DOI: 10.5433/1981-8920.2014v19n3p30

EXTRAÇÃO E TRATAMENTO DE DADOS NA BASE LATTES PARA IDENTIFICAÇÃO DE CORE COMPETENCIES EM DENGUE

IDENTIFICACIÓN DE LAS COMPETENCIAS FUNDAMENTALES EN DENGUE A TRAVÉS DE INVESTIGACIÓN EN LA BASE DE DATOS LATTES

Jorge Magalhães – jorgemagalhaes@far.fiocruz.br
Pós-doutor Recherche en Sciences de l'Information et de la Communication pela Aix Marseille Université, France. Pesquisador da Fundação Oswaldo Cruz (FIOCRUZ).

Luc Quoniam – mail@quoniam.info
Doutor EM *Science de l'information* et de la Communication pela *Université Aix Marseille III*. Professor titular da *Université Du Sud Toulon Var,* França.

Jesús Mena-Chalco - jmenac@gmail.com
Pós-doutor em Ciência da Computação pela Universidade de São Paulo (USP). Professor da Universidade Federal do ABC (UFABC).

André Santos – amsantos@univali.br
Mestre em Administração pela Universidade Federal do Rio Grande do Sul (UFMG). Professor da Universidade do Vale do Itajaí (UNIVALI).

RESUMO

Introdução: A conjuntura de competitividade global requer das organizações práticas cada vez mais ousadas de Inteligência Competitiva, a fim de obter, analisar, tratar e disseminar as informações para auxiliar na tomada de decisão. No ambiente de Big Data na Web, é premente cautela para o resgate e análise de dados a fim de transformá-los em informações essenciais para o gestor.

Objetivo: Identificar e extrair a produção científica, produtos tecnológicos, instituições, redes dos cientistas que trabalham com a doença Dengue.
Metodologia: Usa-se a bibliométrica com técnicas para mensurar a produção e disseminação do conhecimento científico. Analisar os 2,5 milhões de currículos da base Lattes do CNPq, extrair e tratar os que possuem o termo "dengue" com o software ScriptLattes do conceito Web 2.0.
Resultados: A identificação de 15.465 currículos específicos com o tema dengue. Extraiu-se 424 cientistas renomados na área, bem como mais 971 colaborações nacionais e internacionais relacionados com o termo dengue. O método possibilitou extrair dos especialistas a geolocalização, publicações, orientações, dentre outros.
Conclusões: A análise dos resultados mostrou a relação multidisciplinar dos cientistas em Dengue. O método pode ser replicado para qualquer área da ciência. A bibliometria como fonte de auxílio no tratamento de Big Data mostrou-se eficaz através da ferramenta ScriptLattes.

Palavras-chave: Estudos métricos da informação. Bibliometria. Análise da produção científica. ScriptLattes. Base Lattes. Inteligência competitiva. Dengue.

1. INTRODUÇÃO

O conhecimento produzido pela comunidade científica ao longo do tempo pode ser encontrado em diversas bases de dados científicas. No âmbito internacional, um exemplo é a *Web of Science* (ISI), considerada uma das maiores e mais reconhecidas bases de registros científicos, como a ScienceDirect e outras. No Brasil, como exemplo pode-se citar a Scielo, considerada como a maior base da América Latina. Essas bases de dados registram as produções dos cientistas para auxiliar na divulgação do avanço da ciência para a humanidade, permitindo explorar as relações entre ciência, tecnologia e a exploração industrial (DOSI; LLERENA; LABINI, 2006).

Nesse sentido, as bases de dados disponibilizam acesso à informação cientifica como referências aos especialistas ou aos artigos científicos. Elas são chamadas de informações secundárias, pois somente indicam informações individuais do manuscrito (*paper*), e às vezes, quando disponível em termo de *copyright*, permitem acesso ao *paper* completo. Observa-se que essas bases ou são temáticas (abrangem só uma área da ciência) ou generalistas (várias áreas da ciência). Nota-se ainda, que algumas bases incluem citações, haja vista que a mesma é útil ao modelo sociológico da ciência, onde a "boa ciência" se constrói com "boas fundações" (citações), além de sempre haver um custo de produção.

Embora exista uma grande variedade de bases de dados, cada uma possui particularidades e limitações que dificultam obter uma informação abrangente e interdisciplinar sobre determinado assunto científico. Algumas são importantes, pois permitem referenciar a produção científica de uma área, mas falham em relacionar pesquisadores, instituições e atividades de pesquisa. Muitas declaram abranger todas as áreas da ciência, porém quando comparadas a bases temáticas de abrangência internacional, como a *Medline* para a medicina, o *Scifinder* para química, Inspec para física, o *Compendex* para a engenharia, o *Metadex* para Materiais, etc, demonstram-se inferiores em termos de resultados e informações sobre o assunto em determinado país. Enfim, é bastante difícil construir um mapa integrado do conhecimento acerca de uma área científica, pois as informações encontram-se dispersas em diferentes bases e com diferentes propósitos. Porém, no Brasil, o que existe é uma base de dados que tem se aproximado do conceito de base integrada, reunindo pesquisadores, atividades, produção científica e instituições envolvidas com a ciência, no qual configura-se um grande banco de dados de currículos e que, intrinsicamente, possui uma excelente referência à produção bibliográfica em cada currículo elencado na base (WALLACE; LARIVIÈRE; GINGRAS, 2012).

No Brasil, investimentos na área da ciência tem se tornado cada vez maior (MASSARANI, 2013). Nesse âmbito o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) criou a base de dados Lattes, que se distingue de outras bases, principalmente por ter uma característica curricular. Essa particularidade se deve ao depósito compulsório do currículo por parte de todos os cientistas brasileiros no banco de dados da mesma. Esses currículos são constituídos com todos os dados pessoais e profissionais do pesquisador, além de toda a produção bibliográfica (*papers*, patentes, congressos, orientações de alunos etc.) que o mesmo construiu durante sua carreira científica. Além disso, essa base tem por premissa a atualização constante quando da conclusão de um novo trabalho pelo pesquisador. Dessa forma, o governo Brasileiro disponibiliza todos os dados *online*, para consulta sem restrição. Não obstante, a base Lattes proporciona extrações de dados das instituições aos quais os pesquisadores são filiados. Portanto, em síntese, é uma ferramenta para gestão contábil dos currículos.

Se por um lado a plataforma Lattes é uma das mais completas bases de dados sobre pesquisadores, produção científica e instituições de ciência no Brasil, por outro a forma tradicional como estas informações podem ser acessadas e utilizadas restringe enormemente o seu uso. Atualmente, a pesquisa pode ser realizada por palavra(s) chave(s), mas não há a possibilidade de identificar onde a mesma se encontra dentro do currículo. A visualização dos currículos é permitida e demonstra a produção cientifica do pesquisador em questão, contudo não é possível buscar somente a produção do especialista, bem como o tipo de produção que ele realizou em determinado período. Essa produção científica está apresentada em forma de um "bloco" para cada artigo, se tornando difícil o uso dessa referência para posteriores tratamentos, ou montagem de uma base "local" de referências cientificas. Por conseguinte, mesmo esta base sendo robusta em termos de produção cientifica brasileira, ela é considerada somente como uma base de metadados sobre a produção científica brasileira, e desta forma, essas características levam o Brasil a ficar dependente da consideração dos artigos produzidos nas bases de dados internacionais. Este fato se agrava quando a produção científica é feita na língua portuguesa, onde a mesma ainda é pouco considerada nas bases internacionais.

Nesse sentido, o trabalho versa sobre a relevância da plataforma brasileira Lattes e propõe uma metodologia para extração e tratamento dos dados existentes na mesma, a fim de identificar as competências essenciais e respectivas correlações com cientistas nacionais e internacionais, além de resgate automatizado das suas produções científicas. O modelo de Competências Especiais adotado é o de Prahalad e Hamel (1997), onde definem Competências Especiais como o conjunto de habilidades e tecnologias que habilitam uma companhia a proporcionar um benefício particular para os clientes, mais do que uma habilidade ou tecnologia. Essas *core competencies* são oriundas das competências organizacionais, que por sua vez, são obtidas pela existência das competências pessoais de indivíduos especialistas (*core*) (PRAHALAD, 1990; PRAHALAD; HAMEL, 1997)

O êxito nessa identificação e posterior mapeamento se torna uma importante ferramenta de gestão para os tomadores de decisão das organizações (MAGALHÃES; QUONIAM; BOECHAT, 2013). Isso se intensifica quando aplicado à saúde pública, onde nações em conjunto com a Organização Mundial de Saúde (OMS) almejam melhores condições de vida à suas populações através de

programas de saúde, inserção de novos medicamentos, fomento à pesquisa etc. (MOREL et al., 2009; ZHANG et al., 2013). Assim, para melhor compreensão do propósito, elencamos um caso prático de levantamento e tratamento da produção cientifica dasCompetências Especiais na doença negligenciada (DN) dengue.

1.1 Acerca da Doença Dengue

A escolha dessa doença foi concebida por múltiplas razões, dentre elas, por ser endêmica no Brasil e por pertencer ao grupo de DN prioritárias consideradas na Agenda Nacional de Prioridades de Pesquisa em Saúde da Secretaria de Ciência, Tecnologia e Insumos Estratégicos (SCTIE) do Ministério da Saúde Brasileiro (BRASIL, 2008). A figura 1 mostra aproximadamente 100 países onde a dengue pode ser encontrada, envolvendo as Américas, África, Ilhas do Pacífico, Ásia e Mediterrâneo (BRASIL, 2008; WORLD HEALTH ORGANIZATION, 2013a). Considerada como doença grave e potencialmente letal, a Organização Mundial de Saúde estima que cerca de 2,5 bilhões de pessoas correm o risco de contraí-la. O número de infecções aumentou drasticamente nas últimas décadas devido ao aumento da urbanização, o comércio e as viagens em todo globo terrestre (DYE et al., 2013).

Figura 1 - Distribuição de dengue no mundo

Fonte: World Health Organization (2013b).

Os maiores problemas e desafios no controle da dengue são a inexistência de vacinas, áreas extensas de disseminação do mosquito, conhecimento científico insuficiente para a redução das populações do vetor, problemas na detecção e

notificação precoce dos casos da doença e a fragilidade da integração entre a vigilância entomológica e a vigilância epidemiológica. Nesse âmbito, se torna fundamental a integração dos sistemas de pesquisa e serviços de saúde pública (BALES et al., 2011).

Dada essa relevância, optou-se pelo estudo de caso nessa enfermidade com o uso de fontes abertas (ALLARAKHIA; AJUWON, 2012). Assim, a obtenção de dados na base Lattes sobre "dengue" seria relativamente simples, ao contrario de outras áreas científicas. Isto decorre em razão de que a única palavra "dengue" é suficiente por si só, ou seja, não apresenta ambiguidade nem variações no idioma da literatura científica internacional, por exemplo, em inglês ou francês ela permanece "dengue". Deste modo, a pesquisa não será dependente de outros erros quando se foca o levantamento das CE.

Por outro lado, extrair e analisar dados da base Lattes não é trivial. A base consiste em um conjunto de sistemas de informação, bases de dados, *data warehouses*, portais e sistemas de conhecimento voltados ao mapeamento das competências nacionais e das ações de fomento em CT&I.

1. 2 Acerca da Plataforma Lattes

A plataforma Lattes representa a experiência do CNPq na integração de bases de dados de currículos, dos grupos de pesquisa e de instituições em um único sistema de informações. Sua dimensão atual se estende não só às ações de planejamento, gestão e operacionalização do fomento do CNPq, mas também de outras agências de fomento federais e estaduais, das fundações estaduais de apoio à ciência e tecnologia (C&T), das instituições de ensino superior e dos institutos de pesquisa. Além disso, se tornou estratégica não só para as atividades de planejamento e gestão, mas também para a formulação das políticas do Ministério da Ciência, Tecnologia e Inovação e de outros órgãos governamentais da área de ciência, tecnologia e inovação (CNPQ, 2013).

Os currículos extraídos dessa plataforma constituem-se em Currículo Lattes e se tornaram um padrão nacional no registro da vida pregressa e atual dos estudantes e pesquisadores do país. Além do que é adotado pela maioria das instituições de fomento, universidades e institutos de pesquisa do Brasil. Por sua riqueza de informações e sua crescente confiabilidade e abrangência, é considerado

elemento indispensável e compulsório à análise de mérito e competência dos pleitos de financiamentos na área de C&T (MENA-CHALCO; CESAR JUNIOR, 2009; MENDONÇA; SANTOS, 2004).

Desde a criação da Plataforma Lattes, ela oferece um estudo de caso interessante nos currículos por, principalmente, três motivos: (i) os currículos Lattes tornaram-se um padrão para a avaliação das atividades acadêmicas e profissionais de pessoas atuantes no Brasil, (ii) a maioria dos pesquisadores brasileiros estão registrados na mesma – em Julho/2013 foram identificados quase 2,5 milhão de currículos cadastrados no seu banco de dados, e (iii) cada currículo incorpora uma lista de produtividade acadêmica e profissional, por exemplo, de produções bibliográficas, técnicas e artísticas/culturais, projetos de pesquisa, supervisões acadêmicas e participação em bancas de defesa de mestrado ou doutorado (CNPQ, 2013).

O currículo é auto declaratório e não é requerido que as publicações registradas sejam indexadas pelo Scopus, Web of Science, SciELO ou qualquer outra base científica. Por estas razões, a Plataforma Lattes é uma fonte abrangente para análise de várias áreas do conhecimento.

Cabe destacar a potencialidade da base, quando se observa o registro de cerca de 1,2 milhões de currículos com nível de doutorado ou superior, representando mais de 20 milhões de publicações científicas (MENA-CHALCO et al., 2013). Por outro lado, inúmeras outras informações valiosas podem ser mapeadas para a C&T, quando, por exemplo, se almejam obter dados sobre orientações realizadas ou em andamento pelos cientistas, projetos de pesquisa, redes de parcerias, patentes etc. Nesse sentido, é possível dizer que esta base é exaustiva em termo de pesquisadores ativos em território brasileiro, haja vista a possibilidade de cadastrar currículos de cientistas estrangeiros.

Em outro aspecto, a base se torna ainda mais robusta, por considerar que o cientista pode solicitar financiamento em qualquer Agência de fomento para suas pesquisas, desde que esteja cadastrado na base Lattes. Portanto, se torna muito difícil uma atividade qualquer de pesquisa institucional sem o cadastro no currículo Lattes. Assim, pode-se concluir que esse sistema é considerável em termo de representar a “ciência brasileira”.

O sistema de financiamento às pesquisas instituído no Brasil, tanto no ensino como na pesquisa, desenvolvimento e inovação é concorrido (ANTUNES; MAGALHAES, 2008; COSTA; GADELHA; MALDONADO, 2012). Ele está avaliado através de um sistema de avaliação "por pares", onde os currículos tem um papel imprescindível. Esta situação potencializa a representatividade de informações científicas dentro do Lattes Com essa diretriz, o Brasil provoca, compulsoriamente, que cada pesquisador mantenha seu currículo atualizado no sistema e, assim, a base Lattes se torna cada vez mais dotada de um sistema de informação valiosa ao longo do tempo.

O sistema funciona na concepção da "Web 2.0" (QUONIAM; LUCIEN, 2010), ou seja, funciona de modo "participativo" - cada pesquisador mantém seu próprio currículo e como a base é aberta, qualquer pessoa pode consultar o currículo de uma ou várias pessoas. Este trabalho mostra que o Lattes é importante para produzir dados com valor agregado em cima dela própria. Por esse princípio, a base se torna uma parte do "Big data" (LYNCH, 2008; MCAFEE; BRYNJOLFSSON, 2011), onde permite fazer trabalhos similares aos que são feitos no ramo do "data-jornalismo" (GRAY; CHAMBERS; BOUNEGRU, 2012). Cada currículo e identificado com um número único de 16 dígitos (nomeado ID Lattes), o que facilita a identificação única dos pesquisadores.

Diante de toda positividade da base Lattes, ela é contra balançada com lados negativos. De certa forma, há um controle dos currículos no momento de avaliação por pares, ou seja, quando o pesquisador solicita algum financiamento de alguma agência de fomento, a Agência submete seu projeto e seu currículo a avaliação de seus pares. Nesse momento, não há controle nem das informações que foram inseridas pelo cientista, nem quanto à digitação dos dados. É evidente, que se o avaliador puder demandar um tempo maior de análise, poderá tentar buscar/descobrir cada informação contida no currículo acessando os Periódicos em que lá estão inseridos, bem como as bases de patentes, os Grupos de Pesquisa, instituição, etc. Contudo, há que considerar que seria um trabalho extremamente árduo.

Esta realidade demanda certa fragilidade à base Lattes em termos de qualidade, porém, diante da robustez que a mesma apresenta, ela se torna mais relevante em termo de representatividade de uma amostra consistente de toda a

produção da ciência Brasileira do que se pode encontrar em bases indexadas internacionais.

1.3 Acerca do Diretório de Grupo de Pesquisa da Plataforma Lattes

Nesta outra base da plataforma Lattes, se encontra o Diretório dos Grupos de Pesquisa no Brasil, que pode ser considerado como um inventário dos grupos de pesquisa em atividade no país. Dentre as informações constantes na base, identificam-se os recursos humanos constituintes dos respectivos grupos, as linhas de pesquisa propriamente dita, os setores de atividades envolvidos, as especialidades do conhecimento, a produção científica, tecnológica e artística e os padrões de interação com o setor produtivo. Os grupos estão localizados em instituições de ensino superior, institutos de pesquisa, etc.; e as informações individuais de cada participante são extraídas de seus próprios currículo. No que tange às atualizações, elas são realizadas continuamente pelos líderes de grupos, pesquisadores, estudantes e dirigentes de pesquisa das instituições participantes e o CNPq realiza censos bianuais (CNPQ, 2013).

O DGP foi concebido para promover às organizações do Sistema Nacional de C&T e Inovação à condição de usuárias da Plataforma Lattes. Ele registra todas e quaisquer organizações ou entidades que estabelecem algum tipo de relacionamento com o CNPq. A disponibilização pública dos dados da Plataforma na internet dá maior transparência e mais confiabilidade às atividades de fomento do CNPq e das agências que a utilizam, fortalecem o intercâmbio entre pesquisadores e instituições, e é fonte importante de informações para estudos e pesquisas de cientometria. Na medida em que suas informações são recorrentes e cumulativas, além de proporcionar importante papel de preservar a memória da atividade de pesquisa no país (CNPQ, 2013).

O Diretório de Grupos de Pesquisa contém dados sobre um “reagrupamento” de pesquisadores que possuem currículo e, portanto, trabalham em conjunto nessas linhas de pesquisas definidas nesse grupo de pesquisa.

2 OBJETIVO

O artigo tem como objetivo experimentar e demonstrar uma metodologia para o levantamento do Curriculum Vitae de pesquisadores denominados como core competencies em determinado assunto (PRAHALAD, 1990; PRAHALAD; HAMEL, 1997). Dessa maneira, tornando-se possível resgatar os currículos, tratá-los posteriormente e ainda oferecer vários desdobramentos, tais como geoposicionamento dos especialistas, identificação das instituições e também indicadores bibliométricos da produção científica das CE.

3 METODOLOGIA

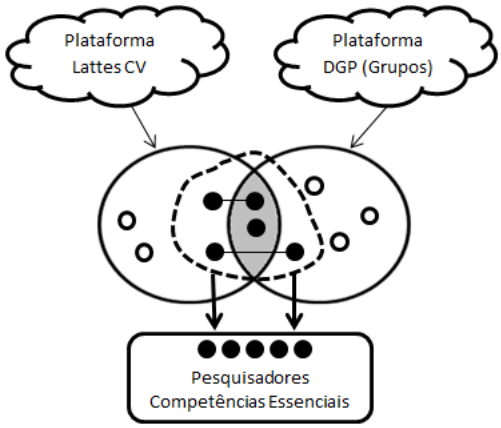
O levantamento de dados curricular por assunto não é uma tarefa trivial e implica no uso de diferentes ferramentas e processos para conseguir resultados válidos. Portanto, esse artigo descreve a metodologia que desenvolvemos para a busca e análise de currículos em busca de *core competencies.* A metodologia tem como pressupostos:

1) Na base curricularlattes, a busca irá retornar os currículos que contenham pelo menos uma ocorrência do termo de busca. Certamente, isto irá depender das informações que o pesquisador insere no seu currículo e da forma como ele as insere.
2) Na base do Diretório de Grupos de Pesquisa, a busca irá retornar os pesquisadores pertencentes aos Grupos de Pesquisa que contenham pelo menos uma ocorrência do termo de busca na descrição do grupo.
3) Os pesquisadores que formam o núcleo de competências essenciais são aqueles que estão presentes em ambos os resultados dos currículos e Grupos de Pesquisa, acrescidos daqueles pesquisadores presentes em apenas um dos conjuntos mas que colaboram diretamente com algum pesquisador do outro conjunto.

Esses pressupostos, representados na Figura 2, levam a parecer razoáveis para poder garantir o levantamento de competências essenciais de uma

determinada área do conhecimento usando como estudo de caso a Plataforma Lattes do CNPq e o Diretório de Grupos de Pesquisa.

Figura 2 - Pressupostos para competências essenciais.



Fonte: Elaboração própria dos autores.

a. Proposta de modelo para mineração de dados

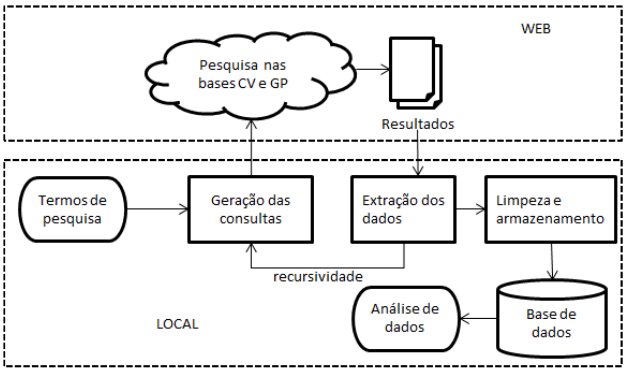
As principais formas de análise de dados presumem uma fonte de dados normalizada e estruturada, geralmente no formato de banco de dados. Porém, os resultados das consultas feitas às páginas de currículos e Grupos de Pesquisa não se encontram desta forma e os métodos tradicionais de análise não podem ser aplicados diretamente a estas fontes de informações (LAENDER; RIBEIRO-NETO; DA SILVA, 2002; LIU; WANG; AGRAWAL, 2012).

Para analisar as informações não estruturadas, como aquelas disponíveis nos resultados de currículos e Grupos de Pesquisa, uma das soluções apontadas pela literatura é o uso de ferramentas específicas para a mineração e análise de informações como, por exemplo, agentes inteligentes e mecanismos de *crawler, mining e scraping* (DE MEO et al., 2013, 2014; DHALL; BHARAT; CHAUDHARY, 2012; MENA-CHALCO; CESAR JUNIOR, 2009; ZHANG et al., 2013). Um sistema mineração de dados na Web é capaz de interagir com uma fonte de dados da Web e extrair os dados armazenados nesta fonte. Estes dados são então filtrados e convertidos para um formato apropriado para análise, geralmente sob a forma de banco de dados (ARASU; GARCIA-MOLINA, 2003; DE MEO et al., 2014; MAGALHAES, et al., 2013).

Uma dificuldade adicional enfrentada ao extrair informações das bases de currículos e Grupos de Pesquisa é a forma dinâmica de recuperação dos resultados. No lugar de páginas estáticas, as informações das bases de currículos e Grupos de Pesquisa são apresentadas de acordo com os argumentos de pesquisa inseridos pelo usuário. A página com os resultados é construída especificamente para responder a pesquisa solicitada e fica disponível apenas de forma temporária. Isto exige uma adaptação das ferramentas de mineração para que possam interagir de forma automatizada com as fontes de informações da web profunda, inserindo argumentos de pesquisa e extraindo o conteúdo das páginas dinamicamente criadas (LIU; WANG; AGRAWAL, 2012; ZHANG et al., 2013).

Dessa forma, o modelo proposto neste trabalho pode ser visualizado sinteticamente na Figura 3. Ele tem início pela entrada dos termos de pesquisa pelo usuário. Com base nos termos solicitados, a ferramenta de mineração utiliza um módulo de pesquisa capaz de gerar consultas específicas para cada uma das fontes de dados da web selecionadas. Uma vez que as respectivas fontes da web forneçam os resultados, o módulo de extração de dados retira as informações relevantes das páginas e as envia para o processo de limpeza e armazenamento. Assim, o processo de limpeza é responsável por eliminar redundâncias e resultados incompletos. Por fim, as informações são estruturadas e gravadas sob o formato de um banco de dados e então analisadas.

Figura 3 - Processo de extração e análise de dados.



Fonte: Elaboração própria dos autores.

Nesse sentido, os métodos utilizados foram desde execução de softwares livres como o scriptLattes (MENA-CHALCO; CESAR JUNIOR, 2009) versão 8.09, o

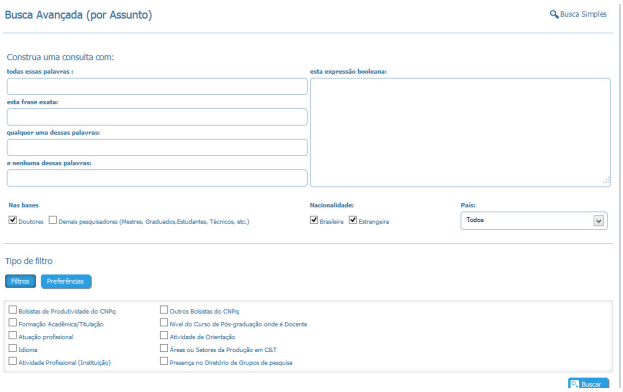
Python para Windows versão 2.7 e o Microsoft Office Excel® 2010 para confecção de planilhas, exclusão de duplicados e cruzamento dos dados, além das bases Lattes e do Diretório de Grupos de Pesquisa do CNPq.

4 RESULTADOS E DISCUSSÃO

a. Experimentação na base Lattes

O sistema de busca para consulta na base possui dois módulos: um simples e um avançado. Na busca avançada está disponível o método de escolha por expressões booleanas, onde há opções de usar operadores como “AND”, “OR”, “NOT”, “NEAR” etc. (até 10 palavras de distância), além de parênteses com a finalidade de combinar esses operadores para uma busca sofisticada.

Figura 4 - Tela da base Lattes do CNPq para busca avançada.



Fonte: CNPq (2013).

As opções de filtragem disponibilizam critérios interessantes, como a possibilidade de restringir a quantidade de resultados ao aumentar a pertinência da busca em questão. Por exemplo, ao escolher a opção para pesquisa, quanto aos currículos Lattes terem doutorado ou não, nacionalidade brasileira e/ou estrangeira, além de diversos outros tipos de filtro como a “presença no diretório de grupos de pesquisa” etc.

Ao efetuar pesquisa nessa tela, aplicando a “busca simples” com o termo “dengue”, puderam-se obter diversos resultados através da aplicação de diferentes critérios. No caso específico da “dengue”, os resultados podem ser considerados semelhantes por busca simples ou avançada, haja vista que a palavra “dengue” se

encontra inserida nos termos de sinonímia da dengue. Dessa forma, foi considerada a "busca simples", onde os resultados podem ser visualizados na tabela 1.

Tabela 1 - Número de cientistas na base Lattes para o termo Dengue.

Bases / Tipo de Filtro escolhidas	**Busca "Dengue"**
Doutores e demais Brasileiros e estrangeiros	15465
Doutores Brasileiros e estrangeiros	4164
Doutores Brasileiros	4077
Doutores Brasileiros e estrangeiros com grupo de pesquisa	3214
Doutores Brasileiros e estrangeiros com grupo de pesquisa e bolsa produtividade (qualquer nível)	568

Fonte: CNPq (2013).

Ao continuar a busca, restringido mais critérios disponíveis na base, não foi possível obter resposta do sistema. Por exemplo: usar "doutores Brasileiros e estrangeiros com nível do curso de Pós-graduação onde é Docente"; "Doutores Brasileiros e estrangeiros com atividade de orientação", o resultado apresentava-se negativo. Esta falha de resposta pode ser oriunda de várias razões, dentre elas o não preenchimento dos dados pelos pesquisadores ou ainda da própria base[1].

Cabe ressaltar que foram identificados 101 pesquisadores caracterizados como "estrangeiros", porém, tal fato pode não corresponder em sua totalidade como verdadeiros. Isto decorre em razão do sistema Lattes somente considerar a naturalidade do pesquisador e não a sua nacionalidade. Nesse sentido, dentre os 101 revelados pela base, pode haver brasileiros nascidos no exterior. Da mesma forma, também não há precisão de onde se encontra o termo buscado no currículo, ou seja, do local onde o pesquisador colocou a palavra em voga na pesquisa; se ela se encontra no título, no nome do periódico, na palavra-chave ou área de pesquisa, bem como não há informação sobre o "peso" dessa palavra-chave dentro do currículo do pesquisador.

Portanto, é possível haver alguma depreciação do levantamento de trabalhos usando a base Lattes, por entender como "ruidosa" a informação no sentido de

[1] Em 12/04/13 enviamos e-mail ao suporte@cnpq.br solicitando auxílio sobre a dificuldade encontrada e nos foi enviado resposta que o problema foi encaminhado à equipe técnica ao qual entraria em contato – fato que não aconteceu até a submissão do presente artigo

clareza dos dados. Por outro lado, esta possível restrição torna-se vantagem quando se deseja ter um olhar holístico ou sistêmico da situação em questão. Por exemplo, quando há necessidade de se avaliar o "ambiente cientifico" do tema de uma determinada área técnico-científica.

Por esta razão, foi efetuada uma segmentação nos dados (por exemplo, um "corte") para análise do tema "dengue" no critério do pesquisador ser doutor, tanto Brasileiro quanto estrangeiro e pertencer a um grupo de pesquisa.

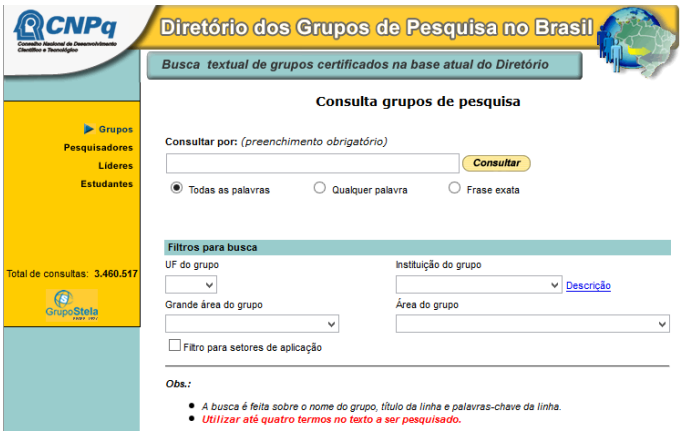
Através desse "corte" resgataram-se 3214 números de Identificação Lattes (ID) dos currículos. Aprofundando a busca, fez-se novo corte dos pesquisadores com grau de doutor, sendo tanto Brasileiro quanto estrangeiro, a fim de obter a diferença que, possivelmente, não estaria ligado a nenhum Grupo de Pesquisa. Nessa etapa foram coletados 4077 para tratamentos de dados posteriores (ver seção 4.3).

b. Experimentação no diretório dos grupos de pesquisa

Igualmente à Plataforma Lattes, o Diretório de Grupos de Pesquisa funciona no modo "Web 2.0", ao qual obedecem as mesmas regras de participação, acesso aberto e a atualização contínua. A interface de busca da base pode ser vista na figura 5, onde foi extratificada em Junho de 2013, quando da realização deste trabalho. Ela proporciona buscar informações por palavras contidas no nome do grupo, título da linha de pesquisa e o termo chave da linha em questão.

Contudo, não há possibilidade de analisar se todos os currículos dos pesquisadores presentes nos Grupos de Pesquisa correspondem à mesma palavra que foi "utilizada" na pesquisa de ID Lattes, haja vista que um Grupo de Pesquisa pode contemplar vários campos de pesquisa ou até incluir membros de outras áreas da ciência por questões de interdisciplinaridade. Esse fato é interessante pela possibilidade de analisar as vertentes interdisciplinares de um determinado "campo cientifico". Da mesma forma, nesta interface do Diretório de Grupos de Pesquisa, ao contrário da base Lattes, não está disponibilizado a opção de efetuar busca por expressão booleana. Assim, dificultando a pesquisa quando for necessário o levantamento de assuntos mais precisos, com o suporte que só os operadores booleanos permitem. No caso deste trabalho, para a dengue não foi necessário o uso de operadores booleanos.

Figura 5 - Tela da base Diretório dos grupos de pesquisa do CNPq.



Fonte: CNPq (2013).

Na tabela 2, se podem observar os resultados da busca com a palavra "dengue" no campo de "consulta a grupos de pesquisa".

Tabela 2 - Número de cientistas em Diretório de Grupos de Pesquisa no Brasil para o termo Dengue

Termo Pesquisado	Grupos encontrados (distintos/únicos)	Função Pesquisadores (distintos/únicos)	Função Estudantes (distintos/únicos)
Dengue	139	1258	1672

Fonte: CNPq (2013).

Como nota-se na tabela 3, há possibilidade de um pesquisador pertencer a vários Grupos de Pesquisa, sendo líder de um grupo e ainda pesquisador ou estar como pesquisador em um grupo e até mesmo estudante em outro. A fim de resguardar a coerência na nossa proposta, o foco ateve-se na busca em pesquisadores.

Tabela 3 - Número de pesquisadores no Diretório de Grupos de Pesquisa no Brasil que estuda Dengue.

Número de Grupos de Pesquisa em que está inserido	Profissionais na função de Pesquisador
1	1152
2	88
3	12
4	5
5	1

Fonte: CNPq (2013).

Os ID Lattes dos pesquisadores presentes nos Grupos de Pesquisa mencionando a palavra "dengue" foram levantados e em seguida retirados os resultados duplicados (presença em vários Grupos de Pesquisa), ao qual se resgatou uma lista de 1258 ID Lattes (ver detalhamento na seção 4.3).

c. Experimentação quanto às "core competencies"

Como observado anteriormente, com ambas plataformas de dados, foi possível levantar dois universos imperfeitos: os currículos que "tem algum registro associado ao assunto ("falam do assunto") e os Grupos de Pesquisa que tem alguma linha de pesquisa associado ao assunto ("trabalham no assunto"). Ao efetuar cruzamento dos dois resultados encontrados, aumentamos a probabilidade de termos um conjunto de pesquisadores que pode representar as Competências Especiais no assunto em questão. Desta forma, aplicando, os 4077 com a palavra no currículo e os 1258 envolvidos no Diretório de Grupos de Pesquisa obtiveram-se 424 currículos em comum.

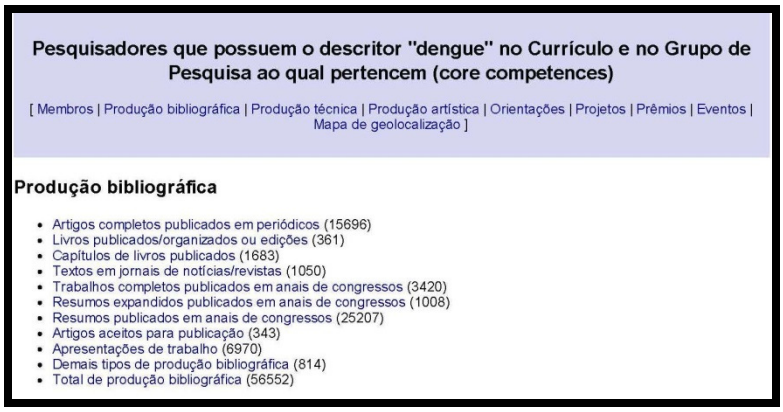
Nessa abordagem, não há como precisar que é a "população total" de Competências Essenciais na doença dengue, mais há garantia de se ter uma amostra mais do que significativa do envolvimento de pesquisadores nesse assunto. Esse fato leva ao interesse em continuar levantando currículos em vez de artigos, proporcionando uma visão mais holística e sistêmica da ciência envolvida ao redor do assunto analisado.

Considerando que, por alguma razão, nem todos os pesquisadores pertençam a um Grupo de Pesquisa, pode-se evidenciar as Competências Especiais estendendo a técnica aplicada. No ramo da ciência, também são consideradas as co-publicações e, às vezes, até um cientista pode ser depreciado por fazer ciência isoladamente. Nesse sentido, ao aplicar o software Scriptlattes nas Competências Especiais obtidas (424) foi possível identificar os colaboradores destas "core competencies", em um total de 10413 e suas respectivas orientações. Conclui-se, portanto, que estes colaboradores podem trabalhar com a doença "dengue" e não estar no Diretório de Grupos de Pesquisa ou vice-versa.

Nessa abordagem, cruzando esta nova lista de colaboradores com a lista inicial de 4077 ID Lattes (aqueles que trabalham na doença dengue e em grupo de pesquisa dengue ou não), foi possível resgatar 1395 pesquisadores que contemplam

pesquisas “essenciais” para com dengue – consideramos como cientistas “core” mais seus colaboradores em primeiro nível. Como exemplo da aplicabilidade da metodologia proposta para identificação de “core competencies”, pode-se observar no quadro 1 e na figura 6, partes da página da web, contendo os resultados dessa rede identificada de Competências Especiais em dengue. Resumo de suas publicações, membros, geolocalização etc.

Quadro 1 - Resultados das competências essenciais em dengue.



Pesquisadores que possuem o descritor "dengue" no Currículo e no Grupo de Pesquisa ao qual pertencem (core competences)

[Membros | Produção bibliográfica | Produção técnica | Produção artística | Orientações | Projetos | Prêmios | Eventos | Mapa de geolocalização]

Produção bibliográfica

- Artigos completos publicados em periódicos (15696)
- Livros publicados/organizados ou edições (361)
- Capítulos de livros publicados (1683)
- Textos em jornais de notícias/revistas (1050)
- Trabalhos completos publicados em anais de congressos (3420)
- Resumos expandidos publicados em anais de congressos (1008)
- Resumos publicados em anais de congressos (25207)
- Artigos aceitos para publicação (343)
- Apresentações de trabalho (6970)
- Demais tipos de produção bibliográfica (814)
- Total de produção bibliográfica (56552)

Fonte: extraído pelos autores através do software ScriptLattes (CNPq, 2013).

Figura 6 - Resultado da geolocalização das competências essenciais em dengue.

Fonte: extraído pelos autores através do software ScriptLattes. (CNPq, 2013).

Os resultados detalhados estão disponíveis em http://vlab4u.info/doencas%20negligenciadas/Dengue/. