manuscripts submitted to Biomedical journals: writing and editing for biomedical publication. ICMJE, 2008. p.1-16. Disponível em: <http://www.icmje.org/#prepare>. Acesso em: 22 oct. 2008.

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

KUHN, Thomas Samuel. **A estrutura das revoluções científicas**. São Paulo: Perspectiva, 2005. (Coleção Debates, 115).

LUNDBLAD, V.; SZOSTAK, J.W. A mutant with a defect in telomere elongation leads to senescence in yeast. **Cell**, v. 57, n.4, p. 633-643, 1989.

MARCONDES, C.H.; MENDONÇA, M.A.R.; MALHEIROS, L.R.; COSTA, L.C.; SANTOS, T.C.P. Ontological and conceptual bases for a scientific knowledge model in biomedical articles. **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 abril 2009.

MEADOWS, Arthur Jack. **A Comunicação científica**. Brasília, DF: Briquet de Lemos, 1999. 268 p.

NATIONAL INSTITUTES OF HEALTH. **The Human Embryonic Stem Cell and the Human Embryonic Germ Cell**. Disponível em: <http://stemcells.nih.gov/>. Acesso em: 8 mar. 2006.

NATIONAL LIBRARY OF MEDICINE. **Unified Medical Language System – Fact sheet**, 2006. Disponível em: < http://www.nlm.nih.gov/pubs/factsheets/umls.html>. Acesso em: 04 de jan. de 2008.

_____. **Unified Medical Language System**, 2008a. Disponível em: < http://www.nlm.nih.gov/research/umls/about_umls.html>. Acesso em: 04 de jan. de 2008.

_____. **Unified Medical Language System - Metathesaurus**, 2008b. Disponível em: < http://www.nlm.nih.gov/research/umls/meta2.html>. Acesso em: 04 de jan. de 2008

NELSON, Stuart J.; POWELL, Tammy; HUMPHREYS, Betsy L. **The Unified Medical System® (UMLS®) project**, 2006. Disponível em:< http://www.nlm.nih.gov/mesh/umlsforelis.html> Acesso em: 04 de jan. de 2008.

NIINILUOTO, Ilkka. Scientific Progress. In: ZALTA, E.N. (Ed.). **The Stanford Encyclopedia of Philosophy**. feb. 2007. Disponível em: <http://plato.stanford.edu/archives/fall2008/entries/scientific-progress/>. Acesso em: 01 fev. 2008.

OLDENBURG, Henry. Introduction. **Philosophical Transactions**, v.1, n.1, p.1-2, 1665.

PRICE, Derek J. de Solla. **O desenvolvimento da ciência**. Rio de Janeiro: Livros Técnicos e Científicos, 1976. 96 p.

SABBATINI, Marcelo. **Evolución histórica de las publicaciones científicas: de la Republique des Lettres hasta la World Wide Web**. Salamanca, 1999. Trabalho de curso apresentado ao Máster CTS, Cultura y Comuncación en Ciencia y Tecnología. Disponível em: <http://www.sabbatini.com/marcelo/artigos/1999sabbatini-republique.pdf>. Acesso em: 02 de junho de 2008.

SMITH, Barry. Ontology(Science). **Nature Precedings**, 2008. Disponível em <http://hdl.handle.net/10101/npre.2008.2027.2.> Acesso em: 1 ago.2009.

STEIN, Lincon D. Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. **Nature Reviews Genetic**, v. 9, p. 678-688, Sept. 2008.

VAN RAAN, Anthony F. J. Sleeping Beauties in Science. **Scientometrics**, v.59, n.3, p.467-472, 2004.

WILLIAMS, Jennifer; ANDERSON, William. Bringing ontology to the Gene Ontology. **Comparative and Functional Genomics**, v. 4, p.90–93, 2003. Disponível em: <http://hindawi.com/GetPDF.aspx?doi=10.1002/cfg.253>. Acesso em 31 jul. 2009.

ZIMAN, John. **Conhecimento público**. Belo Horizonte: Ed. Itatiaia; São Paulo: Editora da Universidade de São Paulo, 1979. (Coleção o Homem e a Ciência).