DOI: 10.5433/1981-8920.2016v21n2p170

DESCOBERTA DE CONHECIMENTO EM ARTIGOS DIGITAIS EM CIÊNCIAS BIOMÉDICAS

DESCOBRIMIENTO DE CONOCIMIENTO EN LOS ARTICULOS DIGITALES EN CIENCIAS BIOMEDICAS

Carlos Henrique Marcondes*
Leonardo Cruz da Costa**
Sergio de Castro Martins***

RESUMO
Introdução: A emergência da Web Semântica vem impactando o ambiente de publicações digitais científicas. O crescimento da literatura biomédica em formato digital suscita a questão de que novas descobertas podem ter origem não só nos laboratórios, mas também nas bases de dados bibliográficas e factuais. O formato textual linear dos artigos científicos, voltado para leitura humana, é inadequado ao tratamento por computadores.
Objetivos: Contextualizar o problema de pesquisa de modelos semântico para artigos digitais em ciências biomédicas, identificando áreas correlatas e produzindo uma revisão da literatura.
Metodologia: A pesquisa é descritiva e exploratória, centrada no problema da de pesquisa, procurando identificar interfaces entre as áreas revistas; a abordagem é qualitativa, os métodos foram pesquisa bibliográfica e documentaria.
Resultados: apresentação do estado da arte e principais questões em cada subárea, mineração de textos biomédicos e publicações estendidas/modelos semânticos de publicações; são áreas complementares mas não integradas.
Conclusões: Estamos em transição de um modelo de publicação centrado no artigo textual linear para um modelo em que o artigo digital, em um novo formato, é um dos elos de uma rede de recursos interligados e acessíveis simultaneamente, processáveis por programas, tirando partido das tecnologias da Web Semântica.

Palavras-chave: Descoberta de conhecimento na literatura. Mineração de textos. Publicações estendidas. Publicações semânticas. Ciências biomédicas.

* Professor do Programa de Pós Graduação em Ciência da Informação da Universidade Federal Fluminense. E-mail: marcon@vm.uff.br
** Professor do Programa de Pós Graduação em Ciência da Informação da Universidade Federal Fluminense. E-mail: leo.cruz@yahoo.com.br
*** Doutorando do Programa de Pós Graduação em Ciência da Informação da Universidade Federal Fluminense. E-mail: sergio.scm@gmail.com

1 **INTRODUÇÃO:** DESCOBERTA E REUSO DE CONHECIMENTOS EM ARTIGOS CIENTÍFICOS BIOMÉDICOS DIGITAIS

Nos últimos anos a capacidade da humanidade de gerar e colecionar informação tem aumentado rapidamente. Pessoas e organizações dependem cada vez mais da capacidade de lidar com dados e informações. Situação comparável se dá com o campo da ciência.

A literatura publicada em ciências biomédicas é, potencialmente, uma fonte de novos conhecimentos. Repositórios digitais como o PubMed (http://www.ncbi.nlm.nih.gov/pubmed) mantém hoje cerca de 25 milhões de artigos, a maior parte deles de texto completo, em formato digital; milhares de novos artigos são adicionados diariamente. Estes artigos reportam as mais recentes descobertas em ciências biomédicas.

Attwood et al. (2009) chamam a atenção para o potencial de conhecimento contido na literatura científica. Hoje em dia esta literatura esta basicamente disponível em formatos como o PDF, “metáforas” digitais do artigo impresso, com limitadas funcionalidades para processamento automático do seu conteúdo. O artigo científico digital continua sendo uma cópia do artigo impresso, adequada para a leitura por pessoas, mas totalmente inadequada para processamento por computadores.

As bases de dados bibliográficas, pela quantidade de informações e pelo seu crescimento, só podem ser eficientemente gerenciadas através de métodos computacionais. No entanto o acesso e reuso deste conhecimento é bastante problemático. Além do formato textual dos artigos digitais, estas bases de dados, em sua maioria, não são interoperáveis. Para serem acessadas utilizam sistemas de recuperação da Informação tradicionais com algoritmos de recuperação baseados na pouco expressiva Lógica Booleana da década de 1970. Por sua vez o processamento do conhecimento contido no texto de artigos com vistas ao reuso - identificação de lacunas, contradições ou concordâncias no conhecimento de determinada área, validação dos resultados de uma pesquisa - é extremamente trabalhoso, por demandar leitura e processamento por humanos.

Estudos com dados recolhidos por extensos períodos (TENOPIR; KING, 2009) comprovam que pesquisadores precisam ler um número cada vez maior

de artigos e dispensam cada vez menos tempos na leitura de cada artigo. Devido à "explosão informacional" tornada crítica coma Web e as publicações científicas digitais, pesquisadores têm sempre lido estrategicamente (RENEAR; PALMER, 2009), trabalhando com diferentes artigos e fontes de informação simultaneamente, comparando-os, analisando os fragmentos de texto. Há uma necessidade urgente de novas ferramentas para melhorar a leitura estratégica e o gerenciamento de conhecimentos no ambiente Web. Projetos de pesquisa integrados e internacionais identificam esta necessidade:

> Semantic text analysis approaches for extraction of knowledge from biomedical literature - Biomedical literature, for example Pubmed, represents a vast and valuable resource for life sciences research. The ability to extract relevant knowledge from biomedical text and its representation in Semantic Web standard formats such as RDF is an important research issue that is being addressed in this project (KNOESIS, 2016).

A descoberta e reuso de conhecimentos no texto de artigos científicos biomédicos digitais por métodos computacionais, também chamada de "automated literature analysis" (REBHOLZ-SCHUHMANN; OELLRICH; HOEHNDORF, 2012), tornam-se assim um problema de pesquisa relevante. Avanços nesta área são previstos e demandados de forma crescente pela comunidade de pesquisa nas Ciências Biomédicas (STEIN, 2008). Alternativas como a mineração de textos têm sido tentadas para processar o conteúdo de textos de artigos científicos, identificar neles conceitos de vocabulários controlados, terminologias padronizadas ou ontologias (FAKUDA et al., 1998; SRINIVASAN, 2001; TANABE et al., 1999).

Outra alternativa é trabalhar no sentido de propor novos mecanismos de publicação e novos formatos de artigos digitais que garantam que conteúdos de artigos científicos possam ser publicados já diretamente de formato estruturado, processável por programas.

A área de Ciência da Informação mostra crescente interesse no tema, como pode ser demonstrado na revisão do ARIST de Ganiz, Pottenger e Janneck (2005) e nas anteriores aí citadas.

Por outro lado o ambiente digital e as tecnologias surgidas com a Web Semântica e LOD - "linked open data" - permitem novas oportunidades e apontam para um reposicionamento do artigo científico como elemento único e central na comunicação científica. O tradicional artigo em formato impresso como praticamente o único registro público da atividade científica, pode evoluir agora para uma ampla variedade de registros (LYNCH, 2009) incluindo além de imagens estáticas e em movimento, uma quantidade massiva de dados digitais provenientes de diferentes instrumentos de coleta, de simulações computacionais, etc. Um conjunto de entidades, antes meramente mencionadas nos textos dos artigos científicos, agora podem ser diretamente "linkadas" ao artigo, desde termos em sua forma padronizada em terminologias e ontologias computacionais, fórmulas químicas, visualizações de moléculas e gens, até os próprios conjuntos dados originais e códigos computacionais utilizados na pesquisa.[1] Diferentes fontes de informação, através dessas tecnologias, podem tornar-se interoperáveis e utilizadas simultaneamente, aumentando o potencial cognitivo e comunicacional dos registros científicos.

Trabalhamos há anos na proposta de um modelo "semântico" para artigos digitais em ciências biomédicas (MARCONDES, 2005) que permita processar e consultar o conteúdo do artigo por programas. Várias áreas de pesquisa são correlatas e concorrem para obter como resultado o processamento do conhecimento em artigos científicos por meios automatizados. O objetivo deste artigo é fornecer um panorama atual desse problema de pesquisa, suas inter-relações e interfaces, na forma de uma revisão.

Metodologicamente a pesquisa tem abordagem qualitativa; é descritiva e exploratória, centrada no problema da descoberta e reuso de conhecimentos no texto de artigos científicos biomédicos digitais por métodos computacionais; os métodos empregados foram pesquisa bibliográfica e documental. Usou-se como material para o levantamento bibliográfico várias revisões sobre os temas.

[1] Ver a iniciativa DATACITE, http://www.datacite.org.

O artigo está estruturado como se segue. Após esta seção, a seção 2 apresenta o a infraestrutura digital que vem sendo desenvolvida para as ciências e os impactos da grande disponibilidade de dados – a nova Ciência de Dados – para a Ciência e a metodologia científica. Nas seções seguintes são apresentados os resultados da revisão de literatura. A seção 3 discute o tema mineração de dados em textos biomédicos, identificação de entidades, relações entre genes, proteínas, doenças e substâncias farmacológicas, e anotação semântica. A seção 4 discute as chamadas publicações estendidas e publicações semânticas e apresenta nossa proposta de pesquisa. Por fim, a seção 5 apresenta os comentários finais, futuras direções de pesquisa e conclusões.

2 CONTEXTUALIZANDO A CIÊNCIA DE DADOS

Há muito se discute as alterações e rupturas pelas quais as sociedades – sobretudo ocidentais – convivem na atualidade, notadamente com a utilização massiva das tecnologias da informação e comunicação (TIC). Se no início do século XX a sociedade industrial primava pela produção industrial em larga escala e efetuava a difusão de informações pelos tradicionais e analógicos meios de comunicação de massa, nas últimas décadas este cenário passou por mudanças paradigmáticas que permeou praticamente todos os aspectos destas sociedades. Sendo uma característica da chamada sociedade ocidental contemporânea, adjetivada de sociedade pós-capitalista, sociedade pós-industrial ou mesmo sociedade da informação (BELL, 1977; DRUCKER, 1995; MATTELART, 2002), o uso intensivo das tecnologias da informação já possui um efeito irreversível sobre o *modus operandi* das práticas sociais, culturais, empresariais, governamentais e científicas da atualidade. Praticamente todos os setores sociais operam com o apoio de aparatos e plataformas de tecnologias da informação para seu funcionamento. Desta forma, este novo tipo de sociedade é pautado pelo consumo de dados e informações em abundância, visando a criação ou melhoria de serviços, bens, atendimentos de demandas e exercícios diversos.

Após o advento da internet pode-se afirmar que uma nova era da informação teve início. Progressivamente, não somente computadores, mas

inúmeros dispositivos e aparelhos de TIC possibilitaram uma interação e interconexão entre as pessoas e entre dispositivos numa escala sem precedentes. Na medida em que criam estas interações e interconexões, estas novas TICs produzem uma quantidade imensurável de dados brutos, sobretudo não estruturados, nos mais variados formatos: textos, imagens, vídeos, sons, dentre outros, fornecendo uma fonte praticamente inesgotável de informação potencial.

Nas últimas décadas o uso de repositórios, armazéns e bases de dados com base nas TICs e na internet tem se intensificado, alterando as formas como dados e informações são acessados, decisões são tomadas, pesquisas são feitas e aprendizados são realizados. Eles possibilitaram novas configurações de relações sociais, além de novas práticas e modelos científicos e empresariais. Conceitos como *e-science, e-commerce, e-business, e-learning, e-government, big data*, dentre outros, têm permeado não somente o jargão econômico e social desta sociedade da informação, mas também têm sido cada vez mais objeto de análises científicas. Considerando que o volume de dados aumenta exponencialmente a cada ano, capturados pelos diversos dispositivos, gerando variados formatos de dados e estocados em repositórios e armazéns de ampla capacidade, novas técnicas e metodologias de gestão de dados e informações têm sido desenvolvidas. Assim, é neste contexto que surge a chamada *Data Science* – ou Ciência de Dados – cujo foco declarado é a modelagem, interpretação e representação de dados, mediante o uso de complexos softwares para o estabelecimento de padrões semânticos inteligíveis.

