Como construir indicadores de Ciência, Tecnologia e Inovação usando *Web of Science, Derwent World Patent Index, Bibexcel e Pajek*?

Terry Lima Ruas

Mestre em Engenharia da Informação pelo Programa de Pós-Graduação em Engenharia da Informação da Universidade Federal do ABC. Bacharel em Ciência da Computação e Ciência & Tecnologia pela Universidade Federal do ABC. Atualmente exerce a função de Delivery Cost Branding & Service Planner para a divisão de Software da IBM Brasil.

Luciana Pereira

Professora do Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas da Universidade Federal do ABC e responsável pelo iLab@UFABC. Realizou pesquisas de Pós-Doutorado no Technological Change Laboratory da Universidade Columbia e no Observatório da Inovação e Competitividade do Instituto de Estudos Avançados da Universidade de São Paulo. Possui formação interdisciplinar, tendo obtido os títulos de Doutora e Mestra em Engenharia de Produção pela Escola Politécnica da Universidade de São Paulo, com Estágio Sanduíche no Instituto de Tecnologia de Massachusetts, e de Bacharel em Economia pela Universidade Estadual Paulista.

http://dx.org/10.1590/1981-5344/1678

O objetivo deste artigo é apresentar um processo que explique como usar softwares livres para buscar, extrair (Bibexcel) e visualizar dados (Pajek) dos repositórios de publicações científicas (Web of Science) e tecnológicas (Derwent World Patent Index) para construção de indicadores de produtividade científica e tecnológica. Espera-se que este processo contribua como um guia metodológico para a realização de estudos cientométricos, de tal modo que permita aos pesquisadores e aos gestores de Ciência, Tecnologia e Inovação (CTI) sem conhecimentos avançados em computação a obter informações de forma mais prática, porém confiáveis, dos

repositórios e que, a partir delas, possam construir indicadores e elaborar avaliações mais precisas de CTI.

Palavras-chave: *Cientometria; Construção de Indicadores de Produtividade Científica e Tecnológica;Gestão da Ciência, Tecnologia e Inovação*

How to build Science, Technology, and Innovation Indicators usingWeb of of Science, Derwent World Patent Index, Bibexcel, and Pajek?

The goal of this paper is to present a process to retrieve (Bibexcel), organize, and visualize (Pajek), information from data repositories for research publications (Web of Science) and technological efforts (Derwent World Patent Index) to build science and technology indicators. The process contributes as a methodological guide for scientometric studies in such a way that enables researchers and managers in Science, Technology and Innovation studies without advanced skills in computer science how to deal with large repositories in a more practical, yet reliable, way. The result is development of indicators that will allow more accurate assessments of science, technology and innovation.

Key-words: *Scientometrics, Science and Technology Indicators, Tools for Management of Science, Technology and Innovation*

Recebido em 09.12.2012 Aceito em 06.04.2014

1Introdução

Uma das principais características da sociedade da informação e do conhecimento é a velocidade com que os avanços científicos e tecnológicos têm ocorrido. De acordo com Gantz e Reinsel (2012), a produção global de dados atingiu a marca de 2,8 zettabytes (ZB) em 2012 - ou 2,8 trilhões GB, dos quais apenas 0,5% serão analisados. Uma questão importante é que por detrás dessa quantidade de informação encontram-se implícitos padrões de comportamento e outras características úteis para traçar cenários e antecipar tendências sobre aspectos sociais e econômicos da sociedade contemporânea. Desse modo, percebe-se a importância dos instrumentos que facilitem a busca, coleta,

processamento e a análise de dados a fim de se extrair algum conhecimento dos mesmos (MEARIAN, 2011).

Dado este cenário, o objetivo deste artigo é apresentar um processo que permita recuperar informação em base de dados bibliográficos e de patentes, que são a forma que os investimentos em política de ciência, tecnológica e inovação se tornam públicos e podem ser mensurados. Para alcançar esse propósito, este artigo apresenta um processo que integra diferentes softwares livres de tal forma que seja possível capturar, organizar e apresentar as informações que podem ser extraídas de diferentes bases de dados bibliográficos e de patentes. Para ilustrar os procedimentos e a funcionalidade do processo, ao final é feita uma aplicação no campo da biofotônica no Brasil, na China e nos Estados Unidos.

Uma base de dados bibliográficos é uma coleção digital que contém os registros da literatura publicada, com informações sobre o que foi publicado (artigo de periódico, conferência, livro), quem publicou (autor, instituição, pais) e onde se publicou (periódico A, B ou C). Quando os esforços de pesquisas resultam numa solução para um problema tecnológico, ao invés de se publicar o resultado na forma de um artigo, ele será patenteado, pois se acredita que a tecnologia possa ser incorporada num produto ou processo e, por isso, deve ser protegida. Assim como os dados bibliográficos, há um sistema similar que são as bases de dados de patentes, as quais fornecem informações sobre inventor, invenção e prazos de validade de um invento.

O caráter inovador deste artigo reside na estruturação do processo que procura mostrar de forma didática como utilizar ferramentas computacionais na extração de dados para fazer um mapeamento de um determinado campo do conhecimento usando bases de dados bibliográficos e de patentes que contém milhares de informações. Enquanto na literatura há vários estudos sobre modelos cientométricos avançados (COCCIA, 2005; GLÄNZEL, 2010) ou análises que discutem campos específicos do conhecimento, tais como os estudos sobre a biotecnologia realizados por De Moura e Caregnato (2011) e da ciência da informação feitos por Oliveira e Gracio (2011), não encontramos artigos que expliquem efetivamente como é o processo de recuperação de informação das bases de dados utilizando ferramentas computacionais do começo ao fim.

Sendo assim, o processo de mineração e visualização dos dados proposto neste artigo foi elaborado para explicar os passos a serem seguidos para extração e filtro dos dados desejados da maneira mais simples possível. Com essas informações é possível fazer análises sobre a evolução de áreas do conhecimento, seja dentro de um país ou em perspectiva comparada.

O presente artigo está organizado da seguinte maneira. A seção introdutória apresenta a contextualização e explicita o objetivo do artigo. Na Seção 2 é feita uma breve fundamentação teórica do trabalho. Na Seção 3 é apresentada a metodologia utilizada para a execução da

pesquisa. Enquanto a seção 4 apresenta o processo e suas especificidades de forma detalhada. A última seção apresenta as principais implicações do trabalho apresentado.

2 Como se mensura a produção científica e tecnológica?

A cientometria pode ser definida como o campo do conhecimento que se preocupa com os métodos e ferramentas que auxiliam no processo de mensuração e análise das atividades de pesquisa científica. Os resultados científicos, frutos dos avanços do conhecimento, comumente são difundidos de duas maneiras principais: i) na forma da literatura científica e ii) em possíveis aplicações tecnológicas. Em geral, se mensura a produção literária a partir da análise de indicadores bibliométricos. Enquanto que conhecimentos com potencial tecnológico para ser aplicados em produtos ou processos acabam por ser patenteados, que é um termo de posse de propriedade intelectual e que visa obter o monopólio daquela aplicação com objetivos econômicos (ZITT; BASSECOULARD, 2008).

A avaliação de trabalho científico é frequentemente medida por meio de indicadores de produtividade científica. Dentre esses indicadores cientométricos, a bibliometria é considerada o instrumento com o maior potencial de fazer um mapeamento acurado dos desenvolvimentos dos mais diversos campos científicos (GLÄNZEL, 2012). Alguns indicadores bibliométricos padrão incluem o número de artigos publicados, o impacto medido pelo número total de citações recebidas, o número médio de citações por artigo, o número de artigos com contagem de citações acima da média e os valores potenciais de artigos adquiridos através fator de impacto dos periódicos onde os artigos foram publicados. Enquanto que na avaliação do status do autor, da instituição e de um país é importante saber em quais periódicos os resultados de pesquisa foram publicados, até que ponto eles foram notados e quem os notou, características mensuradas pelo processo de citação.

A técnica favorita para o mapeamento da ciência é feita por meio de citação entre os documentos, palavras-chave e descritores textuais extraídos do corpo do texto de documentos científicos. Isso pode ser alcançado através de estudos empíricos em publicações e características de citações, noções de qualidade científica, diferenças em práticas de comunicação realizadas pelas diversas disciplinas, comparação com julgamentos qualitativos dos pares, e outras mais. Todas essas atividades têm como foco objetos centrais da pesquisa: a investigação da transferência e disseminação do conhecimento, assim como o estudo do progresso científico e das mudanças em sua relação com a sociedade em diversos aspectos (RAAN, 2000).

O status do periódico em que a pesquisa foi publicada também é um dos indicadores usados para avaliar cientistas e instituições, sendo obtidos através do uso do fator de impacto. O fator de impacto de um periódico é uma medida da frequência com a qual um artigo foi citado durante certo período de tempo, sendo assim considerado um indicador de avaliação da

qualidade de um periódico. Ele não deve ser usado para avaliar um artigo ou cientista de forma individual.

Por sua vez, uma patente é uma concessão pública, conferida pelo Estado, que garante ao seu titular um monopólio temporário, de forma a impedir que outros fabriquem, usem, ou vendam tal inovação livremente. Em contrapartida, as informações sobre patentes, que contém os pontos essenciais sobre as reivindicações que caracterizam a novidade no invento, é de domínio público, e, por isso, é uma valiosa fonte de informações (JAPAN PATENT OFFICE, 2011).

Informações sobre patentes, como a publicação de pedidos de patentes não examinados, apresentam várias vantagens exclusivas como informações técnicas: abrange uma ampla variedade de tecnologias, incluindo o estado-da-arte da tecnologia, bem como informações sobre as invenções estrangeiras no idioma nativo do leitor. Informações sobre patentes também incluem o conteúdo de um direito exclusivo ou um direito de propriedade intelectual, que são, inevitavelmente, uma parte da atividade econômica atual.

Além disso, a informação sobre patentes é uma indicação útil para as estratégias de desenvolvimento tecnológico ou estratégias globais das empresas em resposta à intensificação da concorrência. Consequentemente, as empresas, as universidades e os institutos de pesquisa utilizam informações sobre patentes ainda no estágio inicial de sua P&D, a fim de identificar tendências, para avaliar inovações e evitar infringir patentes em vigor, usando tais dados para gerir sua propriedade intelectual.

Em termos de técnicas para análise de patentes, alguns métodos têm sido aplicados para reconhecer as tendências de desenvolvimento tecnológico. Alguns desses métodos utilizam técnicas de mineração de texto para analisar os dados textuais de documentos de patentes, como o título e resumo, outra técnica utilizada á análise de citação de patentes, que é bastante semelhante à citação bibliográfica (MARCO, 2007).

Ao final, o objetivo das análises bibliométricas e de patentes é mensurar o processo de transformação e aplicação do conhecimento. A partir do seu entendimento é possível identificar redes nacionais e internacionais de colaboração, mapear a evolução de novos campos da ciência e da tecnologia, bem como conhecer a lógica interna de desenvolvimento da ciência. Por essa razão, os métodos cientométricos estão cada vez mais sendo utilizados para analisar a evolução e as tendências em CTI. Apesar disso, fazer a interface entre a utilização de sistemas de informações computacionais e interpretação dos dados obtidos demanda conhecimentos interdisciplinares, que é um dos principais desafios das pesquisas nessa área. Portanto, projetos mal elaborados, cálculos inadequados e avaliações malfeitas de indicadores cientométricos influenciam negativamente a sua apreciação pela comunidade científica, e, assim, prejudicam a aplicação de indicadores cuidadosamente construídos (RAAN, 2005).

Uma vez que a importância das interações entre os diversos campos da CTI é identificada, a necessidade do desenvolvimento de ferramentas que descrevam empiricamente as diversas formas que essas podem tomar fica evidente.

3 Extração de conhecimento de base de dados

A técnica de mineração de dados será utilizada como procedimento metodológico para a realização da pesquisa. O processo de Extração de Conhecimento em Base de Dados comumente chamada de KDD (*Knowledge Discovery in Databases*) tem como objetivo coletar dados que possuam uma relação de interesse por assunto e de validade para cada dado extraído. Em outras palavras, esta técnica utiliza-se de algoritmos de aprendizado de máquina capazes de generalizar os fatos encontrados em um grande repositório de dados, na forma de regras de alto nível compreensíveis ao ser humano e de grande valor para uma tomada de decisão.

Quando se tenta realizar tarefas referentes ao descobrimento de conhecimento em aplicações do mundo real, percebe-se que as mesmas podem ser de extrema complexidade e que a tarefa de mineração de dados representa apenas uma porção, porém de grande importância, do processo global. Assim, o processo de KDD deve ser visto como sendo composto por várias etapas interligadas (PONNIAH, 2001). Sob a perspectiva de análise de dados, a cientometria e o KDD estão fortemente relacionados.

O processo de KDD é interativo e iterativo, envolvendo diversas etapas, sendo que cada etapa gera um conjunto de conhecimentos. Essas etapas normalmente são realizadas de forma sequencial, ou seja, é preciso compreender o domínio de aplicação, selecionar e transformar os dados para depois tentar encontrar padrões nos dados. Por se tratar de um processo interativo, as pessoas envolvidas na sua realização devem possuir um canal de comunicação que viabilize uma troca de informações transparente (FAYYAD, 1996).

3.1 Seleção dos dados

A escolha dos dados a se investigar é uma tarefa importante tendo em vista que, através deles explorar-se-á o sistema social no qual se deseja extrair algum tipo de conhecimento. A internet possui um número quase que ilimitado de dados, portanto a seleção de quais dados serão observados é um pré-requisito indispensável para esse trabalho. Graças a grande quantidade de informações disponíveis o processamento de todo esse repositório seria algo computacionalmente inviável, seja por tempo hábil ou por recursos de *hardware/software.*

3.1.2 Domínio de aplicação

É a área do conhecimento sobre o qual se deseja estudar e do qual serão extraídos os dados que servirão de base para a geração de informação, que será interpretada, transformando-se em conhecimento, que atuará como apoio a tomada de decisões. Nessa etapa é preciso estar familiarizado com o campo a ser estudado.

3.1.3 Seleção dos repositórios

Para analisar a produção científica (artigos) existem diversas bases que armazenam publicações dos últimos anos. Dentre os mais conhecidos citados podemos citar:

ISI Web of Science (ISI WOS)
Scopus
Google Scholar
NLM's MEDLINE

De acordo com (BARÍLAN, 2010), (FALAGAS, PITSOUNI, *et al.*, 2008), (MIKKI, 2010), essas bases não cobrem a área científica da mesma maneira. Isto é, cada base possui uma característica que a difere das outras, seja pelo tipo de dado que essa possui ou pela facilidade em exportar o conteúdo desejado.

Existem repositórios especializados em trabalhos publicados na área médica como no caso do NLM's MEDLINE. Outros se concentram em armazenar trabalhos acadêmicos publicados nas mais diversas áreas do conhecimento, caso do *Google Scholar*. Algumas bases de dados, como por exemplo, a ISI WOS, possuem tamanha complexidade e valor agregado que o seu acesso é restrito à somente algumas instituições que pagam pelo seu uso podem utilizá-las. Há outras características dessa base, tais como funcionalidade. É possível exportar dados armazenados pela ISI WOS automaticamente. Porém, o mesmo não é possível com o sistema *Google Scholar*.

No âmbito das produções tecnológicas (patentes) é possível citar alguns repositórios mais tradicionais, conhecido como a tríade:

a)Escritório Americano de Patentes (USPTO)
b)Escritório Europeu de Patentes (EPO)
c)Escritório Japonês de Patentes (JPO)

A *Derwent World Patent Index* (DWPI) que é um banco de dados que contém os pedidos e concessões de patentes tendo como fonte 44 autoridades mundiais emissoras de patentes. Tendo em vista todas as bases de dados consideradas para o projeto, e os critérios de seleção estudados e explorados, delimitaram-se como fontes deste estudo os repositórios *ISI Web of Science* (WOS) para as publicações científicas e *Derwent Patent Index* (DWPI) para as publicações tecnológicas.

3.2 Coleta dos dados

Para facilitar a tarefa de processamento dos dados optou-se por utilizar softwares gratuitos. Dentre os diversos softwares disponíveis, uma pré-seleção estabeleceu quais os programas seriam testados para avaliação de suas funcionalidades. Tomou-se como base o artigo de Cobo (2011) o qual elenca diversos *softwares* de processamento e visualização. Esse artigo faz uma análise profunda levando em conta as principais características de cada programa, analisando aspectos como: medidas de normalização, algoritmos de agrupamento, capacidade de eliminação de ruídos, tipos de análise bibliométrica, elementos de pré-processamento, métodos de análises, entre outras funcionalidades. Após consultar o estudo apresentado, testes com cada um dos programas foram explorados no intuito de melhor entendê-los e validar as informações. Excluiu-se a utilização do *VantagePoint, CoPalRed* e o *IN-SPIRE*, pois todos esses são *softwares* comerciais. Apesar de muito completos, a obtenção de uma licença para seu uso fica fora do escopo financeiro desse projeto.

3.2.1Programas de processamento de dados

i)Bibexcel

O *Bibexcel*, embora não tenha uma interface intuitiva, permite a construção de uma linha do tempo interessantes que mostra a relação entre o nível de citação entre os autores e o respectivo ano no qual isso ocorreu. Essa informação, atrelada com os trabalhos em si, de cada autor, faz com que seja possível mapear um trabalho/ideia/inovação feita por um desses autores, desde sua primeira publicação até os dias mais recentes.

Também é possível saber o quão importante um trabalho fora conceituado, de modo a tornar-se referência para futuros trabalhos. Uma medida interessante é também saber como trabalhos de um mesmo autor estão relacionados com as referências de outros trabalhos contemporâneos. Com a informação do resumo é possível relacionar como trabalhos, altamente referenciados, propiciam que novas ideias, totalmente divergentes entre si, surjam.

Uma desvantagem desse programa é que basta se equivocar em uma etapa para que o mesmo seja encerrado sem maiores informações. Uma vantagem é a sua flexibilidade de integração com outras ferramentas como *Pajek*, *VOSViewer*, *Mapequation, NetDraw, Ucinet* e outros. Além disso, ele é capaz de pegar dados de diversas fontes, como *Scopus* e WOS (PERSSON, DANELL; SCHNEIDER, 2009).

ii)Science of Science Tool (Sci²Tool)

Ao primeiro contato mostra-se uma ferramenta muito bem estruturada pedindo que seja feito um registro antes de qualquer ação. O *software* é apresentado em diversas conferências e defendido por vários pesquisadores. Para visualização o software atende todos os requisitos, fazendo com que os dados processados possam ser vistos em forma de

Tree View (*prefuse beta*), *Tree Map* (*prefuse beta*), *Balloon Graph* (*prefuse alpha*), *Radial Tree/Graph* (*prefuse alpha*).

O software é muito rápido, amigável, possui diversas funcionalidades, filtros, algoritmos de processamento, visualização. No entanto, o programa gerou muitos erros para os exemplos dados, talvez por alguma incompatibilidade de versão do software.

Por exemplo, na visualização dos dados pelo '*Visualization* > *Networks* > GUESS nenhum comando executado na aba de "*interpreter*" parece funcionar adequadamente. Uma vantagem é poder configurar quanto de memória será alocada para subir o programa (COBO, *et al*, 2011); (SCI² TEAM, 2009).

iii) CiteSpace II

Dentre os programas disponíveis gratuitamente esse é o que oferece mais opções quanto a análise bibliométrica. Essa ferramenta é capaz receber dados de diversas fontes como: WOS, PubMed, arXvid e SAO/NASA *Astrophysics Data System*. Além de também trabalhar com dados de patentes como o DWPI.

É possível conduzir diferentes tipos de análises bibliométricas através de sua utilização como: co-autores, instituições de co-autores, países de co-autores, ocorrência de categorias de assuntos, co-citação de autores, redes de *journals*, entre outras. A única característica técnica que o difere do *Bibexcel* é a utilização de algumas medidas de normalização de dados.

Essa ferramenta possui um arcabouço completo para análise bibliométrica. Ela fornece os materiais necessários para detectar, analisar e visualizar padrões e tendências em registros científicos. No entanto, a sua utilização não é intuitiva e suas documentações são pouco detalhadas fazendo com que o usuário precise investir mais tempo em seu estudo (CHEN, 2003), (CHEN, 2006), (CHEN, 2004).

iv) Leydesdorff's Software

O programa desenvolvido por *Loet Leydesdorff* é bem explorado sendo material de estudo para diversos trabalhos no meio acadêmico. Essa ferramenta possibilita algumas análises bibliométricas como: co-palavras, co-autores, gráficos de acoplamento entre autores e periódicos, e co-citação entre autores.

A documentação para sua utilização é confusa e disponibilizada de maneira muito dispersa o que dificulta o entendimento do mesmo. Outro ponto que vale a pena ressaltar é que o *Leydesdorff's Software*, não possui mecanismos de pré-processamento e normalização de dados, o que pode comprometer alguns estudos bibliométricos dependendo do objetivo do trabalho (LEYDESDORFF, N/A).

v)Network Workbench Tool (NWBT)

Essa ferramenta é muito similar ao *Science of Science Tool (Sci²Tool)*, ambas possuem os mesmos módulos de pré-processamento e normalização de dados. Ambos podem ler diversos formatos de dados, como ISI WOS, *Scopus, Bibtex* e *EndNote Export Format*.

O NWBT possui um excelente manual de usuário que explica a ferramenta em detalhes, além de explorar a importantes aspectos do mapeamento cienciométrico. Tanto o NWBT quanto o *Sci²Tool* são as únicas ferramentas que investigam essa questão. No entanto, assim como no *Sci²Tool* não é possível fazer análises completas de co-autores e co-citações, prejudicando muito a sua escolha como um programa de análise bibliométrica para objetivo desse trabalho (NWB TEAM, 2006).

3.3 Interpretação dos resultados

Os usuários envolvidos devem interpretar os padrões extraídos e, para tal, podem lançar mão de ferramentas estatísticas e de visualização que permitam fazer uma "leitura" precisa sobre os resultados. Isso possibilitará a verificação da validade e novidade, ou mesmo, a (ir)relevância dos padrões encontrados.

3.3.1Programa de mapeamento visual

i)VOSViewer

É uma ferramenta focada na visualização e construção de mapas bibliométricos. Com ele mapas podem ser criados a partir de dados de rede, através das técnicas de mapeamento VOS e agrupamento VOS. O VOSViewer pode ser usado para explorar mapas sob diferentes perspectivas, cada uma enfatizando uma característica em específico como: publicações, autores, mapas de palavras-chave, co-ocorrência de citações, entre outras.

A descrição e a estrutura que suporta essa ferramenta são bem construídas, inclusive é possível executar a ferramenta diretamente da página pela qual o programa é disponibilizado. Fazendo com que o mesmo seja portável para inúmeras plataformas. O próprio programa possui um manual que explica suas características, mas a falta de exemplos práticos faz com que seu entendimento seja restrito a teoria apresentada. Seria interessante se o próprio manual explorasse exemplos práticos para que o usuário fosse capaz de validar os pontos apresentados nos manual. Vale ressaltar que o VOSViewer não é capaz de construir nenhum mapa de redes bibliométricas, apenas visualizá-lo (ECK e WALTMAN, 2009); (ECK e WALTMAN, 2010).

ii)Pajek

Assim como o VOSViewer esse é um programa focado apenas na visualização de dados bibliométricos, ou seja, não é possível construir qualquer tipo de mapa com ele, apenas visualizá-lo. Essa ferramenta é largamente utilizada na comunidade científica, tanto que na grande maioria dos programas de análise bibliométrica existe a opção de exportar os dados para serem utilizados pelo Pajek. Até mesmo o VOSViewer disponibiliza um modo que executa os mesmos arquivos utilizados pelo Pajek (BATAGELJ e MRVAR, 2008).

A existência de um Wiki para o programa agrega muito a sua utilização, pois é possível trocar informações diretamente com os usuários

da ferramenta. Devido a sua larga utilização é possível encontrar muitos documentos que ensinam a utilizá-lo em conjunto com outras ferramentas e inclusive a explorá-lo em detalhes.

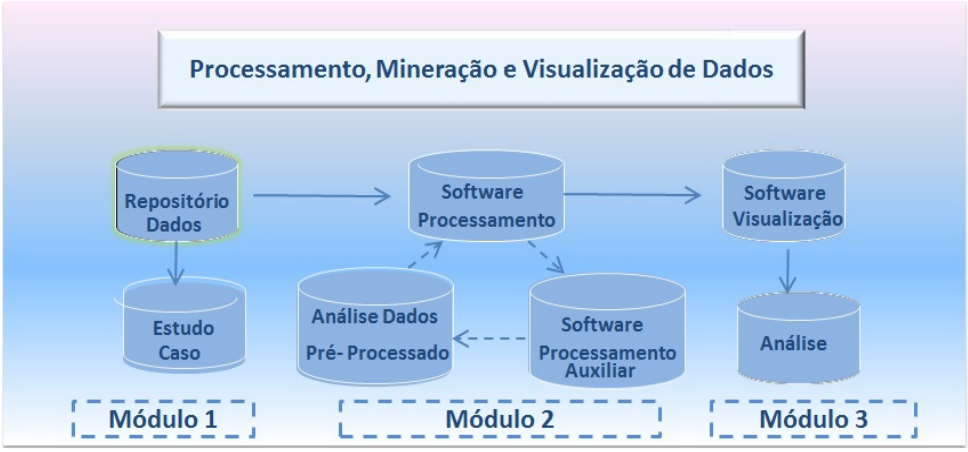
4 Processo de integração de um sistema de informação para mapear Ciência, Tecnologia e Inovação

Nesta seção apresentamos um processo que integra *softwares* livres que, quando acoplados, formam um sistema para obter informações sobre dados de CTI de um campo do conhecimento previamente escolhido. A elaboração de um sistema modular e genérico tem como principal objetivo torná-lo flexível para que esse possa ser aplicado aos mais diversos campos de conhecimento que se deseja estudar. Separando cada etapa do sistema em um componente independente possibilita que as etapas de extração, processamento e visualização de informação se adaptem às necessidades da pesquisa.

Cada componente do sistema é especializado em uma atividade e com atribuições bem definidas. É possível agrupar inúmeras tarefas em um único estágio, ganhando tempo, mas o sistema ficaria dependente de uma ferramenta e/ou base de dados única. Como essa abordagem tornaria o processo muito limitado, optou-se pela escolha de módulos desacoplados. Nesse processo, cada variável do sistema pode ser substituída por outro, caso este desempenhe melhor a função desejada. A Figura 1 ilustra o fluxograma do processo.

O primeiro módulo (1) é o ponto de partida no qual se define o estudo de caso e os repositórios de dados que possuirão os registros extraídos. Nesse módulo é aplicado o filtro de buscas, que recuperará os dados que serão processados nas próximas etapas, após serem extraídos. O segundo módulo (2) tem como ~~por~~ objetivo processar e formatar os dados obtidos através das ferramentas selecionadas. Muitas vezes, um único programa de processamento não é suficiente para que os dados sejam trabalhados a ponto de serem exportados para a próxima etapa do sistema. Desse modo, programas auxiliares podem ser utilizados, no intuito de refinar os resultados obtidos pelos *softwares* de processamento. O último módulo (3) é responsável pela consolidação dos dados obtidos, na forma de mapas e tabelas para que possam ser analisados adequadamente.

Figura 1- Processo de um sistema de informação para mapeamento da CTI



Fonte: Elaborado pelos autores

Nas próximas seções são explicadas as etapas do processo para especificação, obtenção, coleta, pré-processamento, processamento e visualização dos dados para dois repositórios internacionais. O repositório *Web of Science* (WOS) foi escolhido por ser uma base bastante ampla e uma das mais utilizadas em estudos cientométricos. A base *Derwent World Patents Index* (DWPI) é um banco de dados que contém pedidos e concessões de patentes tendo como fonte quarenta e quatro autoridades mundiais emissoras de patentes.

A escolha do *Bibexcel* como programa de processamento e do *Pajek* como de visualização ilustram uma decisão da pesquisa, que teve como base as experiências obtidas com os testes em cada *software*. *Bibexcel* e *Pajek* também apresentaram um melhor custo/benefício quanto a sua facilidade de operação e com seu poder de processamento para o escopo escolhido. Outro quesito importante no processo de seleção foi a compatibilidade entre os *softwares* de processamento (*Bibexcel*) e de visualização (*Pajek*) e os repositórios, pois além de serem livres se mostraram bastante completos na funcionalidades necessárias para realizar um mapeamento das bases WOS e DWPI.

4.1 Módulo de definição do repositório, campo do conhecimento e busca de dados

Antes de coletar qualquer tipo de informação é necessário estudar o campo do conhecimento (*e.g.* nanotecnologia, física das partículas, biofotônica) que se deseja coletar informações. A familiarização com o objeto de estudo é imprescindível para que se consiga investigar os pontos relacionados. Para isso, vê-se necessário: analisar áreas de atuação em que esse campo se concentra, palavras chaves utilizadas sobre o assunto, eventos relacionados à área, entre outros. Qualquer material capaz de enriquecer o portfolio de palavras e termos de busca deve ser considerado. Após coletar o máximo de informações possíveis é

interessante adquirir sinônimos e variações para que os resultados encontrados na busca recuperem mais dados sobre o objeto de estudo.

4.2.1 Formulação de expressão lógica para o processo de busca

Uma vez selecionadas as palavras e os termos é interessante construir uma expressão lógica de forma que o resultado obtido seja relevante para o objeto desejado. Isso pode ser alcançado de diversas maneiras, nesse trabalho optou-se pela utilização de termos separados por operadores lógicos com a seguinte estrutura: *(<campo de busca> = ((<termo principal 1> OR <termo principal 2> OR ... <termo principal N>) AND (<termo auxiliar 1> OR <termo auxiliar 2> OR ... <termo auxiliar M>))) AND <delimitador 1> AND <delimitador 2> AND ... <delimitador P>*

O parâmetro <campo de busca> utiliza o atributo "Tópico" (*Topic*), que engloba quatro características dentro das publicações científicas no WOS e duas no DWPI. No WOS tem-se as seguintes características investigadas: título (*title*), resumo (*abstract*), palavras chave do autor (*author's keyword*), palavras chave especiais (*special keywords*). No DWPI somente o título e o resumo são considerados durante a varredura. Os termos principais e auxiliares são resultados dos estudos feitos anteriormente, no qual se especificou as principais palavras utilizadas dentro do campo de atuação. É importante ressaltar que em ambos os repositórios a busca deve ser feita na alternativa avançada, pois essa opção permite maior flexibilidade na confecção dos filtros de busca (expressões lógicas) utilizadas. Na opção "básica" de ambos os repositórios as buscas só podem ser feitas com filtros pré-estabelecidos, o que limita muito as opções/qualidade de busca.

Os delimitadores servem para direcionar a busca feita para algum alvo em específico. Um exemplo de delimitadores pode ser os países a serem investigados e também a língua em que esses resultados devem ser apresentados. Tanto WOS quanto o DWPI possuem delimitadores distintos para essa tarefa. No WOS utilizou-se como delimitador o parâmetro país, representado pela sigla "CU", e o parâmetro língua, representado pela palavra "*language*". No DWPI utilizou-se o parâmetro de número de patente (*Patent Number-PN*) o qual o formato possui a seguinte estrutura <XXZPTO>. Os primeiros dois caracteres representam o código do país onde a patente foi registrada. Devido a uma restrição de cinquenta caracteres no campo de busca no DWPI, foi necessário desmembrar o filtro inicial utilizado no WOS, de forma a manter sua integridade. Em outras palavras, para um único filtro utilizado no WOS foram necessários cinco filtros no DWPI, alterando-se apenas os termos auxiliares utilizados. Isso fez mais ruídos fossem gerados na amostra de dados coletadas do DWPI.

4.1.2 Selecionando o formato de exportação dos dados

É importante escolher como e em que formato os resultados serão exportados. Essa tarefa influencia todos os passos envolvendo o seu processamento. Tanto o WOS e o DWPI possuem opções com características específicas para exportar os seus resultados. No WOS existem três passos que guiam o processo de exportação dos resultados. No primeiro escolhe-se a quantidade de publicações a serem exportadas: todas que se encontram na página, todas selecionadas ou um período determinado pelo usuário. No entanto, devido a uma limitação do repositório somente é possível exportar 500 resultados por vez (essa limitação também está presente no DWPI). No segundo passo, escolhe-se a granularidade dos resultados, podendo variar desde algumas informações como título, fonte e resumo, até o seu registro completo contendo todas as informações do resultado em adição as referências citadas. No ultimo passo escolhe-se o formato em que os resultados são exportados, esses podem ser: html, *bibtex*, utf-8, *Windows*, Mac, entre outros. Para conseguir um conjunto que pudesse ser aproveitado por diferentes ferramentas de análise o formato escolhido foi o de texto plano (*plain text*).

A única diferença entre o WOS e DWPI nessa etapa encontra-se no passo dois. Para o DWPI tem-se a opção de obter os registros com o número da patente, título, cessionários (*assignees*) e inventores em adição ao resumo, ou o registro completo. Ao escolher salvar o resultado, a ferramenta disponibiliza um arquivo de extensão .txt (texto) com o seguinte nome: savedrecs.txt. Esse arquivo possui os registros obtidos com o filtro de busca.

4.2 Pré-Processamento dos dados obtidos

A etapa de pré-processamento dos registros obtidos consiste em preparar os dados para que esses possam ser processados de maneira efetiva e que os ruídos sejam minimizados. As operações para contornar esses "ruídos" devem compreender, entre outros, os seguintes aspectos: i) padronização dos valores dos atributos, ii) remoção de registros duplicados, iii) tratamento e eliminação de ruídos e iv) tratamento de valores ausentes.

4.2.1Estruturação dos conjuntos obtidos

Cada repositório possui um conjunto de características que descrevem os dados armazenados neles. Todo conjunto ou subconjunto obtido através da utilização do filtro na forma de arquivos possui, para cada registro, diversas entradas como: nome do autor, inventor, tipo de registro, ano de publicação, referências citadas, resumos, entre outros. Devido à limitação na exportação do conjunto de registros é possível que se tenha inúmeros arquivos, contendo diversos registros. No entanto,

deseja-se consolidar todo o conteúdo de uma só vez. Para isso, é necessário trabalhar com um único arquivo, contendo todos os registros indexados.

Desse modo, antes de inserir todos os registros em um único arquivo, é necessário converter esses subconjuntos em um formato que se possa trabalhar. O intuito dessa atividade (realizada através *software* livre *Bibexcel*,) é transformar os arquivos com extensão .txt em arquivos com extensão .doc, que facilita a consolidação e o processamento da informação obtida. Uma vez aberto, dentro do programa deve-se navegar pela estrutura de arquivos até o diretório no qual se encontram os arquivos savedrecs.txt obtidos. Essa atividade deve ser repetida até que todos os sub-arquivos com os resultados possuam a extensão .doc. No anexo A há um tutorial que explica esta tarefa.

4.2.2 Consolidação dos resultados

Dependendo do filtro de buscas utilizado a quantidade de arquivos com extensão .doc pode ser maior do que um. Desse modo, ainda como atividade de pré-processamento, é necessário consolidar todas as ocorrências em um único arquivo. Para adicionar todos os conteúdos de todos os arquivos em um único seleciona-se os arquivos .doc desejados. Para selecionar mais do que um arquivo, bastar manter a tecla "*ctrl*" pressionada. Esses arquivos são mostrados no canto superior esquerdo do programa, logo ao lado da árvore de diretórios. Em seguida, é necessário inserir um nome e extensão para o arquivo resultado dessa operação. Isso deve ser feito logo abaixo do campo "*Type new file name here*" no canto direito do programa. Em seguida, clica-se em *File->Append all selected files to another*, e também no OK que aparece após a operação.

Essa etapa fará com que um arquivo, aqui nomeado de "consolidado.doc", seja criado no diretório raiz do programa. Esse arquivo possuirá em seu conteúdo todas as entradas dispostas nos arquivos separados. Com os dados consolidados em somente um único arquivo é necessário editá-lo para que as entradas "EF" dos subconjuntos de dados sejam excluídas, mantendo somente a ultima ocorrência. Essa entrada delimita o fim do arquivo. Então, deve-se mantê-la somente para o ultimo registro do conjunto total de dados. Caso contrário pode-se ter um processamento limitado ou deficitário. Terminada essa etapa existe agora um arquivo único que pode ser explorado a fim de obter informações relevantes sobre os dados coletados.

4.3 Módulo de processamento dos dados obtidos

Uma vez que o conjunto ou subconjunto estejam prontos, o processamento de todas as informações coletadas pode ser iniciado. Esse processamento é feito em dois grandes grupos para cada repositório: o primeiro chamado de referência simples e o segundo de referência cruzada. No primeiro, observa-se a frequência de um determinado atributo em relação às ocorrências para o conjunto de dados extraído. No

segundo grupo analisa-se a correlação entre dois atributos em relação à frequência dentre os registros obtidos. Somente o atributo "referências citadas", do conjunto obtido pelo WOS, precisou passar por um processamento personalizado. Essa característica possui algumas especificidades em relação à utilização do Bibexcel que necessitam de uma tarefa de processamento personalizada.

4.3.1Extração de referências simples

Para analisar as características classificadas como simples dentro do arquivo obtido pelo WOS e pelo DWPI escolheram-se os seguintes atributos, por serem essenciais para a análise de dados referente à CTI.

WOS: Autores (AU); Referências citadas (CR); Tipo de documento (DT); Tipo de publicação (PT); Editores (PU); Agências de financiamento (FU); Categoria de assunto (SC); Nome de publicação (SO); Ano de publicação (PY).

DWPI: Inventores (AU); Código de classe *Derwent*- Campo de pesquisa (DC); Cessionários (AE); Número primário de aquisição *Derwent*– Ano de registro (GA); Número de patente (PN).

Uma vez selecionado o atributo, é necessário processar o arquivo.doc consolidado para se obter a distribuição frequência desse entre os registros contidos. No entanto, antes de processar a frequência propriamente dita uma atividade intermediária se faz necessária. Nessa, obtém-se o segundo arquivo mais importante para o *Bibexcel*, de extensão .out, que assim como o de extensão .doc, atua como base para muitas outras sub atividades. A etapa de confecção do arquivo .out contendo o atributo que se deseja verificar por registro é a mesma, tanto no repositório WOS quanto no DWPI. As diferenças entre os dados de um e de outro residem na natureza dos atributos estudados e seus valores. Há no DWPI um atributo GA, (Número primário de aquisição *Derwent*) que mostra o ano que a patente ou processo foi registrado. Porém, há uma peculiaridade desse atributo que precisa ser trabalhada individualmente. Para obter-se o arquivo base .out ver tutorial (ANEX0) B e atributo GA (ANEXO C).

4.3.2 Extração de referências cruzadas

O processo de obtenção das referências cruzadas pode ser classificado como uma tarefa mais complexa em relação às referências simples, que levam em consideração apenas um atributo. Nesse processo, correlacionam-se atributos diferentes do mesmo documento entre si. Mesmo com poucos registros, essa tarefa é de difícil conclusão e a ajuda de um programa externo se mostra necessária. Tal programa pode ser feito em qualquer linguagem que convenha ao usuário, desde que esse consiga correlacionar informações entre si.

Para iniciar o processo de referências cruzadas é necessário ter os arquivos base (.out) dos atributos que se deseja correlacionar. A ideia é transformar esses arquivos em um formato que o programa desenvolvido

consiga manipulá-los. Desse modo, transformou-se esses arquivos com a extensão .out em arquivos do tipo .csv, nos quais os campos são separados pelo caractere ";". Essa transformação pode ser feita utilizando um editor de planilhas como: *Open Office, BrOffice, Excel*, etc. Uma vez aberto o arquivo .out basta salvá-lo no formato .csv usando como separador o ";".

Em seguida, é necessário correlacionar as ocorrências desses registros em conjunto com o índice dos mesmos, fazendo com que dois atributos distintos de um mesmo registro possam ser processados. A ideia nesse tipo de extração é de evidenciar quando dois atributos ou mais ocorrem ao mesmo tempo.

Feito isso, o arquivo resultante, nesse artigo chamado de br.autores.anopublicacao.txt, possui o mesmo formato base dos arquivos com extensão .out e pode ser utilizado pelo Bibexcel no intuito de processar a distribuição de frequência para essa nova característica.

4.3.3 Limpeza dos dados

A limpeza dos dados pode ser feita tanto no conjunto consolidado quanto em seus subconjuntos. Os dados que estão no WOS e no DWPI já se encontram em um formato no qual os ruídos são mínimos. No entanto, é importante ter em mente que a tarefa de limpeza de dados propiciará um resultado mais refinado e fiel, por isso sua explicação precede as etapas seguintes. Uma vez que o arquivo .out seja obtido é possível remover eventuais duplicadas que possam aparecer durante o processamento dos dados e gerar um novo arquivo .out. (ANEXO E) para realizar esse procedimento.

4.4 Programas de processamento auxiliares

Ao longo da realização desse trabalho foram detectadas algumas limitações do software *Bibexcel* em relação ao: i) processamento de atributos cruzados e i) formato dos registros de patentes do DWPI quanto à característica "número primário de aquisição *Derwent* (GA)". Para sanar essas lacunas foram desenvolvidos dois programas na linguagem de programação JAVA chamados de *ReadWriteBuffer* e *FixDate*. Esses programas foram registrados no Instituto de Propriedade Intelectual, porém seu uso é livre. Mais informações podem ser obtidas no Anexo D.

4.5 Visualização dos dados

A visualização dos dados é a tarefa que apresenta as informações extraídas dos dados coletados e processados até o momento. A exibição dos dados, nesse trabalho, é feita de duas maneiras: através de tabelas (contendo os detalhes sobre as características e situações analisadas) e através da construção de mapas de visualização. As tabelas resultam diretamente das etapas apresentadas nas seções anteriores. No entanto, os mapas precisam de algumas etapas adicionais e programas específicos

para sua construção. Nesse artigo, os mapas de visualização têm como objetivo ilustrar as correlações entre as citações referenciadas dentro dos registros das publicações cientificas extraídas do WOS.

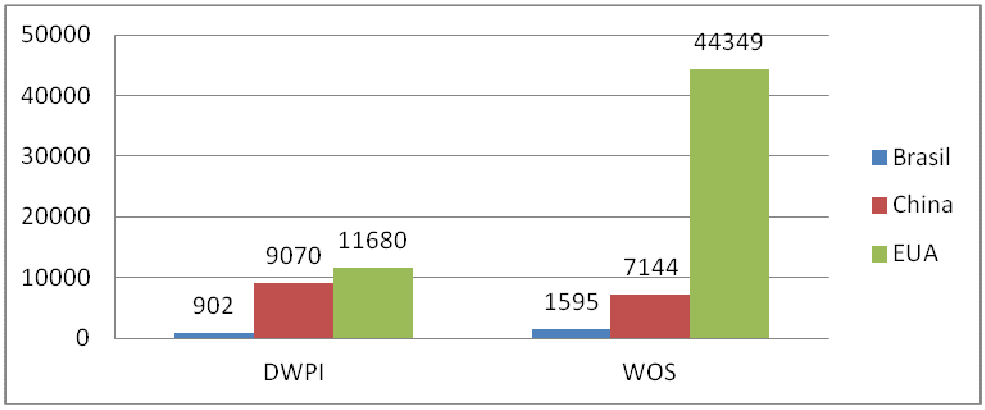
4.5.1Construção de mapas de visualização

Para a visualização dos mapas que trazem as redes de colaboração entre os trabalhos publicados e seus autores, foi utilizado um software livre chamado *Pajek*. Esse processo foi dividido em duas sub-tarefas: a primeira concentra-se no processamento dos dados feito pelo Bibexcel e a segunda foca na visualização dos dados exportados para o *Pajek.*

4.6 Módulo de apresentação e análise

Essa seção apresenta de forma sintética o tipo de informação que será obtida após a aplicação do processo proposto no artigo. Embora ela se chame análise, não haverá discussão do caso estudado por não ser esse o escopo do artigo. Os experimentos apresentados foram realizados para testar o processo. Os dados apresentados referem-se ao campo do conhecimento chamado de biofotônica. Após aplicar os filtros sob os repositórios apresentados foi obtido um conjunto de dados representativos para os países em questão: Brasil, China e EUA. Para o WOS, que abrange as publicações científicas, o Brasil apresentou 1595 registros, a China 7144 e os EUA 44349. Para o DWPI, no qual se encontram os dados referentes às publicações técnicas, o Brasil apresentou 902 registros, a China 9070 e os EUA 11680. A Figur 2 apresenta um resumo dos dados obtidos.

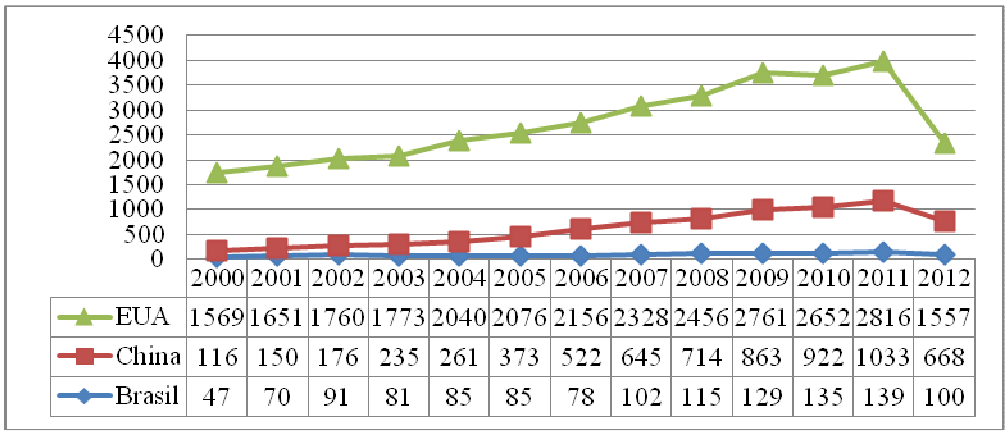
Figura 2 - Quantidade de registros obtidos nos repositórios WOS e DWPI para o período selecionado



Fonte: Elaborado pelos autores

A figura 3 apresenta o atributo simples país e o ano de publicação. Os dados obtidos através dos números de registro - mostrados na Figura 2- contemplam publicações desde 1963 até 10 de setembro de 2012, que é parcial, pois os dados foram coletados ainda no início do terceiro quartil do mesmo.

Figura 3- Relação de quantidade de publicações por ano entre Brasil, China e EUA.

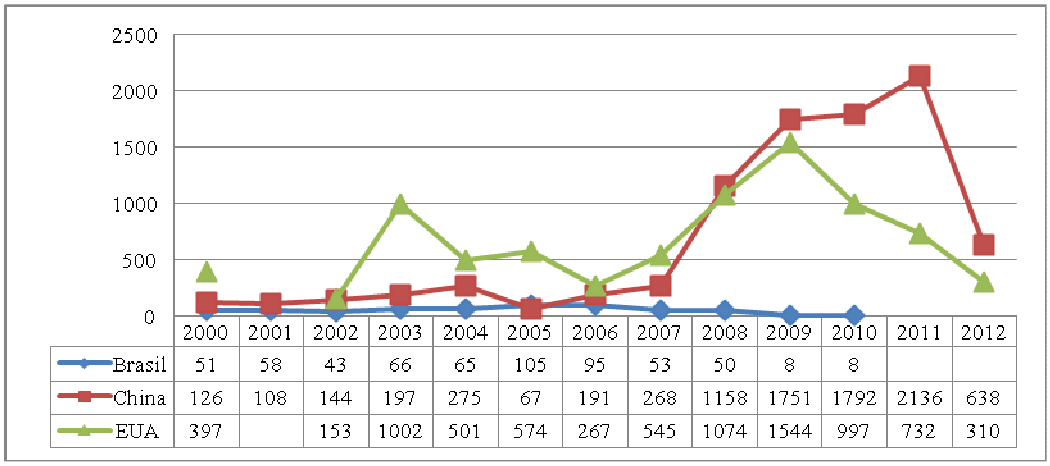


	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
EUA	1569	1651	1760	1773	2040	2076	2156	2328	2456	2761	2652	2816	1557
China	116	150	176	235	261	373	522	645	714	863	922	1033	668
Brasil	47	70	91	81	85	85	78	102	115	129	135	139	100

Fonte: Elaborado pelo autores

A figura 4 mostra um atributo simples para a base DWPI, a qual retrata a relação da quantidade de patentes registradas nos últimos doze anos no Brasil, China e EUA.

Figura 4 - Quantidade de patentes por ano entre Brasil, China e EUA.

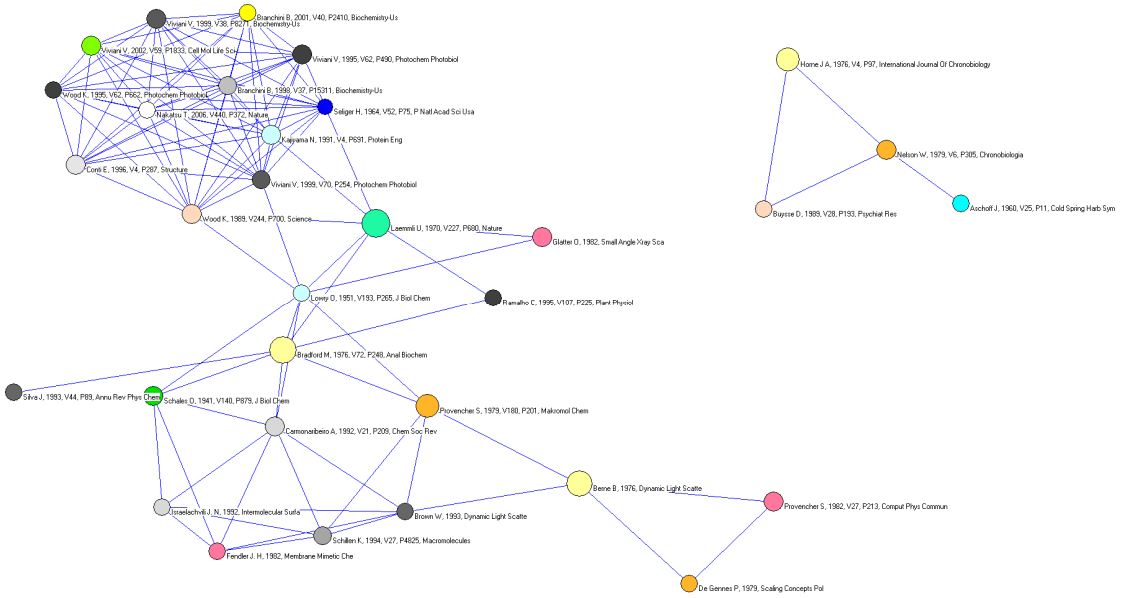


	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Brasil	51	58	43	66	65	105	95	53	50	8	8		
China	126	108	144	197	275	67	191	268	1158	1751	1792	2136	638
EUA	397		153	1002	501	574	267	545	1074	1544	997	732	310

Fonte: Elaborado pelos autores

Por fim, apresentamos os mapas de visualização para detectar as redes de colaboração entre os autores mais citados em Biofotônica no Brasil, China e EUA. O Anexo F mostra em detalhes as etapas necessárias para a construção das redes de colaboração obtidas. A Figura 5 mostra uma rede brasileira bastante esparsa chegando a apresentar até mesmo uma sub-rede, isolada do restante dos demais.

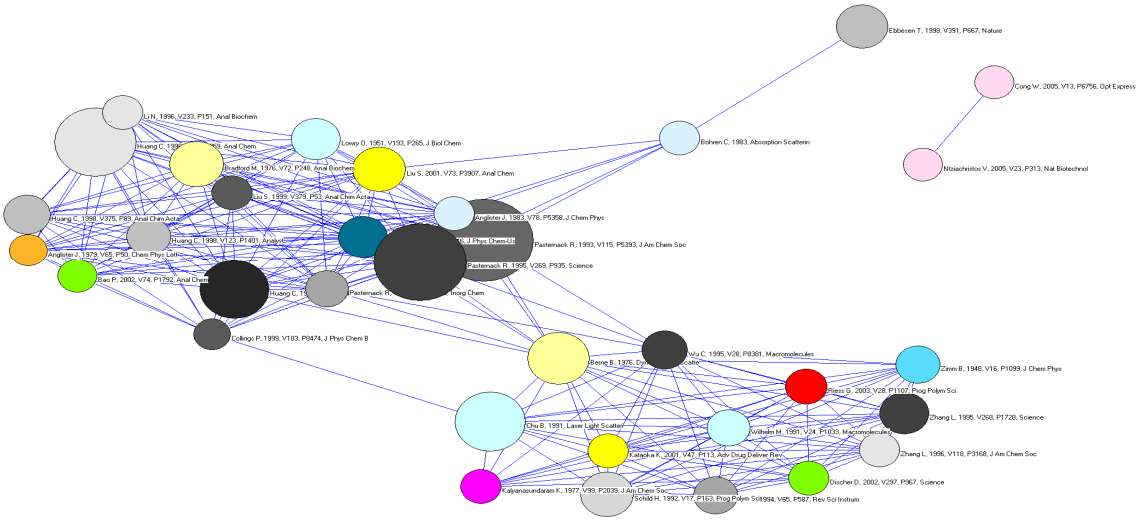
Figura 5 - Rede de colaboração entre autores e trabalhos citados em Biofotônica no Brasil para publicações científicas.



Fonte: Elaborada pelo autor

Na China, o mapa construído apresenta um perfil diferente em relação ao Brasil. Na Figura 6 é mostrada uma rede mais concentrada e com blocos mais definidos, além de ser evidente que o número de ocorrências por registros acompanhou o número de registros do país, uma vez que as dimensões dos círculos também são maiores se comparadas com o Brasil.

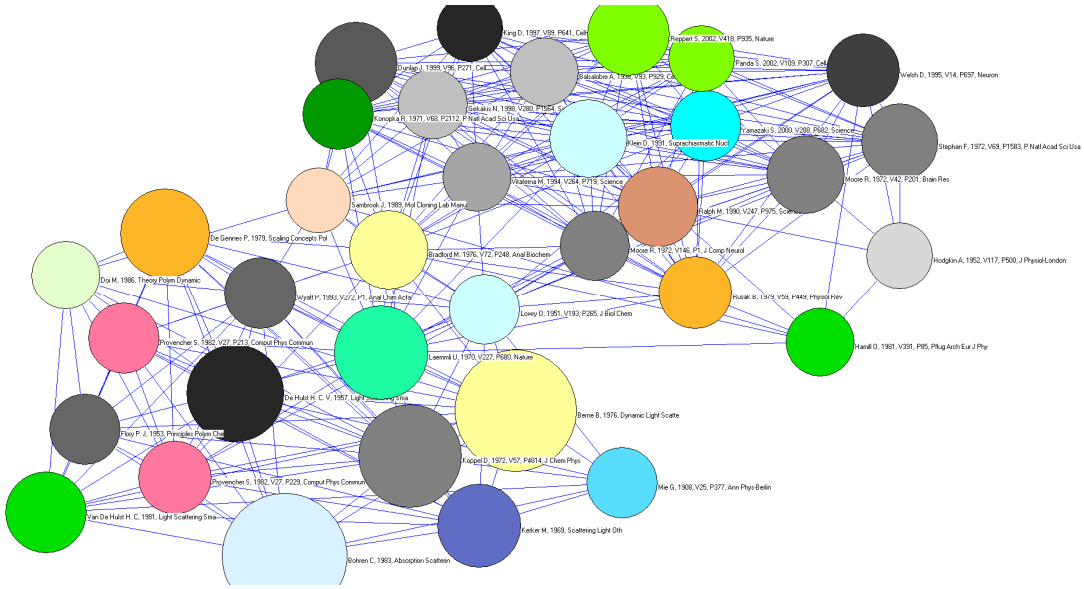
Figura 6 - Rede de colaboração entre autores e trabalhos citados em Biofotônica na China para publicações científicas.



Fonte: Elaborado pelos autores

A Figura 7 apresenta a rede de colaboração formada pelos registros extraídos dos EUA. Pode-se visualizar que os EUA apresentaram uma composição mais densa que Brasil e China. Enquanto que o perfil brasileiro é formado por nós pequenos, a China possui nós mesclados (pequenos, médios e alguns grandes), a rede norte-americana apresenta uma distribuição mais homogênea.

Figura 7 - Rede de colaboração entre autores e trabalhos citados em Biofotônica nos EUA para publicações científicas.



Fonte: Elaborado pelos autores

5 Considerações finais

A quantidade de dados que hoje é produzida mostra-se muito maior do que a capacidade de processamento dos mecanismos tradicionais. Um dos grandes desafios de uma pesquisa cienciométrica é justamente explorar o processo de recuperação de informação das bases de dados através de ferramentas computacionais, do começo ao fim. Por isso, o principal objetivo deste artigo é alcançado ao apresentar de forma didática um processo que auxilie os gestores da inovação sem conhecimento avançados em computação a recuperar e a organizar uma grande quantidade de dados em qualquer área do conhecimento usando softwares livres.

Hoje é possível identificar duas frentes principais na area da cientometria: (i) na análise de funcionalidades das ferramentas bibliométricas (programas de processamento e visualização de dados) e (ii) nos estudos de comportamentos e tendências de um objeto de estudo frente a um cenário escolhido, mas com informações já processadas. No entanto, nenhum deles efetivamente explica como o processo de obtenção da informação é realizado. Além disso, embora estudos cientométricos têm sido construídos, os mesmos têm sido criticados por sofrerem sérias limitações, tais como o período avaliado, falta de comparação seja em relação às diferentes áreas do conhecimentos, seja em relação à regiões

ou países, assim como as próprias especificidades do conhecimento inter e intra áreas.

Numa análise geral, este artigo explica de forma detalhada as principais etapas necessárias para operacionalizar um estudo cienciométrico, o que significa um avanço em relação aos trabalhos encontrados na literatura. Nessa estrutura, os procedimentos necessários à sua aplicação foram agrupados em torno de três elementos constitutivos: Exportação dos Dados dos Repositórios, Processamento e Visualização dos Dados e Apresentação e Análise da Informação. Para realizar esse processo, é preciso considerar uma série de etapas. Dentre essas podemos elencar: i) familiarização com o estudo de caso, ii) especificação dos repositórios de dados pertinentes, iii) extração dos dados, iv) limpeza, v) pré-processamento, vi) normalização, vii) processamento, reprocessamento, maturação dos dados, exportação, viii) visualização e ix) análise dos mesmos.

Todas essas atividades foram condensadas de modo a culminarem no processo de extração e visualização proposto. Através da pesquisa conduzida foi possível elaborar um processo que utiliza *softwares* livres de processamento e visualização de dados para base de dados bibliométricos e de patentes. Desse modo, entende-se que o processo possui atributos capazes de dar um suporte metodológico para àqueles que almejem construir uma estrutura que permita a aplicação nas mais diversas áreas do conhecimento. O resultado será a construção de indicadores de CTI mais robustos. Além disso, o processo foi elaborado de maneira a permitir sua escalabilidade e portabilidade, dependendo do objeto de estudo escolhido e dos componentes que o integram. Aplicado em todas as etapas do tratamento de dados, o modelo atua como uma ponte entre os estudos conduzidos por (i) e (ii).

Durante a pesquisa percebeu-se a necessidade de entender e explicar cada componente do processo e como cada um corrobora para que o resultado final possibilite uma análise mais criteriosa. Desde a escolha e familiarização do objeto de estudo até a apresentação das informações obtidas, todas as etapas são essenciais e precisam ser executadas de modo a possibilitarem uma conclusão sólida. O estudo cienciométrico feito também mostra a importância de se analisar mais do que um indicador e como o cruzamento de indicadores enriquecem a pesquisa. A apresentação através de gráficos e mapas de colaboração ilustram de maneira sistêmica como a rede se comporta e como atributos, antes implícitos, são importantes para consolidar o conhecimento.

Em estudos futuros seria interessante contemplar outros repositórios, além do WOS e o DWPI, possibilitando que uma pesquisa comparativa entre ambos. Aplicando o mesmo filtro de busca nas diferentes bases de dados resultaria em conjuntos de dados distintos, enriquecendo a análise dos mesmos. De maneira análoga, esse cenário pode ser aplicado às ferramentas de processamento e visualização de dados utilizados. Processar e analisar o mesmo conjunto de dados por

aplicações distintas permitirá obter resultados complementares entre si, tornando as informações obtidas mais robustas.

Outra oportunidade de aplicar o processo proposto seria através de uma análise de fontes de dados não estruturadas como: jornais, blogs, portais web, sites de notícias, entre outros. Todas essas fontes não estão sujeitas ao processo rigoroso de seleção que WOS e DWPI estão sujeitas. Então, conduzir um estudo cienciométrico tendo como base repositórios estruturados e não estruturados pode abrir novas possibilidades quanto à validade entre os conteúdos dos mesmos. Tendo em vista a quantidade de informações armazenadas nos repositórios, seria interessante estudar as redes colaborativas entre os principais pesquisadores levando em consideração a posição geográfica de sua filiação e como isso influencia no tipo e classificação de estudo conduzido.

Agradecimentos

Este artigo é resultado de projetos de pesquisa financiados pelos processos nº 2011/14745-2, 2010/12119-4; 2009/10039-6 Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

ANEXO A

Arquivo base .doc

Após selecionado o arquivo .txt desejado segue-se as seguintes instruções:

Selecionar Edit doc file->Replace line feed with carriage return . Responder sim fará com que o arquivo te extensão.txt original seja mantido e um .txt2 seja gerado.

Misc->Convert to dialog format->Convert from Web of Science . Um arquivo de extensão .doc será gerado. Basta responder OK na janela apresentada.

Essa atividade deve ser repetida até que todos os sub-arquivos com os resultados possuam a extensão .doc.

ANEXO B

Arquivo base .out

Para obter-se o arquivo base .out segue-se as seguintes etapas:

Selecionar o .doc já convertido;

Clicar em "View file";

No campo "old TAG" inserir o atributo que se deseja analisar (e.g. AU, PY, DT, etc);

No campo direito, acima de "the box" selecione o separador que é considerado/utilizado para o atributo escolhido

Clique em "Prep";

Será gerado um arquivo com extensão .out com as ocorrências da "old tag" (atributo) escolhida em cada registro dentro do arquivo base selecionado.

A etapa de confecção do arquivo .out contendo o atributo que se deseja verificar por registro é a mesma, tanto no repositório WOS quanto no DWPI.

É importante ressaltar que cada atributo possui uma característica específica de formatação. Por exemplo, o nome dos autores de um artigo publicado ou de inventores de uma patente, possuem seus nomes e sobrenomes separados por espaços em branco. Enquanto que, os nomes de diversos autores e inventores são separados entre si através de um ";". Nesse caso usar a opção de espaços em branco ("*Blank-separated words (e.g. title)*") como um separador (item (d)) pode não ser uma alternativa interessante. Por outro lado, usar a opção de qualquer caractere como separador ("*Any ; separated filed*") pode trazer informações mais reais dos dados analisados.

Além de cada atributo possuir um separador distinto é importante mencionar que para cada característica analisada perguntas relacionadas ao seu tratamento serão feitas antes que arquivo com extensão .out seja concluído.

Após obter o .out com as informações de cada registro e também o seu respectivo valor(es) pode-se trabalhá-lo a fim de obter a frequência do dado atributo. Para isso seguem-se as seguintes etapas:

Selecionar o arquivo .out contendo o atributo que se deseja rastrear a frequência

Clicar em "View file"

Em "Frequency distribution Select type of unit" Selcionar "Whole String"

Selecionar a opção "Sort descending" (para ordenar em ordem decrescente);

Clicar em "Start"

Será gerado um arquivo com extensão .cit com frequência das ocorrências da "old tag" (atributo) escolhida e discriminada no arquivo .out.

ANEXO C

Especificações do atributo GA – DWPI

As diferenças entre os dados de um e de outro residem na natureza dos atributos estudados e seus valores. No DWPI existe um atributo que, devido a seu formato característico, precisou ser trabalhado especificamente. O atributo GA (Número primário de aquisição *Derwent*) mostra o ano que a patente ou processo foi registrado. No entanto, esse campo possui o seguinte formato:<ano>--<identificador>. Graças a esse identificador não foi possível extrair a distribuição de frequência desse atributo diretamente. Para tal objetivo, construiu-se um programa na linguagem de programação JAVA, que utiliza o arquivo de extensão .out (contendo o atributo GA e seu valor completo), o processa de maneira a manter o número de registro o qual essa entrada pertence e o associa com o ano contido no formato original. Essa tarefa faz uso de recursos como *arrays, strings* e *substrings* a fim de ter como produto final um arquivo similar o .out original. Esse arquivo mantém somente o <ano> do formato original para o campo GA. Fazendo com que o valor do atributo GA passe de <ano>--<identificador> para <ano> e assim através dos

passos ilustrados no Anexo 2 também é possível processar sua distribuição de frequência.

ANEXO D

Programas Auxiliares

Por conta de limitações de produto, o Bibexcel não trabalha de maneira satisfatória quanto: (i) ao processamento de atributos cruzados e com (ii) o formato dos registros de patentes do DWPI quanto à característica "número primário de aquisição Derwent (GA)" Então, construiu-se dois programas visando atuar nessas lacunas (i e ii), chamados de *ReadWriteBuffer* e *FixDate* respectivamente.

Ambos os programas foram desenvolvidos na linguagem de programação JAVA. Em (i) trabalha-se com dois arquivos extraídos do *Bibexcel* contendo apenas as entradas referentes a um determinado atributo, e o seu respectivo índice (que representa a qual registro esse pertence). Em seguida esses arquivos são comparados e quando existe a relação de igualdade entre os índices um novo arquivo é alimentado contendo o registro do índice comparado e o valor do mesmo, agora representado pela combinação dos dois valores dos arquivos iniciais.

Em (ii) utiliza-se apenas um arquivo de entrada, responsável por armazenar todos os registros referentes ao atributo "número primário de aquisição Derwent (GA)", que contém a informação de quando (ano) a patente foi registrada. Esse programa utiliza o arquivo exportado pela ferramenta de processamento (*Bibexcel)*, extrai somente os dados pertinentes para que se tenha a informação desejada (data de registro da patente) e constrói um novo arquivo contendo tal informação que será processado normalmente.

ANEXO E

Limpeza de dados – Pré processamento

As etapas seguintes ilustram uma das maneiras de refinar os dados obtidos através de sua limpeza:

Selecionar o arquivo .out/oux/cap/etc desejado;

Clicar em "View file";

Em "Frequency distribution Select type of unit" Selcionar "Whole String"

Marcar as opções "Remove duplicates" e "Make new out file"

Clicar em "Start";

Selecionar "OK" na janela que será aberta;

Será gerado um arquivo com extensão .oux com o mesmo conteúdo do arquivo .out, mas com as duplicatas removidas.

O arquivo gerado (.oux) pode ser trabalhado da mesma maneira que o sua origem (.out). Em outras palavras, todas as funções disponíveis pelo programa *Bibexcel* disponíveis para o arquivo .out continua

m válidas para o arquivo .out.

ANEXO F

Redes de colaboração/ Mapas de colaboração

Para a construção das redes de colaboração utilizam-se as seguintes etapas:

Selecionar o arquivo .doc já convertido;

Clicar em "View file";

No campo "old TAG" inserir o atributo que se deseja analisar (nesse caso CD);

No campo direito, acima do item "the box" selecione o separador que é considerado/utilizado para o atributo escolhido. Para esse utiliza-se o "Any ; separated field"

Clique em "Prep";

Será gerado um arquivo com extensão .out com as ocorrências da "old tag" (atributo) escolhida em cada registro dentro do arquivo base selecionado.

Selecionar o .arquivo .out gerado;

Clicar em "View file";

Em "Edit out-files" seleciona-se "Keep only author's first initial". É gerado um arquivo com extensão .1st contendo o mesmo conteúdo do do arquivo .out, mas agora mostrando apenas a primeira letra dos nomes do atuores;

Seleciona-se o arquivo .1st gerado;

Clicar em "View file";

Em "Edit out-files" seleciona-se "Convert Upper Lower Case" > "Good for Cited reference strings". É gerado um arquivo com extensão .low fazendo com que todas as entradas do arquivo .1st tenham o mesmo padrão;

Seleciona-se o arquivo .low gerado;

Clicar em "View file";

Em "Frequency distribution Select type of unit" Selcionar "Whole String"

Marcar as opções "Remove duplicates" e "Make new out file"

Clicar em "Start";

Selecionar "OK" na janela que será aberta. Será gerado um arquivo com extensão .oux com o mesmo conteúdo do arquivo .low, mas com as duplicatas removidas.

Selecionar o arquivo .oux contendo a correlação de atributos que se deseja rastrear a frequência;

Clicar em "View file";

Em "Frequency distribution Select type of unit" Selcionar "Whole String";

Selecionar a opção "Sort descending" (para ordenar em ordem decrescente);

Clicar em "Start". Será gerado um arquivo com extensão .cit com frequência da correlação dos atributos contidos no arquivo .oux;

Selecionar o arquivo .cit gerado e clicar em "View file";

Na parte direita inferior do programa, no campo "The List" selecionar as ocorrências que deseja-se mapear. Para isso basta clicar no registro e selecioná-lo. Se for necessário selecionar mais de uma ocorrência basta segurar a tecla "Ctrl" no teclado e escolher outras ocorrências. Para esse exemplo escolheu-se as primeiras 15 entradas.

Clicar na opção "Copy" que fica a direita do campo "The List";

Clicar me "Clear" e clicar novamente dentro do campo "The List";

Clicar em "Paste", que se encontra ao lado do botão "Copy". Isso fará com que os registros copiados sejam colado no campo "The List";

Selecionar o arquivo .oux;

Clicar em "Analyze" > "Co-occurance" > "Make pairs via listbox". Certifique-se de responder "no" para a pergunta que aparecer. Será gerado um arquivo de extensão .coc contendo as co-ocorrências entre as entradas selecionadas e todas as outras entradas.

Ao termino dessas etapas todos os pré-requisitos básicos para a preparação dos dados foram feitos. A partir desse ponto inicia-se o processo de construção dos arquivos que serão utilizados pelo Pajek. Para isso citam-se os seguintes passos:

Selecionar o arquivo . coc gerado;

Clicar em "View file";

Em "Mapping" selecionar "Create .net file for Pajek". Quando a pergunta sobre arcos direcionados for exibida clique em "no" uma vez que não se está utilizando esse tipo de métrica. Na segunda pergunta sobre a troca de valores basta responder "yes". Será gerado um arquivo com extensão .net e um com extensão .net. A partir desse ponto já é possível trabalhar a visualização dos dados em si. No entanto, para um modelo mais robusto é necessário ainda algumas tarefas;

Selecionar o arquivo de extensão .cit o qual escolheu-se as entradas para a elaboração do arquivo com extensão .coc;

Clicar em "View file";

Em "Mapping" selecionar "Create .vec file". Será gerado um arquivo de extensão .vec;

Selecionar novamente o mesmo arquivo .cit e em "Edit out-files" seleciona-se "Extract publication year from references". Será gerado um arquivo com extensão .dpy contendo as datas de publicação das entradas do arquivo .cit;

Selecionar o arquivo .dpy gerado;

Clicar em "View file";

Em "Mapping" selecionar "Create .clu file". Será gerado um arquivo de extensão .clu;

Nesse momento todos os arquivos base necessários para a visualização dos dados pelo programa PAJEK foram concluídos. Agora basta importá-los vec, .vet, .net e .clu) no programa para que a rede de colaboração possa ser construída. Ainda no PAJEK na opção "Networks" clicar no ícone do diretório (a esquerda do disquete) e selecionar o arquivo de extensão .net gerado pelo Bibexcel. Em "Partitions" e "Vectors" faça o mesmo processo, mas para os arquivos de extensão .clu e .vec respectivamente. Em seguida na opção "Draw" executa-se a tarefa "Draw-Partition-Vector" para a visualização da rede de colaboração.

Referências

BATAGELJ, V.; MRVAR, A. *Pajek Wiki*, 2008. Disponivel em: <http://pajek.imfm.si/doku.php>. Acesso em: 15 ago. 2012.

BARÍLAN, J. Citations to the "Introduction to informetrics" indexed by WOS, Scopus and Google Scholar. *Scientometrics*, , Budapest, Hungira, v. 82. , n. 3 , p. 495-506, 2010.

CAMPBELL, D. *et al*. Bibliometrics as a performance measurement tool for researchevaluation: the case of Research Funded by the National Cancer Institute of Canada. Disponivel em:*American Journal of Evaluation*, N/A, v. 31, n. 1, p. 66-83, mar. 2010.

CHEN, C. CiteSpace. *Visualizing Patterns and Trends in Scientific Literature,* 13 Setembro 2003. Disponivel em: <http://cluster.cis.drexel.edu/~cchen/citespace/>. Acesso em: 20 ago. 2012.

CHEN, C. Searching for intellectual turning points: Progressive knowledge domain visualization. *In:* NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, 2004. *Proceedings…* [S. l.]; [s. n.], 2004. of the p. 5303-5310.

CHEN, C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, Publicado online,v. , n. , p. 359--377, 2006.

COBO, M.J.; LÓPEZ-HERRERA, A.G.; HERRERA-VIEDMA, E.; HERRERA, F. Science mapping software tools: Review, analysis, and cooperative study among tools*. Journal of the American Society for Information Science and Technology,* New York, USA, v. 62 , n. 7 , p.1382-1402, 2011.

COCCIA, M. A scientometric model for the assessment of scientific research performance within public institutes. *Scientometrics*, Budapest, Hungira, v. 31, n. 1v. 65, n. 3, p. 307-321, 2005.

ECK, N. J. V.; WALTMAN, L. *VOSviewer.* Welcome to the VOSviewer web site*,* 2009. Disponivel em: <http://www.vosviewer.com/>. Acesso em: ago. 2012.

ECK, N. J. V.; WALTMAN, L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics,* N/A, v. 84, n. 2, p. 523--538, 2010.

FALAGAS, M. E; PITSOUNI, E. I; MALIETZIS, G. A; PAPPAS, G. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *The FASEB Journal,* N/A, v. 22, n. 2 338-342, 2008.

FAYYAD, U. Data Mining and Knowledge Discovery: Making Sense Out of Data. *IEEE Expert*: Intelligent Systems and Their Applications, Piscataway, USA, v. 11, n. 5, p. 20-25, Outubro 1996.

GANTZ, J.; REINSEL, D. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. Framingham, USA , IDC iView: IDC Analyze the Future, 2012.

GLÄNZEL, .W. On reliability and robustness of scientometrics indicators based on stochastic models. An evidence-based opinion paper. *Journal of Informetrics* N/A, v. 4, n. 3, p. 313-319, 2010.

JAPAN PATENT OFFICE. *Introduction to Patent Map Analysis, 2011*. Disponivel em: <http://www.training-jpo.go.jp/en/uploads/text_vtr/pdf/Introduction%20to%20Patent%20Map%20Analysis2011.pdf**>.** Acesso em: 25 out. 2012.

LEYDESDORFF, L. *Communication and Innovation in the Dynamics of Science & Technology, University of Amsterdam, N/A.* Disponivel em: <http://www.leydesdorff.net/software.htm>. Acesso em: ago. 2012.

MARCO, A. The Dynamics of Patent Citations. *Economics Letters*, N/A, v. 94, n. 2, p. 290-296, 2007.

MEARIAN, L. World's data will grow by 50X in next decade, IDC study predicts. *Compter World*, N/A, junho, 2011. Disponivel em: <http://www.computerworld.com/s/article/9217988/World_s_data_will_grow_by_50X_in_next_decade_IDC_study_predicts>. Acesso em: 03 jul. 2012.

MOURA, A. M. M.; CAREGNATO, S. E. Co-autoria em artigos e patentes: um estudo da interação entre a produção científica e tecnológica. Perspectivas em Ciência da Informação, Belo orizonte,Horizonte, Horizonte, v. 16, n. 2, p. 153-167, 2011.

NWB TEAM. Network Workbench. A Workbench for Network Scientists, N/A, 2006. Disponivel em: <http://nwb.slis.indiana.edu>. Acesso em: 15 ago. 2012.

OLIVEIRA, E.; GRACIO, M. C. Indicadores bibliométricos em ciência da informação: análise dos pesquisadores mais produtivos no tema estudos métricos na base Scopus. *Perspectivas em Ciência da Informação*, Belo Horizonte, v. 16, n. 4, p. 16-28, 2011.

PERSSON, O.; DANELL, R.; SCHNEIDER, J. W. *Celebrating Scholarly Communication Studies.* A Festscrift for Olle Persson at his 60th Birthday. [S. l.]: The Authors, 2009.

PONNIAH, P. *DataWarehousing Fundamentals: A Comprehensive Guide for IT Professionals*. [S. l.]: AWiley-Interscience Publication, 2001.

RAAN, A. F. J. V. The interdisciplinary nature of science: theoretical framework and bibliometric-empirical approach. *In*: WEINGART, P.; STEHR, N. *Practising Interdisciplinarity.* Toronto: University of Toronto

Press, 2000. p. 66-78. Disponivel em: http://www.cwts.nl/TvR/documents/AvR-PractInterdisc.pdf. Acesso em: 03 jul. 2012

RAAN, A. F. J. V. Measurement of central aspects of scientific research: performance, interdisciplinarity, structure. *Measurement*: Interdisciplinary Research and Perspectives, Leiden, Netherlands, v. 3, n. 1, p. 1-19, 2005.

SCI² TEAM. Sci² Tool. *A Tool for Science of Science Research & Pratice*, 2009. Disponível: ,https://sci2.cns.iu.edu/user/index.php> . Acesso em: 20 ago. 2012.

RUAS, T.L ; PEREIRA, L. "*ReadWriteBuffer*": Registro Programa de Computador. Número: BR512014000077-0,: 31 jan. 2014, Instituto Nacional da Propriedade Industrial.

RUAS, T.L ; PEREIRA, L. "*FixDate*": Registro Programa de Computador. Número: BR512014000078-8, 31 jan. 2014, Instituto Nacional da Propriedade Industrial.

ZITT, M.; BASSECOULARD, E. Challenges for scientometric indicators: data demining, knowledge-flow measurements and diversity issues. *Ethics in Science and Environmental Politics*, Luhe, Germany,v. 8, n. 1 p. 49-60, 2008.