Conquanto já seja discutida há alguns anos nas diversas mídias, redes sociais e no mundo corporativo, as análises acadêmicas sobre a Ciência de Dados estão em relativo início, considerando o tempo de pesquisa requerido para tal empreitada. Ademais, nesse contexto também um novo profissional ou cientista emerge: o cientista de dados, cuja função é estudar e aplicar aspectos teóricos e práticos da Ciência de Dados (DAVENPORT, 2014; STANTON, 2012). No que tange ao seu status como disciplina acadêmica, Smith (2006) observou que, tal como aconteceu com a Ciência da Computação, a Ciência de Dados é uma disciplina que se encontra em franco processo de consolidação,

visto que não somente periódicos, mas também uma crescente literatura tem sido desenvolvida para cuidar dos temas da área, além da proliferação de cursos em nível de graduação e pós-graduação. Também várias instituições na área de Ciência de Dados têm sido criadas para o fomento de atividades, intercâmbio, eventos, estabelecimento de padrões e outros aspectos relevantes. De acordo com o *Journal of Data Science*, por "*Data Science, we mean almost everything that has something to do with data: collecting, analyzing, modeling... yet the most important part is its applications, all sorts of applications*." (ABOUT…, 2015).

Agarwal e Dhar (2014) reconhecem que as questões inerentes à Ciência de Dados não são necessariamente novas, pois o tratamento e a mineração de dados já existem há muito tempo. Entretanto, alguns aspectos significativos provocaram a emergência deste campo sob as características que ora se apresentam. Entre estes aspectos, incluem-se a necessidade de processamento de uma massa de dados extremamente volumosa, gerada ou capturada pelos mais diversos aparatos tecnológicos contemporâneos. Além da volumosa massa de dados, a velocidade e a variedade com que são criados ou capturados, aliados à veracidade dos mesmos, configuram o que se convencionou chamar de *Big Data* (DAVENPORT, 2014; SCHUTT; O'NEIL, 2014; STANTON, 2012). O conceito de *Big Data* é central para a Ciência de Dados, que utiliza um processamento através de modelização matemática por algoritmos a serem aplicados aos dados para obtenção ou extração de informações relevantes ao usuário. Esta modelização se dá pela análise de dados em seus repositórios, que podem ser coleções, armazéns ou bancos de dados e sua representação se dá predominantemente de maneira gráfica, permitindo uma visualização mais intuitiva do potencial informativo dos dados. Este processamento consiste na construção de modelos matemáticos mediante utilização de algoritmos a serem aplicados na interpretação de dados para obtenção de padrões, tendências, vetores, estatísticas e outras correlações, que serão então representados para o usuário, em consonância com o interesse desejado na busca ou consulta. De acordo com Ribeiro (2014), a utilização de modelagens é necessária pelo fato do ser humano ter uma

capacidade limitada de interpretação de volumes massivos de dados, tarefa esta cabível aos modelos matemáticos e programas computacionais.

Para uma melhor compreensão do impacto da superabundância de dados no contexto científico, apresentaremos um breve histórico do processo de aquisição e transmissão do conhecimento ao longo do tempo, bem como as realidades e possibilidades que ora se apresentam. Para Thomas Kuhn (2007), o conhecimento científico se dá através de revoluções, onde teorias sucedem umas às outras consolidando consensos da comunidade científica de determinadas épocas. Em adição, Karl Popper (MILLER, 2010) constatou que teorias vão resistindo a testes, experimentos e observações até seu ponto de saturação, quando então não mais se mostram satisfatórias na explicação de fenômenos e são, assim, substituídas por novas teorias.

Na ótica de Jim Grey (HEY; TANSLEY; TOLLE, 2009), o processo de conhecimento científico passou por revoluções paradigmáticas, muitas vezes acompanhado de progressos técnicos e tecnológicos de determinadas épocas. O primeiro consistia na observação dos fenômenos e sua sistematização pela compreensão racional, sendo então transmitidos oralmente, na maioria das vezes. Este processo de conhecimento, observação-compreensão-sistematização-ensino, perdurou por séculos. O segundo deu-se no início da Revolução Científica (séc XVI-XVIII), estabelecendo uma nova etapa nas práticas de aquisição e compreensão dos fenômenos, através de modelagens e abstrações teóricas visando a generalização. Importa ressaltar que nesta etapa houve uma mudança não somente no processo de compreensão e sistematização teórica do mundo, mas também no processo de comunicação do conhecimento atingido. O terceiro surge em meados dos anos 1940, quando então a Ciência da Computação começou a tomar corpo. No ambiente científico, em particular, a Computação possibilitou a prática de simulações de eventos complexos, quando computadores cada vez mais poderosos permitiram modelagens baseadas em processamento de dados com utilização de equações e algoritmos.

A partir dos anos 2000, como se viu, uma série de fatores contribuiu para uma nova mudança paradigmática: a convergência das TICs, aumentando exponencialmente o poder de processamento e comunicação de computadores

e dispositivos; a ampla utilização da internet; as tecnologias de miniaturização e armazenamento, provocando uma explosão de repositórios e armazéns digitais diversos e possibilitando a virtualização de dados no que se conhece hoje como computação em nuvem; o poder de processamento paralelo e distribuído de dados, aumentando significativamente a capacidade computacional; a massificação de dispositivos diversos de captura de dados, gerando o equivalente a vários *petabytes* ao dia, dentre outros. Na ciência, tais fatores fornecem o cenário para o que Grey denominou de quarto paradigma do conhecimento, consistindo num fazer científico baseado em utilização massiva de dados – a chamada *e-science*. Este novo modelo baseia-se na captura de uma massa colossal de dados proveniente das mais diversas fontes, de maneira ininterrupta e cada vez mais acelerada, trazendo à tona uma série de questões a serem consideradas.

A primeira delas refere-se às considerações dispensadas aos dados. Se no modelo científico tradicional os dados eram coadjuvantes no processo de construção teórica e experimental, muitas vezes expostos de maneira opaca nos registros e publicações, no novo modelo de *e-science* os dados adquirem um protagonismo sem precedentes. Essa condição modifica a própria prática científica, uma vez que os dados passam, assim, a orientar e a condicionar reflexões teóricas. Assim, temos um novo modelo de ciência orientada a dados, onde cientistas efetuam simulações e formulam teorias de acordo com a observação e reconhecimento de padrões e indicações obtidos pelos mesmos.

Uma segunda questão é o processo de captura de dados, sobretudo dados não-estruturados. Sensores e tecnologias diversas já permitem que dados sejam coletados num determinado universo empírico de maneira ininterrupta, massiva e diversa, nos mais diferentes formatos e padrões. Na mineração tradicional os dados são coletados em frequência e quantidade pré-determinadas, ocasionando uma série de limitações em relação ao novo modelo orientado a dados massivos. Essas limitações refletem amostragens esparsas e pontuais, gerando um lapso de tempo de resposta significativo. Em contrapartida, o modelo orientado a dados massivos trabalha com dados inseridos ininterruptamente, coletados no ambiente desejado e sob recortes seletivos específicos, possibilitando interpretações e reflexões em espaço de

tempo cada vez menores. Este fato impacta o tempo de resposta das pesquisas científicas ante um problema apresentado, diminuindo significativamente o lapso apresentado no modelo tradicional. De acordo com Gillam et al. (2014), descobertas científicas na área biomédica que levavam séculos ou décadas para serem incorporadas ao currículo e protocolo médico têm agora levado cada vez menos tempo de incorporação, na razão de apenas alguns anos ou poucos meses.

Uma terceira questão no novo modelo científico é a capacidade de processamento dos dados massivos coletados. Como se viu, a Ciência de Dados apresenta-se como o domínio de tratamento desses dados, utilizando-se do avanço técnico da Ciência da Computação e da Microeletrônica no que se refere ao poder de processamento paralelo e armazenamento. A Ciência de Dados tem utilizado poderosos softwares de tratamento semântico de dados complexos, tanto no âmbito corporativo e comercial quanto no âmbito científico, constituindo o que Hey, Tansley e Tolle (2009) chamam de *sistemas de gerenciamento de informações laboratoriais*. Tais sistemas permitem que no momento da publicação, dados possam ser acessados para reconstituição de experimentos e simulações, ilustrando as teorias propostas e, ao mesmo tempo, fornecendo subsídios para novos testes. Na área biomédica, em particular, softwares como o *Pubmed*, *Molecular Biosystems* e *Reflect* permitem variados recursos, como a adição de camadas de marcação para interoperabilidade, além de outras aplicações para interações semânticas, que podem ser visualizados sobre os dados sem comprometer a integridade dos mesmos.

Como visto, na ciência moderna o registro e a publicação de teorias e experimentos em relatórios, artigos e outros meios é uma atividade essencial ao fazer científico, pois não há ciência sem comunicação. Desta forma, uma nova questão se apresenta como relevante no novo modelo de *e-science*: a curadoria de dados. Segundo Sayão e Sales, a curadoria digital “assegura a sustentabilidade dos dados para o futuro, não deixando, entretanto, de conferir valor imediato a eles para os seus criadores e para os seus usuários” (SAYÃO; SALES, 2012, p.185). A curadoria reflete preocupações próprias da consolidação dos dados como um elemento central no novo modelo científico.

Segundo os autores, a curadoria é um gerenciamento de dados e seus metadados, primando por aspectos como gestão do ciclo de vida e ações sequenciais. Além disso, a curadoria de dados tem sua importância intimamente relacionada a questões como softwares, processamento, armazenamento, interoperabilidade, ontologia e semântica, bem como a segurança de informação e aspectos éticos. Estas questões, assim, revelam-se como relevantes na prática científica orientada a dados massivos.

A área biomédica tem sido apontada como uma das mais desenvolvidas na prática orientada à utilização de dados massivos e aplicações semânticas. Para Ginsparg (2009) outras áreas deveriam espelhar-se nos exemplos pioneiros das Ciências Biológicas e da Vida, acrescentando também que os dados, tanto quanto os textos, devem ser considerados de maneira igualitária e equivalente. Entendendo a comunicação científica como uma operação essencial na ciência, os registros são os elementos-chave nesta prática, na qual os dados – estruturados ou não – surgem como recursos indispensáveis para o encurtamento das distâncias entre as reflexões teóricas e a aplicação prática. Como atestou Lynch, em 2009,

> But now we are also seeing scholars engage the scientific record in the large, as a corpus of text and a collection of interlinked data resources, through the use of a wide range of new computational tools […] we will see this change many aspects of scientific culture and scientific publishing practice, probably including views on open access to the scientific literature, the application of various kinds of markup and the choice of authoring tools for scientific papers, and disciplinary norms about data curation, data sharing, and overall data lifecycle. Further, I believe that in the practice of data-intensive science, one set of data will, over time, figure more prominently, persistently, and ubiquitously in scientific work: the scientific record itself (LYNCH, 2009, p. 182-183).

Nos últimos anos, para os desafios da prática científica que ora se apresenta – a *e-science* – a Ciência de Dados vem formando pessoal versado em habilidades de interpretação e representação de dados e informações afluentes de áreas diversas como computação, ciência da informação, matemática e estatística, dentre outras, cujo objetivo central é a tradução da massa inteligível de dados em informação relevante. Também tem contado

com diversos softwares para processamento semântico de dados, de modo a extrair padrões e relações causais que possam potencializar teorias, simulações e experimentos para uma prática de pesquisa mais completa e abrangente, na qual a replicação e reprodução de experimentos e simulações tornam-se possíveis em tempo real. Da mesma forma, aspectos como a computação em nuvem tem possibilitado não somente espaço praticamente ilimitado para armazenamento de dados, mas também a execução de softwares e serviços online (*Software as a Service*), permitindo ao pesquisador processar dados de qualquer lugar e a qualquer momento. Este tipo de prática permite a transcendência dos ambientes clássicos de laboratórios e, além disso, tem possibilitado uma maior integração não somente entre a comunidade científica, mas também entre esta e o público. Ademais, como se viu, há uma grande possibilidade de encurtamento do tempo entre a pesquisa – elaboração de teorias e execução de experimentos e simulações – e a prática – aplicação real em benefício da sociedade.

3 MINERAÇÃO DE DADOS EM TEXTOS BIOMÉDICOS

Grandes são os interesses na aplicação de mineração de textos (MT) na área de biomédica, devido à existência de grandes recursos informacionais em formato de texto (COLLIER et al., 2006; ERHARDT; SCHNEIDER; BLASCHKE, 2006; ZWEIGENBAUM et al., 2007). Na área das ciências médicas, basicamente, objetiva identificar, distinguir, extrair e relacionar diferentes fatos, a fim de auxiliar os médicos no diagnóstico de doenças, tendo, como base de informações, os prontuários e as fichas de acompanhamento. Nesse contexto específico, dos os prontuários e as fichas de acompanhamento, Collier (2006) afirma que a MT “penetra” mais lentamente por razões que envolvem, principalmente, a privacidade dos dados dos pacientes. De maneira contrária, nas ciências biológicas, a mineração de texto auxilia na identificação de entidades biológicas e no lidar com a complexa nomenclatura existente, principalmente a de genes e proteínas, daí ser utilizada como uma das principais áreas de sua aplicação.

A indexação e recuperação de documentos dependem do sucesso da identificação de termos biomédicos (*Named Entity Recognition* – NER).

Segundo Erhardt, Schneider e Blaschke (2006), essa identificação é um "gargalo", pois a terminologia biomédica muda dinamicamente. Assim, segundo os autores, o processo de identificação é um tópico de interesse para aplicação de técnicas automáticas. Todavia, a estratégia utilizada para lidar com essa dificuldade faz uso de vocabulários controlados, aproximando-se de iniciativas como da *Gene Ontology* e da UMLS[2] (SMITH, 2007).

A MT é também aplicada não só para identificar fatos nos textos, como também para ajudar nos relacionamentos indiretos entre entidades biológicas. Esse procedimento tem sido referenciado como *Literature-based discovery* (COHEN, et al. 2008), um método para gerar automaticamente hipóteses para pesquisa, detectando conexões implícitas, negligenciadas na pesquisa da literatura. "[..] consiste em descobrir, ligações indiretas e "escondidas": isto é denominada frequentemente "descoberta baseada em literatura". Estas ligações podem ser propostas como hipóteses científicas potenciais" (ZWEIGENBAUM et al., 2007, p. 368).

As descobertas são geradas sob a forma de relações entre dois conceitos primários, como, por exemplo, uma droga como um tratamento para uma doença ou um gene como a causa de uma doença.

3.1 Abordagens para o processamento de textos

De uma maneira simplificada, o processamento de textos pode ser abordado sob duas perspectivas: a Modelagem Matemática, usando de forma expressiva a abordagem estatística e a do PLN que focam sobre a estrutura sintática implícita da linguagem natural nos textos.

A abordagem estatística busca representar o documento através de seus termos, que são valorados através de sua frequência e peso, eles não são considerados de forma abrangente no contexto onde aparecem. Tem como característica fundamental estimar probabilidades, para auxiliar a tomada de

[2] A UMLS – Unifyed Medical Language System –, é uma grande e amplamente usada base terminológica no domínio da biomedicina.

decisão (ARANHA, 2007), aproximando-se das técnicas de mineração de dados.

Na perspectiva do PLN, ou da Linguística Computacional, busca-se o significado das palavras através dos aspectos morfológicos, sintáticos, semânticos, pragmáticos e do contexto em geral.

3.1.1 Abordagem estatística

Essa abordagem busca representar o documento através de uma descrição numérica, utilizando os termos contidos no texto que são valorados através de sua frequência e peso, isto é, os termos descrevem o conteúdo de um texto e essa característica é capturada, atribuindo para cada termo um peso que age como um indicador da sua importância na descrição do conteúdo.

Trabalhos iniciados entre as décadas de 40 e 50, como os de Shannon (1948), Zipf (1949) e Luhn (1958) são marcos iniciais dessa abordagem e uma variedade de técnicas foram desenvolvidas, ao longo dos anos para determinar a importância dos termos, tais como: Algoritmos Genéticos (ROBERTSON; WILLETT, 1996), uso de conceitos para pesagem (KEEN, 1991), Neural Networks (BOGER et al., 2001), Modelos Probabilísticos (GREIFF; MORGAN; PONTE, 2002; MELUCCI, 1998; PONTE; CROFT, 1998), Espaço Vetorial (SALTON; YANG; YU, 1975), Modelos Algébricos, *Latent Semantic Indexing Model*, (GORDON; DUMAIS, 1998), entre outros.

3.1.2 Abordagem linguística

A linguística Computacional se desenvolve com a perspectiva teórica da gramática gerativa de Chomsky (1957). Chomsky definiu uma linguagem como um conjunto infinito de sentenças gramaticais e, de acordo com ele, uma gramática é um jogo finito de regras, gerando infinitas sentenças gramaticais.

O autor propõe entender gramáticas como dispositivos que reúnem pedaços de orações, de acordo com regras precisas, “gerando” orações bem formadas.

Segundo Chomsky (1957), a sintaxe é o estudo dos princípios e processos que presidem a construção de frases em línguas particulares. O estudo sintático de uma determinada língua tem como objetivo a construção de uma gramática, que pode ser encarada como um mecanismo de produção de frases da língua em questão. Sobre esse viés, uma análise linguística automática visa a operar sobre as sentenças do texto, uma vez que se apresenta como um enunciado dotado de expressão completa de sentido. Adotar a sentença como uma unidade de análise é reconhecer que cada uma de suas partes constituintes foi gerada a partir da gramática da linguagem.

Analisar a sentença é, então, analisar suas partes menores, seus constituintes que se organizam de maneira hierárquica, representada por uma estrutura arbórea, a estrutura interna da sentença. Nos níveis intermediários estão os elementos sintáticos formativos da sentença (sintagmas) e na base da estrutura, os itens lexicais correspondentes (as palavras).

O sintagma consiste num conjunto de elementos que constituem uma unidade significativa dentro da oração e que mantém entre si relações de dependência e de ordem. Os sintagmas se organizam em torno de um elemento fundamental, o núcleo, que pode, por si só, constituir o sintagma. Um núcleo pode ser um elemento nominal (nome ou pronome), definindo um sintagma nominal, um elemento verbal, definindo um sintagma verbal. A regra definida para a sentença (Sentence → NP + VP) aparece como constituintes obrigatórios, o sintagma nominal e verbal. Porém, outros tipos de sintagmas podem utilizados na oração, como: preposicional, adverbial, etc.

3.2 Infraestrutura para PLN

Diversas ferramentas de código aberto que auxiliam o PLN estão disponíveis e acessíveis na *web*, como as do *Stanford Natural Language Processing and Computational Linguistics Group*, *Unified Medical Language System* (*UMLS*), *LingPipe*, *MorphAdorner* da *Northwestern Universit.* A adoção de uma infraestrutura para o PLN permitirá um ganho de qualidade no trabalho, pois permitirá explorar aquilo de melhor que cada uma tem.

- SPECIALIST NLP Tools

A ferramenta *SPECIALIST NLP Tool* é desenvolvida pelo *The Lexical Systems Group* do *The Lister Hill National Centre for Biomedical Communications* cujo grupo objetiva investigar as contribuições que as técnicas de PLN podem trazer às tarefas de mediação entre a linguagem dos usuários (linguagem natural) e a linguagem para recursos de informação biomédicos on-line (linguagem de consulta, *query*) (UMLS, 2008).

- LingPipe

LingPipe (2009) é um conjunto de bibliotecas de Java para PLN, de código aberto. Bastante rápido e seguro, com diversas funções disponíveis, sendo possível reconhecer entidades (pessoas, proteínas, etc.), classificar textos, corrigir ortografia, quebrar em sentenças, identificar idioma, entre outros.

- WordNet

WordNet (PRINCETON UNIVERSITY, 2008) é um dicionário semântico para a língua inglesa, desenvolvido na Universidade de *Princeton*, contendo elementos tais como substantivos, verbos, adjetivos e advérbios. O *Wordnet* forma uma rede de conceitos, em que cada um corresponde a um conjunto de palavras que são sinônimos entre si (*synsets*). Através dos *sysnsets*, são codificados os significados para cada palavra. Os conceitos estão organizados de acordo com relações semânticas, que podem ser: Sinonímia, Antonímia, Hiponímia/hiperonímia e Meronímia/Holonímia, troponímia. Existem ainda relações morfológicas representadas por plural, feminino, conjugação de um verbo e associações a conceitos de acordo com o lema (FELLBAUM, 1998).

3.3 Extração de Informação

Diferente das informações armazenadas em bancos de dados que são altamente estruturadas, estando disponíveis explicitamente e podendo ser obtidas através de consultas SQL, nos textos, a informação pode estar implícita, escondida e difícil de ser encontrada. Devido a essa natureza não estruturada (ou semiestruturada) do texto, são necessários mecanismos computacionais diferentes dos utilizados para os bancos de dados. A Extração de Informação (*Information Extraction*) é um mecanismo utilizado para lidar com este tipo de dado (JONES; WILLETT, 1997). O objetivo da Extração de Informação (EI) é classicamente associado à extração de específicos tipos de informação de um documento (RILOFF; LEHNERT, 1994).

Diferente da Recuperação da Informação (RI) que tem por objetivo encontrar textos e documentos relevantes, de acordo com a consulta do usuário, a EI trata de solucionar o problema de achar informações dentro dos textos. Difere, também, da PLN porque é mais específico, visando a extrair determinados tipos de informação (obter informação pré-especificada), geralmente direcionada para extrair características do domínio (termos, objetos, entidades, relações) no qual o texto está inserindo. A EI é diferente, ainda, da Extração de Conhecimento (Descoberta de Conhecimento - *Knowledge Discovery in Databases* – KDD), porque não visa a deduzir regras.

A área se estabelece como uma coleção de diferentes problemas, tais como, nomeação de entidades ou *named entity recognition* (NER), extração de relação (ER) e extração de eventos (EE). A NER trata do problema de identificação e nomeação de entidades (pessoas, organizações, lugares, doenças, genes, proteínas, etc.) no texto; a ER trata da identificação de associações entre entidades; e, finalmente, a EE trata de identificar atividades ou ocorrências, tais como: fusões, aquisições, desastres e atividades terroristas. Somam-se ainda, os problemas relacionados com *co-reference resolution* (resolução de correferência) na qual se buscam resolver referências

anafóricas[3] e os problemas relacionados com a identificação de expressões que se referem a uma mesma entidade. A extração de palavras do texto para a construção de vocabulários controlados, taxonomias, indexação derivativa e ontologias é um emprego conhecido da EI.

Embora a definição dada anteriormente por Riloff e Lorenzen (1999) seja utilizada largamente, atualmente, o objetivo da EI não está voltado necessariamente para um domínio específico, "um sistema de extração de informação ideal deveria ser independente do domínio ou pelo menos portátil a qualquer domínio com uma quantidade mínima de esforço de engenharia." (MOENS, 2006, p.3).

3.4 Visão Geral dos Métodos de Extração de Informação

Segundo Sarawagi (2008), os métodos usados para extração de informação podem ser reunidos em dois grupos: *hand-coded* ou baseados em aprendizagem e baseados em regras ou estatísticos.

- Hand-coded

Os métodos baseados em *hand-coded* necessitam de especialistas em domínio, para definir regras ou expressões regulares e *snippets program*. Expressões regulares são escritas numa linguagem formal que pode ser interpretada por um processador de expressão regular que examina um texto, identificando partes que casam com a especificação dada (EXPRESSÃO REGULAR, 2016). Por exemplo: ^([a-z0-9_]|\\-|\\.)+@(([az0-9_]|\\-)+\\.)+[a-z]{2,4}$ é uma expressão regular que reconhece um endereço de e-mail, onde cada símbolo possui uma função na expressão regular.

Um exemplo da aplicação dessa estratégia é o sistema DRIPE (*Dual Iterative Pattern Expansion*) (BRIN, 1999) que tem por objetivo extrair uma tupla, formada pelo nome de uma organização e a sua localização, de uma determinada coleção de documentos.

[3] Anáfora é a repetição da mesma palavra ou grupo de palavras no princípio de frases ou versos consecutivos.

- Métodos baseados em aprendizagem

Permitem ao computador aperfeiçoar seu desempenho em alguma tarefa. Requerem que textos sejam manualmente rotulados (etiquetados), para servirem de base de treino para os modelos de extração; nesse caso, é necessária a participação de especialistas no domínio, para que possa identificar as características importantes no texto de forma correta (SARAWAGI, 2008). É comum utilizar bases de treinos disponíveis na internet, tais como, BioInfer *corpus* da *University of Turku* (PYYSALO et al., 2007), Genia *corpus* do *Tsujii Laboratory* da *University of Tokyo*, anotado como um subconjunto de substâncias e os locais biológicos envolvidos em reações entre proteínas; as anotações têm como base a GENIA *ontology*[4] e BioScope *corpus* (BIOINFER..., 2006) que consiste de textos médicos e biológicos anotados com indicações de termos voltados para a negação e a especulação (SZARVAS et al., 2008).

- Métodos Baseados Em Regras ou Padrões Léxico-Sintáticos

Os métodos baseados em regras registram os "comportamentos" ou regras de um determinado domínio para interpretar a informação que está sendo processada, isto é, elementos textuais obtidos são comparados com as regras armazenadas em um conjunto que definem os comportamentos. Estas regras podem ser criadas manualmente ou através de mecanismos estatísticos automáticos e representadas, por exemplo, através de gramáticas ou diagramas de estados. O uso de padrões e regras é encontrado em diversos trabalhos, MindNet (RICHARDSON; DOLAN; VANDERWENDE, 1998), Snowball (AGICHTEIN; GRAVANO, 2000), KnowItNow (CAFARELLA et al., 2005).

- Métodos baseados em estatísticas

Utilizam a noção de pesagem e frequência, buscando estabelecer o grau de relevância de determinado elemento no texto. Collier, Nobata e Tsujii (2000)

[4] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=GENIA+Project

treinaram um HMM com bigramas baseado no léxico e em características sintáticas utilizando um *corpus* relativamente pequeno de 100 *abstracts* do MEDLINE com anotações, realizada por especialistas do domínio, para identificar nomes de proteínas e produtos gênicos. O sistema obteve 0,73 de *f-score*.

Sistemas de EI podem ser baseados em técnicas de PLN que visam a aumentar o grau de entendimento do texto, a fim de melhorar a precisão da extração dos dados relevantes. Basicamente, um sistema de EI baseado em PLN utiliza as seguintes técnicas/fases da PLN: um "tokenizador" que tem como objetivo a identificação de todos os termos no texto, em que uma primeira etapa; a segunda etapa é análise léxica e/ou morfológica, onde cada termo é classificado morfologicamente (substantivo, verbo, artigo, etc.), podendo ter outras características (feminino, plural, o reconhecimento de nomes próprios, data, hora e outros itens que podem ter uma estrutura interna, como dosagem de remédio, pressão arterial, etc.).

A fase seguinte utiliza um analisador sintático/semântico, em que são definidos os padrões de extração, consistindo na criação de um conjunto de regras de extração, específico para o domínio. Geralmente, os padrões se baseiam em restrições sintáticas e semânticas, aplicadas aos constituintes das sentenças (MUSLEA, 1999). Gramáticas regulares, *parsers* (analisadores de estruturas gramaticais) e autômatos finitos podem ser utilizados para o reconhecimento das informações, através de análise léxico sintático (LOH, 2001). Após superar a etapa de análise sintática/semântica, é efetuada a "análise do discurso" que tem como objetivo relacionar diferentes elementos do texto, tais como, reconhecer e interpretar apostos e outros grupos nominais complexos, resolução de co-referência, relacionamento entre as partes do texto, descrevendo, por exemplo, uma rede associativa.

A ER também tem contribuído para a descoberta de relações, através do reconhecimento de padrões que podem ser automaticamente explorados. Um exemplo é a extração de coocorrências de palavras (proximidade de palavras) que podem indicar um relacionamento léxico. Como cada palavra pode denotar um conceito, tal relacionamento poderia indicar uma relação semântica.

Uma técnica para extrair coocorrências é a janela textual (*textual windows*), isto é, são selecionados grupos de palavras (3, 4 ou 5 palavras consecutivas) que são analisadas com relação ao fato de ocorrerem juntas. Fazendo uso de dicionários, vocabulários controlados, taxonomias e ontologias, pode-se inferir a existência ou não da relação. O problema da técnica da janela textual é que toda palavra é potencialmente relacionada à outra palavra e a distância entre elas pode variar. Outro problema é que a técnica baseia-se nos agrupamentos de palavras que ocorrem certo número de vezes.

3.5 Extração de Relação

A extração de relações é a tarefa de procurar relacionamentos entre duas entidades no contexto textual. Segundo Chen et al. (2005), o conceito foi introduzido em MUC-6 (*Sixth Message Understanding Conferences*), a sexta conferência de uma série envolvidas na avaliação de sistemas de extração de informação e na definição de tarefas comuns, patrocinadas pela ARPA (*Advanced Research Project Agency)* para medir o progresso na área de EI.

A extração de relações (RE) é motivada pela necessidade de obter informações, como, manter-se atualizado sobre informações do dia a dia de companhias, como informado em jornais (por exemplo, Pessoa X trabalha para Organização Y, organização Y possui Organização Z) ou mantendo informações sobre proteínas e interações de genes obtidas através da literatura científica, por exemplo, Gene X codifica Proteína Y, Proteína Y liga Proteína Z. (HACHEY, 2009, p. 9).

Os trabalhos de Swanson (1986, 1990), Sawanson e Smalheiser (1997), Swanson, Smalheiser e Torvik (2006) são seminais no campo da descoberta de associações entre entidades. A proposta de Swanson se mostrou fértil e teve outros seguidores (WEEBER, et al., 2003). Swanson explora a ideia de conhecimento público não descoberto. Através da Análise Bibliográfica, ele procura ligações entre informações já publicadas cuja associação ainda não é conhecida pelos pesquisadores. Um exemplo é a identificação de uma relação entre dois *corpus*: um sobre distúrbio do sistema circulatório - a doença de

Raynaud e o outro sobre o uso de óleo de peixe para melhorar a circulação sanguínea. Nenhum pesquisador ainda tinha usado óleo de peixe para tratar a doença de Raynaud. Swanson sugere que há muitos outros fragmentos desconectados de conhecimento na literatura para a qual a análise poderia fazer conexões.

- Métodos de Extração de Relação

Conrad e Utt (1994) desenvolveram um método para, automaticamente, extrair características de textos e identificar, através de estatística, os relacionamentos e associações entre essas características. Inicialmente, identificam-se informações específicas (nomes de pessoas, nomes de companhias formando *joint venture*, nomes de novas companhias, localização da nova companhia, os produtos da nova companhia e o capital financeiro envolvido). O próximo passo é identificar o relacionamento ou associação entre eles. As associações ocorrem, quando as palavras estão próximas umas das outras. Os autores propõem uma janela textual (*text windows*) de tamanho 51 e 201 palavras à direita ou à esquerda de uma determinada menção de entidade (nome da pessoa ou empresa). Os tamanhos, 51 e 201 palavras, são escolhidos empiricamente e próximos do tamanho médio de parágrafos e documentos, respectivamente. Em seguida, utilizam duas medidas estatísticas de associação para classificar ("rankear") os pares de entidade, são elas; *Pointwise Mutual Information* (PMI) que compara a probabilidade de observar duas palavras junto com a probabilidade de observá-las independentemente e o *phi-squared* (ϕ^2) um teste estatístico para análise de desvio da expectativa em dados.

Outro trabalho, no campo da extração de relações, é o de Jenssen et al. (2001) que desenvolveram a ferramenta PubGene, identificadora de expressões e agrupamentos de genes em um texto, relacionando-os com a literatura existente. O texto (título e *abstract*) é pesquisado, buscando identificar a coocorrência de pares de genes. O reconhecimento dos nomes é realizado, através de um processo de busca em dicionários e bases de dados, tais como, HUGO (2010), LocusLink (NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION, 2009), Genome Database (2009) e GENATLAS (PARIS,

2010). A partir da frequência das coocorrências é criado um mapeamento para índices do MEDLINE. PubGene compila agrupamentos de textos e tenta "casar" os pares de genes já descritos nos dicionários, buscando descobrir na literatura relações que podem ser formuladas como hipóteses para pesquisa.

Através dos artigos de revisão (COHEN; HERSH, 2005; YEH et al., 2003; ZWEIGENBAUM et al., 2007) sobre a pesquisa em descoberta de conhecimento em texto, especificamente em biomedicina, percebe-se que o campo concentra três abordagens:

- **Contexto linguístico do texto**, identificação de nomes de entidades, tais como: proteínas, doenças, gene, etc. e a identificação da ocorrência de elementos sintáticos entre os nomes de entidades;
- **Casamento de padrão**, modelos que "casam" estruturas linguísticas específicas para reconhecer e extrair informação de interação entre entidades;
- **Coocorrência de palavras** que extrai coocorrências de entidades e usa-as para predizer as conexões baseado em estatísticas.
- **Problemas Identificados nos Métodos de Extração**
  As estratégias desenvolvidas para a tarefa de extração de relação estão baseadas, de uma maneira geral, na criação de regras ou de modelos de aprendizagem de máquina e ambos são difíceis de suportar novos domínios.

Na abordagem da criação de regras, escrevê-las requer o esforço de um especialista familiarizado com o domínio em que elas se inserem. Representar todo conhecimento humano disponível sobre um domínio, em computador, de forma legível, não é, até este momento, uma realidade (ERHARDT; SCHNEIDER; BLASCHKE, 2006). Assim, é comum restringir o contexto de atuação da extração estabelecendo um conjunto específico de relações a ser extraído, como em Hearst (1998), MindNet (RICHARDSON; DOLAN; VANDERWENDE, 1998), Snowball (AGICHTEIN; GRAVANO, 2000), KnowItNow (CAFARELLA et al., 2005; MARTIN, 1995; MARKOWITZ et al.,

1992), BITOLA (BIOMEDICAL DISCOVERY SUPPORT SYSTEM, 2010), GENESCENE, (PUSTEJOVSKY et al., 2002); ou restringindo as entidades sobre as quais se buscam as relações , como em Conrad e Utt (1994), Filatova e Hatzivassiloglou (2003), Hasegawa, Sekine e Grishman (2004), Chen et al. (2005), Hachey (2009), Fukuda et al. (1998); ou ainda, restringindo o domínio de aplicação, como em Webfoot (SODERLAND, 1997), um preprocessor que analisa gramaticalmente páginas na *web*, agrupando em segmentos logicamente coerentes baseado no layout da página e Popescu e Etzioni (2005) que analisa opiniões sobre produtos, relatados em *newsgroup* (*opinion mining*).

Essas restrições não se configuram apenas como característica da abordagem de regras; elas também são encontradas em sistemas que utilizam modelos de aprendizagem de máquina. De modo geral, modelos baseados em regras possuem boa precisão e baixa cobertura.

No caso da aprendizagem de máquina, o especialista, também, é necessário para efetuar a anotação nos textos (*hand-tagged*) que servem como base de treino, ajustando os parâmetros para o modelo (HACHEY, 2009, p.3). Esses pré-requisitos são limitações bem conhecidas (AGICHTEIN; GRAVANO, 2000; BRIAN, 1998; SUDO; SEKINE; GRISHMAN, 2003). Outra abordagem é a utilização de *corpus* que estão disponíveis na *web*.

Corpus auxilia na produção de estatísticas sobre as informações contidas e, a partir delas, programas podem tomar decisões sobre um dado aspecto. A máquina “aprende” a linguagem, através da regularidade estatística no *corpus*. Os termos e frases que representam entidades biológicas são extraídos do texto e análise estatística das coocorrências no *corpus* é usada para estabelecer associações entre os termos.

A alta dependência do *corpus* para aprendizagem de máquina é clara e isto pode trazer alguns problemas: eles são projetados, em função do objetivo da pesquisa e as anotações são construídas com fim específico[5]; sua

[5] O PennBioIE (http://bioie.ldc.upenn.edu/publications/latest_release/) corpus contém, no total, 2257 abstracts do Medline em dois domínios: genética molecular em câncer com 1157

construção necessita de especialistas e de tempo para produzir as anotações; é comum o reconhecimento de um conjunto específico de relações entre termos de acordo com relações anotadas; a precisão da estatística depende do tamanho do *corpus* e o texto a ser processado deve ter características próximas a dos textos anotados.

Uma característica comum em vários sistemas, principalmente na área de biomedicina tem sido a de se valer da estrutura do documento, concentrando em certas seções (por exemplo, *Abstract*, Resultados), como meio de limitar a procura de características ou padrões (ERHARDT; SCHNEIDER; BLASCHKE, 2006; LIDDY, 2000; MEDELYAN; WITTEN, 2008; RACUNAS et al., 2004; ROSARIO; HEARST, 2001; YEH; HIRSCHMAN, MORGAN, 2003). É comum a utilização dos abstracts do *Medline* como fontes para a extração de relação, porém isso pode trazer problemas, pois sendo o texto contido no *abstract* limitado, a informação disponível pode ser restrita, dependendo do objetivo desejado.

Dessa forma, a identificação e extração das relações no texto dependem de diferentes fatores, como, as relações a serem identificadas, a formatação e o tipo do texto, o assunto do texto, o domínio ou área, a abordagem utilizada.

4 PUBLICAÇÕES ESTENDIDAS E MODELOS SEMÂNTICOS DE PUBLICAÇÕES

Esta seção trata de pesquisas, projetos e propostas de modelos de publicações científicas digitais que objetivam superar o atual modelo de publicações científicas textuais e lineares, herdado do período das publicações impressas.

O linguista americano Noan Chomsky (1957), influenciado pelo Estruturalismo, foi um dos pioneiros que cogitou que a linguagem humana

abstracts e inibição de enzimas da classe CYP450 com 1100 abstracts. Todos os abstracts são anotados manualmente indicando parágrafos, orações, classes de palavras, e entidades biomédicas para uso específico em cada domínio. Já o O Genia Corpus contém 1999 abstracts do *Medline* selecionados a partir dos termos: *human, blood cells, e transcription factors do Mesh*.

poderia ter uma estrutura e, mais ainda, uma estrutura profunda – semântica -, comum a todas as línguas, na qual as estruturas ditas superficiais das diferentes línguas seriam manifestações. Seu trabalho influenciou a maioria das pesquisas posteriores que analisavam a estrutura de textos em linguagem natural.

Kintsh e Van Dijk (1972) analisaram a estrutura de diferentes gêneros de textos, identificando níveis de análise que chamaram de microestrutura, macroestrutura e superestrutura; associaram estes elementos às intenções retóricas dos diferentes gêneros de textos. Um artigo científico é um discurso retórico (GROSS, 1990; BAZERMAN, 1988; HUTCHINS, 1977; SKELTON, 1994; NWOGU, 1997; DE WAARD, 2006) com um objetivo preciso de convencer o leitor da acuidade os métodos empregados em um determinado experimento e, consequentemente, na veracidade, necessidade e universalidade das afirmações feitas na sua conclusão. Se este objetivo é atingido, ao longo da "vida" do artigo – o período desde o qual ele avaliado por pares, aceito para publicação, publicado e, eventualmente citado, o conhecimento contido no artigo (em especial, na sua conclusão) passa a ser incorporado ao acervo de conhecimentos de determinada área. Pode então servir de base e fundamento para novos conhecimentos e hipóteses. Na construção prática deste discurso, na escrita do artigo, cada seção tem uma função, definida a partir do Método Científico e de sua função retórica, como pode ser visto também na citação do The International Committee of Medical Journals Editors[6].

> The text of observational and experimental articles is usually (but not necessarily) divided into sections with the headings Introduction, Methods, Results, and Discussion. This so-called "IMRAD" structure is not simply an arbitrary publication format, but rather a direct reflection of the process of scientific discovery.

[6] The International Committee of Medical Journals Editor, http://www.icmje.org/recommendations/ browse/manuscript-preparation/preparing-for-submission.html.

Gardin (2001) foi um pioneiro na compreensão de que textos de artigos científicos possuíam uma estrutura semântica profunda, para além da estrutura padronizada de seções. Sua proposta de escrita logicista propunha adotar esta estrutura profunda, lógica, de modo a tornar o conteúdo dos artigos científicos processável por computadores, de modo que o artigo pudesse ser consultado como uma base de dados.

Com o objetivo de identificar os elementos lógico-semânticos contidos na estrutura linguística dos artigos científicos estão os trabalhos de Kando (1997), (1999). Os trabalhos de Murray-Rust e Rzepa (1999, 2002), e de Hucka et al. (2003) propõe marcar o texto dos artigos científicos com base na linguagem XML que surgia então. Outro importante experimento nesta direção é a TEI – Text Encoding Initiative, que usa XML para marcação de textos acadêmicos em literatura e linguística com o objetivo de facilitar a recuperação e a preservação de publicações eletrônicas. A DDI - Data Documentation Initiative (2004) – objetiva estabelecer um padrão internacional em baseado em XML para o conteúdo, preservação, transporte e preservação de documentação em bases de dados em ciências sociais. Estes trabalhos, juntamente com o uso crescente de resumos estruturados, apontam para uma padronização da estrutura semântica dos artigos científicos.

Já mais recentemente, sob a influência crescente da Web Semântica e de suas tecnologias, novas possibilidades surgem para as publicações científicas. As tecnologias da Web Semântica (BERNEERS-LEE, 2001) propõem um passo adiante para a questão da recuperação e processamento semânticos de conteúdos em ambientes computacionais. Segundo esta proposta a descrição do conteúdo de um documento na Web não é mais uma questão de combinar palavras-chave, como em ambientes computacionais convencionais desde os anos 60, mas consiste em conjuntos estruturados de conceitos ligados por relações de significado preciso, dado por padrões como

em RDF7 e RDF Schema8. Construídas com base no RDF Schema, ontologias computacionais, codificadas na linguagem OWL9, organizam o conhecimento em domínios específicos, registrando conceitos acordados por comunidades, organizados em hierarquia de classes e subclasses, em propriedades desses conceitos, em relações entre eles e em regras lógicas para aplicá-los a esse domínio. Esse rico esquema de representação "semântica" permite que agentes de software executem "inferências" e tarefas sofisticadas com base no conteúdo de documentos.

Entre os projetos inovadores que avançam na direção de tirar partido do novo ambiente proporcionado pela Web estão os seguintes.

Communications in Physics (2001), da Universidade de Amsterdam, objetiva "[…] propose a new modularity to Physics publications that will reflect the possibilities of electronic tools." The Scholarly Ontologies Project (2004), desenvolvido na Open University, Reino Unido, se propõe a "build and deploy a prototype infrastructure for making scholarly claims about the significance of research documents" (http://projects.kmi.open.ac.uk/scholonto/summary.html). O trabalho de De Waard et al. (2006), que propõe modelar os elementos retóricos em textos científicos, de modo a explicitar e resgatar estes elementos. HyBrow (RACUNAS et al., 2004), um sistema que objetiva auxiliar cientistas na formulação e avaliação de hipóteses com relação ao conhecimento prévio. O trabalho de Hunter et al. (2008), que objetiva identificar conceitos para extração de relações de interação entre proteínas em textos biomédicos. O projeto MachineProse (DINAKARPADIAN et al., 2006), propõe formalizar, com base numa ontologia de relações, as afirmações feitas em artigos científicos, codificando-as em formato "inteligível" por programas. O sistema Information Hyperlinked Over Proteins (iHOP), desenvolvido por Hoffmann e Valencia (2005), utiliza genes e proteínas como hiperlinks entre as frases e resumos extraídos da base de dados PubMed. As sentenças são mostradas dentro do resumo do qual foram extraídas e relacionadas à referência bibliográfica do

[7] RDF Resource Description Framework. http://www.w3.org/RDF/
[8] RDF Schema Specification, http://www.w3.org/TR/2000/CR-rdf-schema-20000327/.
[9] OWL – Ontology Web Language, http://www.w3.org/2001/sw/wiki/OWL.

artigo correspondente, preservando, assim, o contexto completo da sentença. Textpresso, sistema de mineração de texto para literatura científica que utiliza uma ontologia baseada em categorias e subcategorias de termos, que são classes de conceitos biológicos, para recuperar documentos científicos (MULLER; KENNY; STERNBERG, 2004). É um exemplo do uso e integração de ontologias biomédicas na formatação e recuperação de artigos científicos. De Waard et al. (2009) propões formalizar conhecimento científico através de relações entre hipóteses e evidencias.

Além desses, podem ser mencionados ainda: Research in Semantic Scholarly Publishing (2005), da Biblioteca da Universidade Erasmus de Rotterdam; o projeto Writing in the Context of Knowledge" (CARR et al., 2004), do Laboratório de Inteligência, Agentes e Multimídia da Universidade de Southampton, Reino Unido; a Ontologia para a autopublicação de experimentos da Scientific Publishing Task Force (2006) (https://www.w3.org/wiki/HCLS/ScientificPublishingTaskForce); o projeto SWAN (GAO et al. 2006); **o projeto** ArkeoteK (GARDIN, 2001); a EXPO – uma ontologia para experimentos científicos (SOLDATOVA; KING, 2006).

Sendo a comunicação científica uma fase decisiva em qualquer pesquisa, foi desenvolvido o projeto OBI (2008) – Ontology for Biomedic Investigations - uma ontologia mantida pela OBO Foundry (2010) que tem como objetivo ser uma referência de alto nível relativa à pesquisa biológica. Uma evolução e especialização da OBI é a IAO – Information Artifact Ontology (http://obofoundry.org/ontology/iao.html), com foco especificamente nas chamadas entidades informacionais, as presentes nos processos semióticos envolvendo representação de entidades do mundo.

Outros exemplos com foco nas publicações científicas são os seguintes. O projeto Prospect[10], da Royal Society of Chemistry, no qual termos no texto de artigos que se referem a entidades químicas ou biológicas possuem "links" para ontologias ou dicionários que as definem. O grupo editorial Elsevier desenvolve

[10] Disponível em http://www.rsc.org/Publishing/Journals/ProjectProspect/.

o projeto Article of the Future[11] em cima do periódico biomédico Cell, com o objetivo de adicionar diversas funcionalidades aos artigos, incluindo mudança de forma de apresentação – apresentações hierárquicas -, resumos gráficos, uma seção "Highlights", onde são destacadas de forma sucinta as conclusões do artigo, etc., facilidades estas que só são possíveis num ambiente Web de artigos digitais. Na página do projeto estão disponíveis dois artigos experimentais, que ilustram as facilidades implementadas pelo projeto. Shotton et al. (2009) descrevem a experiência de uso de diferentes tecnologias semânticas na publicação[12] PLoS – Public Library of Science - incluindo ontologias biomédicas, comentários nos artigos e uma ontologia de tipos ou motivos para citações. Um número crescente de publicações científicas, em especial na área biomédica[13], como o BMJ – British[14] Medical Journal, JAMA – Journal of American Medical Association, entre outros, vem usando resumos estruturados (GUIMARÃES, 2006) como forma de otimizar a apreensão do conteúdos dos artigos.

A pesquisa em Ciência da Informação em linguagens de indexação, indexação coordenada e recuperação da informação, dá especial atenção às relações como elementos-chave para representar significados. A proposta de Indexação Relacional de Farradane (1980) estabelece que "Meaning, [is] considered as relations between terms." De acordo com Brookes (1980), "knowledge is a structure of concepts linked by their relations and information is a small part of such a structure." Sheth (2003) afirma que "Relationships are fundamental to semantics – to associate meaning to words, items and entities. They are a key to new insights. Knowledge discovery is about discovery of new relationships". Segundo Miller (1947) o conhecimento científico consiste no estabelecimento de novas relações entre fenômenos. Um fenômeno pode ser definido como um "perceptible fact, a sensible occurrence" (BUNGE, 1998, p. 173).

[11] Disponível em http://beta.cell.com/.
[12] Disponível em http://www.plos.org/.
[13] Disponível em http://www.bmj.com/.
[14] Disponível em http://jama.ama-assn.org/.

Nossa alternativa de pesquisa para viabilizar a consulta, descoberta, reuso e gestão do conhecimento contido no texto de artigos biomédicos por programas foi propor um modelo semântico para o artigo científico, baseado nas tecnologias da Web Semântica e dados abertos interligados, em especial, RDF. Grande parte dos textos citados nesta revisão foram visitados por nossa pesquisa.

Relações são o elemento essencial do esquema de representação do de unidades de conhecimento no modelo proposto; são expressas por três elementos: dois *relata* e um *tipo de relação*. Os dois *relata* – chamados de Antecedente e Consequente - podem ser: dois fenômenos distintos ou um fenômeno e alguma de suas características. O *tipo de relação* guarda a semântica da relação, por exemplo, causa-efeito, sintoma-doença, método-o que é viabilizado pelo método, droga-efeito, etc. As afirmações feitas pelo autor no artigo são representadas como Antecedente-Tipo de Relação-Consequente. O modelo formaliza os elementos semânticos do raciocínio científico – Questões, Hipóteses e Conclusões -, representa-os sob a forma de relações. Por exemplo:

- Papiloma Vírus Humano (Antecedente, um fenômeno) causa (tipo de relação) Câncer de Colo do Útero (Consequente, outro fenômeno);
- Encurtamento dos telômeros (Antecedente, um fenômeno) esta associado a (tipo de relação) senescência celular (Consequente, outro fenômeno).
- Extremidade dos telômeros (Antecedente, um fenômeno) tem como composição molecular (tipo de relação) ‘TTGGG’ (Consequente, uma característica do fenômeno expresso pelo Antecedente).

Destes elementos semânticos o modelo dá ênfase especial às conclusões como unidade de conhecimento essencial que sintetiza a contribuição do artigo ao conhecimento de determinada área científica. De um modo geral também, as conclusões são facilmente identificáveis no texto dos artigos: “The conclusion sections of biomedical abstracts seem like a gold-mine

for automated key assertion identification, since the relevant portion of text can be identified easily" (SAMWALD, 2009).

A representação de unidades de conhecimento na forma de tríades Antecedente-Tipo-de-relação-Consequente torna-as suscetíveis de serem representadas em RDF, podendo ser armazenadas em bancos de triplas e consultadas através da linguagem SPARQL[15].

Com base neste modelo um protótipo (COSTA, 2010) foi desenvolvido, que implementa parcialmente o modelo. Este protótipo usa o processamento de linguagem natural em uma parte curta, mas considerada essencial do texto do artigo, sua conclusão, para formata-la como uma relação. Além de formatá-la, o protótipo procura mapear (identificar) os termos de cada elemento da relação em termos da UMLS Semantic Network, marcando também os que não foram mapeados. A UMLS Semantic Network é o esquema de classificação do UMLS Metathesaurus, organiza os conceitos em árvores hierárquicas, cada uma tendo como raiz um Tipo Semântico (Semantic Types). A UMLS SN usa 54 Tipos de Relação (Relation Types) para exprimir relações semânticas entre conceitos nas hierarquias de Tipos Semânticos e determinar quais relações são permitidas entre os Tipos Semânticos.

A UMLS vem caminhando na direção de se tornar uma ontologia formal, na qual termos biomédicos estão relacionados por relações formais, de semântica precisa e definida (BODENREIDER, 2008). Em sua documentação pode-se encontrar a seguinte afirmação: *"*The purpose of NLM's Unified Medical Language System (UMLS®) is to facilitate the development of computer systems that behave as if they "understand" the meaning of the language of biomedicine and health*".*[16]

Em outro desdobramento da nossa pesquisa, esta representação é utilizada como subsídio para a identificação de novas descobertas. A hipótese é que, se os elementos da conclusão de um artigo não são mapeados ou são parcialmente mapeados na UMLS/UMLS SN, tal fato pode ser um indício de

[15] SPARQL Protocol and RDF Query Language, https://www.w3.org/TR/rdf-sparql-query/,
[16] Disponível em http://www.nlm.nih.gov/pubs/factsheets/umls.html.

que o artigo traz uma nova descoberta, um acréscimo no conhecimento (MALHEIROS, 2010; MALHEIROS; MARCONDES, 2013). Maiores detalhes sobre nossa proposta de pesquisa podem ser encontrados em Marcondes e Costa (2016).

5 CONSIDERAÇÕES FINAIS

Nas seções anteriores vimos experiências de descoberta e reuso de conhecimento em artigos científicos digitais que se baseavam em duas alternativas. Na primeira, mineração de textos de artigos para a identificação de entidades possivelmente interrelacionadas encontradas nos textos ou resumos dos artigos. Esta alternativa, mesmo utilizando recursos como terminologias e ontologias computacionais para identificação, não leva em consideração outros elementos contextuais fornecidos pelo texto dos artigos; identificavam entidades, mas não retém seus contextos dentro do texto artigo.

Na segunda alternativa, a formalização dos elementos semânticos, retóricos ou metodológicos de artigos biomédicos na forma de ontologias ou modelos, encontradas nas diversas propostas, que servem basicamente para anotar semanticamente (de forma manual) os artigos. Estas anotações semânticas, embora importantes, demandam ainda leitura e processamento humanos, fazendo com que o processamento semântico do conteúdo dos artigos científicos se baseie unicamente em anotações manuais. Estas tecnologias formalizam os elementos semânticos, retóricos ou metodológicos de artigos biomédicos, mas, em geral, dependem de anotações manuais para as identificarem no texto dos artigos.

A Ciência da Informação desde seus primórdios sempre teve como uma de suas áreas de pesquisa o tratamento e recuperação de artigos científicos por métodos automatizados. Estamos em transição de um modelo de publicação centrado no artigo, de formato textual linear voltado para leitura por humanos, para um modelo em que o artigo é um dos elos de uma rede de recursos interligados e acessíveis simultaneamente, cujos conteúdos possam ser processados por programas.

Crescem as iniciativas na direção de converterem dados biomédicos em triplas RDF, como o Biogateway[17] e o Chem2Bio2RDF[18]. O Biogateway "… provides an entry point to access a data warehouse where biological data is gathered in the form of triples (using RDF). The systems can be queried using SPARQL". Segundo Chen et al. (2010), o portal Chem2Bio2RDF contém 25 datasets em diferentes domínios relacionando química e biologia, com um total de 78 milhões de triplas RDF. Um fator conservador para se evoluir em direção a um ambiente semântico de publicações científicas é o legado de milhões de publicações em formato textual linear mantido nas bases de dados bibliográficas. Como realizar a transição do antigo ambiente para o novo?

Percebe-se cada vez mais interfaces entre as áreas de mineração de textos biomédicos e publicações estendidas/publicações semânticas. É necessário avançar na integração das pesquisas, em especial nas duas áreas revistas, para que se possa obter resultados práticos e ferramentas que permitam atingir um novo patamar nas publicações científicas biomédicas no ambiente da Web Semântica.

REFERÊNCIAS

ABOUT JDS. Journal of Data Science. Disponível em: <http://www.jds-online.com/about>. Acesso em: 2 maio 2015.

AGARWAL, Ritu; DHAR, Vasant. Big Data, Data Science, and Analytics: the opportunity and challenge for IS research. **Information Systems Research**, Providence, v. 25, n. 3, p. 443-448, Sep. 2014.

AGICHTEIN, Eugene; GRAVANO, Luis. Snowball: extracting relations from large plain-text collections. In: ACM INTERNATIONAL CONFERENCE ON DIGITAL LIBRARIES, 5., 2000, New York. **Proceedings**… New York, 2000. Disponível em: <http://www.cs.columbia.edu/~gravano/Papers/2000/dl>. Acesso em: 4 set. 2009.

[17] http://www.semantic-systems-biology.org/biogateway/querying.
[18] http://cheminfov.informatics.indiana.edu:8080/c2b2r/

ARANHA, Christian Nunes. **Uma abordagem de pré-processamento para mineração de textos em português**. 2007. Tese (Doutorado em Engenharia Elétrica) - Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007.

ATTWOOD, Teresa K. et al. Calling international rescue: knowledge lost in literature and data landslide! *Biochemical Journal*, London, v. 424, n. 3, p. 317-333, Dec. 2009.

BAZERMAN, Charles. **Shaping written knowledge**: the genre and activity of the experimental article in science. Madison, Wisconsin: The University of Wisconsin Press, 1988.

BELL, Daniel. **O advento da sociedade pós-industrial**. São Paulo, Cultrix, 1977.

BERNERS-Lee, T.; HENDLER, J; LASSILA, O. 2001. "The semantic web". *Scientific American, v.* 284, n. 5, 2001.

BIOINFER: bio information extraction resource. 2006. Disponível em: <http://mars.cs.utu.fi/BioInfer/>. Acesso em: 5 ago. 2016.

BIOMEDICAL DISCOVERY SUPPORT SYSTEM. **Purpose**. Disponível em:<http://ibmi.mf.uni-lj.si/bitola/>. Acesso em: 12 maio 2010.

BODENREIDER, O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. *IMIA Yearbook of Medical Informatics*, p. 67-79, 2008.

BOGER, Zvi et al. Automatic keyword identification by artificial neural networks compared to manual identification by users of filtering systems. *Information Processing & Management*, Elmsford, v. 37, n. 2, p.187-198, 2001.

BRIAN, Sergey. Extracting patterns and relations from the world wide web. In: SELECTED PAPERS FROM THE INTERNATIONAL WORKSHOP ON THE WORLD WIDE WEB AND DATABASES, 1998, Valencia. **Proceedings…** Valencia, Spain, 1998. Disponível em: <http://bolek.ii.pw.edu.pl/~gawrysia/WEDT/brin.pdf>. Acesso em: 4 set. 2009.

BROOKES, Bertran. The foundations of information science. Part I. Philosophical aspects. *Journal of Information Science*, Cambridge, v. 2, p. 125-133, 1980.

BUNGE, Mario. *Philosophy of science*. London: Transaction Publishers, 1998.

CAFARELLA, Michael J. et al. KnowItNow: fast, scalable information extraction from the web. In: HUMAN LANGUAGE TECHNOLOGY CONFERENCE; CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2005, Vancouver. **Proceedings…**Vancouver: Association for Computational Linguistics, 2005.p. 563-570.

CARR, Leslie et al. The case for explicit knowledge in documents. In:**Proceedings of the 2004 ACM symposium on Document engineering**. ACM, 2004. p. 90-98.

CHEN, Bin et al. **Chem2bio2rdf: A linked open data portal for chemical biology**. 2010. Disponível em: <https://arxiv.org/ftp/arxiv/papers/1012/1012.4759.pdf>. Acesso em: 18 abr. 2016.

CHEN, Jinxiu et al. Automatic relation extraction with model order selection and discriminative label identification. In: INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING, 2., 2005, Jeju Island, Korea. **Proceedings…** Jeju Island, Korea, 2005.

CHOMSKY, N. **Syntactic structures**. The Hague: Mouton & Co., 1957.

COHEN, Aaron M.; HERSH, William R. A survey of current work in biomedical text mining. **Briefings in Bioinformatics**, London, v. 6, n. 1, p. 57-71, Mar. 2005.

COLLIER, Nigel et al. Recent advances in natural language processing for biomedical applications. **International Journal of Medical Informatics**, Shannon, v. 75, n. 6, p. 413-417, 2006.

COLLIER, Nigel; NOBATA, Chikashi; TSUJII, Jun-ichi. Extracting the names of genes and gene products with a hidden Markov model. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS,18., 2000, Saarbrücken. **Proceedings…** Saarbrücken, 2000. v.1.
COMMUNICATIONS in physics. 2001. Disponível em: <http://www.science.uva.nl/projects/commmphys>. Acesso em 15 mar. 2005.

CONRAD, Jack G.; UTT, Mary H. A system for discovering relationships by feature extraction from text databases. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 17., 1994, Dublin. **Proceedings…** Dublin: City University, 1994. p. 260-270.

COSTA, Leonardo Cruz. **Uma proposta de processo de submissão de artigos científicos às publicações eletrônicas semânticas em ciências biomédicas**. 2010. Tese (Doutorado em Ciência da Informação) - Universidade Federal Fluminense, Niterói, 2010.

DAVENPORT, Thomas. H. **Big data no trabalho**. Rio de Janeiro: Campus, 2014.

DE WAARD, A. et al. Modeling rhetoric in scientific publications. In: the International Conference on Multidisciplinary Information Sciences and Technologies, InSciT2006, October 2006, Merida, Spain. **Proceedings**… 2006; Merida, Spain, 2006. p. 25-28. Disponível em: <http://www.instac.es/inscit2006/ papers/pdf/133.pdf>. Acesso em: 30 mar. 2007.

DE WAARD, Anita. et al. Hypotheses, evidence and relationships: the HypER approach for representing scientific knowledge claims. InINTERNATIONAL SEMANTIC WEB CONFERENCE, 8th, 2009, Washington DC. **Proceedings…** Washington DC: Springer Verlag Berlin, 2009. p. 818-832.

DINAKARPADIAN, Deendayal et al. MachineProse: an ontological framework for scientific assertions. **Journal of the American Medical Informatics Association**, Philadelphia, v. 13, n. 2, Mar./Apr. p. 220-232, 2006.

DRUCKER, Peter. **Sociedade pós-capitalista**. São Paulo: Pioneira, 1995.

ERHARDT, Ramón A-A; SCHNEIDER, Reinhard; BLASCHKE, Christian. Status of text-mining techniques applied to biomedical text. **Drug Discovery Today**, Oxford, v. 11, n. 7-8, Apr. 2006. Disponível em: <http://www.drugdiscoverytoday.com/ echoice/jun2008/erhardt. pdf>. Acesso em: 23 jun. 2009.

EXPRESSÃO REGULAR. In: Wikipédia. Disponível em: <https://pt.wikipedia.org/wiki/Express%C3%A3o_regular>. Acesso em: 4 out. 2016.

FAKUDA, K. et al. Towards information extraction: identifying protein names from biological papers. **Pacific Symposium on Biocomputing**, p. 707-718, 1998. Disponível em: <http://www.ncbi.nlm.nih.gov/ pubmed/9697224>. Acesso em: 20 out. 2008.

FARRADANE, Jason E. L. Relational indexing. Part I. **Journal of Information Science,** v. 1, p. 267-276, 1980.

FELLBAUM, Christiane (Ed.). **WordNet**: an electronic lexical database. Cambridge: The MIT Press Cambridge, 1998.

FILATOVA, Elena; HATZIVASSILOGLOU, Vasileios. Domain-independent detection, extraction, and labeling of atomic events. In: RECENT ADVANCES IN NATURAL LANGUAGE PROCESSING, 2003, Borovetz, Bulgaria. **Proceedings**… Borovetz, Bulgaria, 2003.

GANIZ, Murat Can; POTTENGER, William M.; JANNECK, Christopher D. Recent advances in literature based discovery. **Journal of the American Society for Information Science and Technology**, New York, v. 56, 2005. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.77.6842&rep=rep1&type=pdf>. Acesso em: 6 abr. 2016. 7-11 September.

GARDIN, J-C. Vers un remodelage des publications savantes: ses rapports avec sciences de l'information. In: CHAUDRION, S.; Fluhr, C. (Ed.). **Filtrage et Résumé Automatique de l'Information sur les Reseaux - Actes du 3ème Colloque du Chapitre Français de l'ISKO**. Paris: Université de Nanterre-Paris X, 2001.

GENOME DATABASE. Disponível em: <http://www.gdb.org/>. Acesso em: 13 abr. 2009.

GILLAM, Michael et al. **The healthcare singularity and the age of semantic medicine**. 2009. Disponível em: <https://www.cs.umd.edu/users/ben/papers/Gillam2009healthcare.pdf>. Acesso em: 12 jan. 2016.

GINSPARG, Paul. Text in a data-centric world. In: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin. **The fourth paradigm**. Washington: Microsoft Research, 2009.

GORDON, Michael D.; DUMAIS, Susan. Using latent semantic indexing for literature based discovery. **Journal of the Association for Information Science and Technology**, United States, v. 49, n. 9, p. 674-685, 1998.

GREIFF, Warren R.; MORGAN, William T.; PONTE, Jay M. The role of variance in term weighting for probabilistic information retrieval. In: CIKM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 2002, New York. **Proceedings**... New York, 2002.p. 252-259.
GROSS, Alan G. **The rethoric of science**. London: Harvard Univerity Press, 1990.

GUIMARÃES, Carlos Alberto. Structured abstracts: narrative review. **Acta Cirúrgica Brasileira**, São Paulo, v. 21, n. 4, p. 263-268, ago. 2006. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-86502006000400014 >. Acesso em: 20 abr. 2009.

HACHEY, Benjamin. **Towards generic relation extraction**. 2009. Tese (Doctor of Philosophy) - Institute for Communicating and Collaborative Systems, School of Informatics University of Edinburgh, Edinburgh, 2009.

HASEGAWA, Takaaki; SEKINE, Satoshi; GRISHMAN, Ralph. Discovering relations among named entities from large corpora. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 42., 2004, Barcelona. **Proceedings...** Barcelona, 2004. p. 415-422.

HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin. (Ed.). Jim Gray on e-science: a transformed scientific method. In: _____. **The fourth paradigm**. Washington: Microsoft Research, 2009.

HOFFMANN, R.; VALENCIA, A. Implementing the iHOP concept for navigation of biomedical literature.*Bioinformatics*, Oxford, v. 21, n. 2, p. ii252-ii258, 2005. Disponível em: <https://nar.oxfordjournals.org/content/35/suppl_2/W21.full>. Acesso em: 24 abr. 2010.

HUCKA, Michael et al. **System biology markup language (SBML) level 1**: structures and facilities for basic model definitions. 2003. Disponível em: <http://www.sbml.org/specifications/sbml-level-1/version-2/sbml-level-1-v2.pdf>. Acesso em: 2 nov. 2005.

HUGO. **Nomenclature committee**. Disponível em: <http://www.gene.ucl.ac.uk/nomenclature/>. Acesso em 12 maio 2010.

HUNTER**,** Lawrence et al**.** Concept recognition for extracting protein interaction relations from biomedical text. **Genome Biology**, London, v. 9, Suppl 2, 2008.

HUTCHINS, John. On the structure of scientific texts. In: UEA PAPERS IN LINGUISTICS, 5., 1977, Norwich. **Proceedings…** Norwich, UK: University of East Anglia,1977. p. 18-39. Disponível em: <http://ourworld.compuserve.com/homepages/ wjhutchins/UEAP/L-1977.pdf>. Acesso em: 20 mar. 2006.

JENSSEN, T-K.; LÆGREID, A.; KOMOROWSKI, J.; HOVIG, E. A literature network of human genes for high-throughput analysis of gene expression. **Nature Genetics**, New York, v. 28, p. 21-28, 2001.

JONES, Karen Sparck, WILLETT, Peter. **Readings in information retrieval**. San Francisco: Morgan Kaufmann Publishers, 1997.

KANDO, N. Text-level structure of research papers: implications for text-based information processing systems. In: ANNUAL BCS-IRSG Colloquium on IR ResearchIRSG COLLOQUIUM ON IR,19th, 1997, Aberdeen. **Proceedings….** Aberdeen, Scotland: Springer-Verlag, 1997.

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

KEEN, P. Keynote address: relevance and rigor in information systems research. In: NISSEN, Hans-Erik; KLEIN, Heinz K.; HIRSCHHEIM, Rudy (Ed.). **Information systems research**: contemporary approaches and emergent traditions. North Holland: Elsevier Publishers, 1991. p. 27-49.

KINTSCH, Walter; van DIJK, Teun A. Towards a model of text comprehension and production. **Psycological Review***,* Washington, v. 84, n. 5, p. 363-393, 1972.

KNOESIS. **Semantics and services enabled problem solving environment for Tcruzi**. Disponível em: <http://knoesis.org/?q=projects/tcruzi>. Acesso em: 4 out. 2016.

KUHN, Thomas. **A estrutura das revoluções científicas**. 9. ed. São Paulo: Perspectiva, 2007.

LIDDY, Elizabeth. Text mining. **Bulletin of The American Society for Information Science**, Washington, v. 27, n.1, p. 13-14, 2000.

LINGPIPE. **Home**. Disponível em: <http://alias-i.com/lingpipe/>. Acesso em: 21 ago. 2009.

LOH, Stanley. **Abordagem baseada em conceitos para descoberta de conhecimento em textos**. 2001. Tese (Doutorado em ciência da Informação) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2001. Disponível em: <http://www.lume.ufrgs.br/handle/10183/1849>. Acesso: em 17 ago. 2009.

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

LYNCH, Clifford. Jim Gray's fourth paradigm and the construction of the scientific record. In: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (Ed.). **The fourth paradigm**. Washington: Microsoft Research, 2009. p. 177-184.

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

MALHEIROS, Luciana Reis; MARCONDES, Carlos Henrique. Identificación de indicios de descubrimientos científicos en artículos biomédicos mediante análisis de contenidos. **Revista Española de Documentación Científica**, Madrid, v. 36, n. 2, abr./jun. 2013. Disponível em: <http://dx.doi.org/10.3989/redc.2013.2.915>. Acesso em: 5 jan. 2014.

MARCONDES, Carlos Henrique. From scientific communication to public knowledge: the scientific article Web published as a knowledge base. In: EGELEN, Jan; DOBREVA, Milena (Ed.). ICCC ElPub - INTERNATIONAL CONFERENCE ON ELECTRONIC PUBLISHING, 9, 2005, Leuven, Bélgica. **Proceedings...** Leuven, Bélgica, 2005. p. 119-127. Disponível em: <http://elpub.scix.net>. Acesso em: 10 maio 2010.

MARCONDES, Carlos Henrique; COSTA, Leonardo C. A model to represent and process scientific knowledge in biomedical articles with semantic web technologies. Knowledge Organization, Wurzburg, Alemanha, v. 43, n. 2, p. 86-101, 2016.

MARKOWITZ, Judith; NUTTER, J. Terry; EVENS, Martha W. Beyond is-a and part-whole: more semantic network links. **Journal of Computers and Mathematics with Applications**, v. 23, n. 6, p. 377-390, 1992.

MARTIN, Philippe. Knowledge acquisition using documents, conceptual graphs and a semantically structured dictionary. In: GAINES, B. R. (Ed.). **Proc. of KAW'95**. Canada: University of Calgary, 1995.

MATTELART, Armand. **História da sociedade da informação**. 2. ed. São Paulo: Loyola, 2002.

MEDELYAN, Olena; WITTEN, Ian H. Domain-independent automatic keyphrase indexing with small training sets. **Journal of the American Society for Information Science and Technology**, New York, v.59, n.1, p.1026-1040, 2008.

MELUCCI, Massimo. Passage retrieval: A probabilistic technique. **Information Processing & Management**, Elmsford, NY, v. 34, n. 1, p. 43-68, 1998.

MILLER, David. **Explanation versus description.** ***Philosophical Review***, Ithaca, NY, v. 56, n. 3, p. 306-312, 1947.

MILLER, David. **Popper**: textos escolhidos. Rio de Janeiro: Contraponto, PUC, 2010.

MOENS, Marie-Francine. **Information extraction**: algorithms and prospects in a retrieval context. Dordrecht: Springer, 2006.

MULLER, Hans-Michael; KENNY, Eimear; STERNBERG, Paul W. Textpresso: an ontology-based information retrieval and extraction system for biological literature. **Plos Biology**, San Francisco, v. 2, n. 11, 2004. Disponível em: <http://www.plosbiology.org/article/info:doi/10.1371/ journal.pbio.0020309>. Acesso em: 9 mar. 2012.

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

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

MUSLEA, Ion. Extraction patterns for information extraction tasks: a survey. In: AAAI-99 WORKSHOP ON MACHINE LEARNING FOR INFORMATION EXTRACTION, 1999. **Proceedings...** Orlando, Florida, 1999. Disponível em: <http://www.ai.sri.com/~muslea/PS/ml4ie-aaai99.pdf> Acesso em: 17 ago. 2009.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. **LocusLink**. Disponível em: <http://www.ncbi.nlm.nih. gov/LocusLink/>. Acesso em: 13 abril 2009.

NWOGU, Kevin Ngozi. The medical research paper: structure and functions. **English for Specific Purposes**, v. 16, n. 2, p. 119-138, 1997.
PARIS. René Descartes University. **Genatlas**. Disponível em: <http://genatlas.medecine.univ-paris5.fr//>. Acesso em: 12 maio 2010.

PONTE, Jay M.; CROFT, W. Bruce. A language modeling approach to information retrieval. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 21st, 1998, New York, p. 275-281, 1998. **Proceedings…** New York, 1998.

POPESCU, Ana-Maria; ETZIONI, Oren. Extracting product features and opinions from reviews. 2005. Disponível em: <http://turing.cs.washington.edu/papers/ emnlp05_opine.pdf>. Acesso em: 17 ago. 2009.

PRINCETON UNIVERSITY. **WordNet**. Disponível em: <http://wordnet.princeton.edu/>. Acesso em: 10 set. 2008.

PUSTEJOVSKY, J. et al. Robust relational parsing over biomedical literature: extracting inhibit relations. In: PACIFIC SYMPOSIUM ON BIOCOMPUTING, 2002, Hawaii. **Proceedings…**Hawaii, 2002. p. 362-373.

PYYSALO, Sampo et al. BioInfer: a corpus for information extraction in the biomedical domain. **BMC Bioinformatics**, London, v. 8, n.50, p. 1-24, fev. 2007.

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

REBHOLZ-SCHUHMANN, Dietrich; OELLRICH, Anika; HOEHNDORF, Robert. Text-mining solutions for biomedical research: enabling integrative biology. **Nature Reviews Genetics**, London, v. 13, n. 12, p. 829-839, dez. 2012. Disponível em: <http://www.nature.com/nrg/journal/v13/n12/full/nrg3337.html>. Acesso em: 9 mar. 2016.

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

RIBEIRO, Claudio, J. S. Big data: uma investigação com uso de dados abertos sobre acidentes de trabalho. In: ENANCIB ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 15., 2014, Belo Horizonte. **Anais...** Belo Horizonte: ECI/UFMG, 2014. p. 4116-4131.

RICHARDSON, Stephen D.; DOLAN, William B.; VANDERWENDE Lucy. **MindNet**: acquiring and structuring semantic information from text. 1998. Disponível em: <http://www.aclweb.org/anthology/C98-2175>. Acesso em: 9 mar. 2016.

RILOFF, Ellen; LEHNERT, Wendy. Information extraction as a basis for high-precision text classification. **ACM Transactions on Information Systems**, New York, v.12, n.3, p. 296-333, jul. 1994.

RILOFF, Ellen; LORENZEN, Jeffrey. Extraction-based text categorization: generating domain-specific role relationships automatically. In: STRZALKOWSKI, T. (Ed.). **Natural Language Information Retrieval**. London: Kluwer Academic Publishers, 1999. v. 7, p. 167-196.

ROBERTSON, Alexander M.; WILLETT, Peter. An upperbound to the performance of ranked-output searching: optimal weighting of query terms using a genetic algorithm. **Journal of Documentation**, London, v. 52, n.4, p. 405-420, 1996.

ROSARIO, Barbara; HEARST, Marti. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2001, Pittsburgh. **Proceedings…** Pittsburgh, PA, 2001.

SALTON, Gerald; YANG, C. S.; YU, C. T. A theory of term importance in automatic text analysis. **Journal of the American Association Science**, v. 26, n.1, p. 33-44, 1975.

SAMWALD, Matthias. Extracting conclusion sections from PubMed abstracts for rapid key assertion integration in biomedical research. *Nature Proceedings* v. 3775, n.1, 2009.

SARAWAGI, Sunita. Information extraction. **Journal Foundations and Trends in Databases**, Hanover, MA, v. 1, n. 3, p. 261-377, Mar. 2008.

SAYÃO, Luis F.; SALES, Luana F. Curadoria digital: um novo patamar para preservação de dados digitais de pesquisa. **Informação & Sociedade**, João Pessoa, v. 22, n. 3, p. 179-191, set./dez. 2012.

SCHUTT, Rachel; O'NEIL, Cathy. **Doing Data Science**. Sebastopol, CA: O'Reilly Media, 2014.

SHANNON, Claude E. A mathematical theory of communication. **Bell System Technical Journal**, New York, v. 27, p. 379-423, 1948.

SHETH, Amit; ARPINAR, I. Budak; KASHYAP, Vipul. Relationships at the heart of semantic web: Modeling, discovering, and exploiting complex semantic relationships. In: **Enhancing the Power of the Internet**. Springer Berlin Heidelberg, 2004. p. 63-94.

SHOTTON, David. Semantic Publishing: the coming revolution in scientific journal publishing. Learned Publishing, v. 22, n., p. 85–94, April, 2009. Disponível em: <doi:10.1087/2009202>. Acesso em: 2 jul. 2012.

SKELTON, John. Analysis of the structure of original research papers: an aid to writing original papers for publication. **British Journal of General Practice**, London, v. 44, p. 455-459, 1994.

SMITH, Barry et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. **Nature Biotechnology**, New York, v. 25, p.1251–1255, 2007. Disponível em: <http://www.nature.com/nbt/journal/v25/n11/full/nbt1 346.html>. Acesso em: 25 março 2009.

SMITH, F. Jack. Data science as an academic discipline. **Data Science Journal**, Paris, v. 5, p. 163-164, Oct. 2006.

SODERLAND, Stephen. Learning to extract text-based information from the World Wide Web. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING - KDD-97, 3th, 1997 Newport Beach, California. **Proceeding…** Newport Beach, Califórnia, 1997.

SOLDATOVA, Larisa N.; KING, Ross D. Are the current ontologies in biology good ontologies?. **Nature biotechnology**, v. 23, n. 9, p. 1095-1098, 2005.

SRINIVASAN, Padmini. MeSHmap: A Text Mining Tool for MEDLINE. **Journal of Biomedical Informatics**, Philadelphia, p. 642-646. 2001. Disponível em <http://mingo.info-science.uiowa.edu/padmini/Papers/amia01.doc>. Acesso em 20 out. 2008.

STANTON, Jeffrey M. **An introduction to data science**. New York: Syracuse University, 2012.

STEIN, Lincoln D. Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. **Nature Reviews Genetic**, Londres, v. 9, 2008.

SUDO, Kiyoshi; SEKINE, Satoshi; GRISHMAN, Ralph. An improved extraction pattern representation model for automatic IE pattern acquisition. In: ANNUAL MEETING ASSN. COMPUTATIONAL LINGUISTICS, 41., 2003, Sapporo. **Proceedings…** Sapporo, Japan, 2003. Disponível: <http://nlp.cs.nyu.edu/publication/papers/sudo-acl03.pdf>. Acesso: 4 set 2009.

SWANSON, Don R. Fish oil, raynaud's syndrome, and undiscovered public knowledge. **Perspectives in Biology and Medicine**, Baltimore, v. 30, p. 7-18, 1986.

SWANSON, Don R. Medical literature as a potential source of new knowledge. **Bulletin of the Medical Library Association**, Chicago, v. 78, n. 1, p. 29, 1990.

SWANSON, Don R.; SMALHEISER, Neil R. An interactive system for finding complementary literatures: a stimulus to scientific discovery. **Artificial Intelligence**, v. 91, n. 2, p. 183-203, 1997. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0004370297000088>. Acesso em: 4 abr. 2016.

SWANSON, Don R.; SMALHEISER, Neil R; TORVIK, Vetle I. Ranking indirect connections in literature based discovery. The role of Medical Subject Headings. *JASIST,* Malden, v. 57, n. 11, p.1427-1439, 2006. Disponível em: <https://www.researchgate.net/profile/Neil_Smalheiser/publication/220435241_Ranking_indirect_connections_in_literature-based_discovery_The_role_of_medical_subject_headings/links/00b7d53550ec3e8dd8000000.pdf>. Acesso em: 30 set. 2009.

SZARVAS, György et al. The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts. **BMC bioinformatics**, London, v. 9, Suppl 11, 2008.

TANABE, L. et al. MedMiner: an internet text-mining tool for biomedical information, with application to gene expression profiling. **BioTechniques**, London, v. 27, p. 1210-1217, Dec.1999.

TEI: Text Encoding Initiative. Disponível em: http://www.tei-c.org. Acesso em: 7 out. 2016.

TENOPIR, Carol et al. Electronic journals and changes in scholarly article seeking and reading patterns. In: **Aslib proceedings**. Emerald Group Publishing Limited, 2009. p. 5-32.

UMLS. **Specialist natural language processing**. Disponível em:<http://lexsrv3.nlm.nih.gov/ SPECIALIST/index.html>. Acesso em: 11 julho 2008.

WEEBER, Marc et al. Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. **Journal of the American Medical Informatics Association**, Philadelphia, v. 10, n. 3, p. 252-259, 2003. Disponível em: <https://lhncbc.nlm.nih.gov/files/archive/pub2003034.pdf>. Acesso em: 19 out. 2005.

YEH, Alexander S. A.; HIRSCHMAN, Lynette; MORGAN, Alexander A. Evaluation of text data mining for database curation: lessons learned from the KDD challenge cup. **Bioinformatics**, Oxford, v. 19, suppl. 1, p.331-39, 2003. Disponível em: <http://bioinformatics.oxfordjournals. org/cgi/reprint/19/suppl_1/i331>. Acesso em: 10 jul. 2009.

ZIPF, George Kingsley. **Human behaviour and the principle of least effort**. New York: Addison-Wesley, 1949.

ZWEIGENBAUM, Pierre et al. Frontiers of biomedical textmining: current progress. **Briefings In Bioinformatics**. Oxford, v. 8. n. 5. p.358-375, 2007. Disponível em: <http://bib.oxfordjournals.org/cgi/reprint/8/5/358>. Acesso em: 23 jun. 2009.

Title

Knowledge discovery in digital articles in biomedical sciences

Abstract
Abstract
Introduction: The emergence of Semantic Web is impacting the scientific digital publications scenario. The growth of biomedical literature in digital format raises the question that new discoveries may originate not only from laboratories but from bibliographic and factual databases too. The textual linear format of scientific articles, for human reading, is not adequate to be processed by computers.
Objective: To contextualize the research problem on semantic models of digital articles in biomedical sciences, identifying related areas and producing a review.
Methodology: This research is descriptive and exploratory, with focus in the former research problem, trying to identify interfaces between the areas reviewed; the approach is qualitative and bibliographic and documental research methods were used.
Results: A state-of-the-art and the main trends in each area reviewed – biomedical text mining and enhanced/semantic publications - are presented. Both areas are complementary although not integrated.
Conclusions: We are in a transition from a publication model centered in the linear textual article towards a model in which a new format of digital article is just one of many links in a network of resources, simultaneously accessed and processed by programs, taking full advantage of Semantic Web technologies.

Keywords: knowledge discovery in literature. text mining. enhanced publications. semantic publications. Biomedical Sciences.

Título

Descobrimiento de conocimiento en los articulos digitales en ciencias biomédicas

Resumen
Introducción: La emergencia de la Web Semántica está afectando el ambiente de las publicaciones científicas digitales. El aumento de la literatura biomédica en formato digital plantea la cuestión de que pueden surgir nuevos descubrimientos no sólo en laboratorios, sino también en la literatura y las bases de datos fácticos. La forma textual lineal de artículos, hecha para la lectura humana, se presenta inadecuada en su tratamiento por las computadoras.
Objetivo: Contextualiza los problemas de investigación de modelos semánticos para contenidos digitales de las ciencias biomédicas, la identificación de áreas relacionadas y sus interfaces, además de producir una revisión de la literatura.
Metodología: La investigación es descriptiva y exploratoria, con énfasis en el problema de investigación, buscando identificar las interfaces entre las áreas revistas; el enfoque es cualitativo y los métodos fueron la búsqueda bibliográfica y documentaria.
Resultados: Exposición del estado del arte y temas principales en cada subarea, minería de textos biomédicos y publicaciones extendidas / modelos semánticos de las publicaciones; las areas son complementarias pero no integradas.
Conclusiones: Estamos en la transición de un modelo de publicación centrado en el artículo de texto lineal a un modelo en el que el elemento digital, en un nuevo formato, es uno de los eslabones de una red de recursos interconectados y con acceso simultáneo, procesables por programas, aprovechando las tecnologías de web semántica.

Palabras clave: Descubrimiento de conocimiento en la literatura. La minería de texto. Publicaciones extendidas. Publicaciones semánticas. Ciencias biomédicas.

Enviado em: 17.07.2016.
Aceito em: 20.11.2016.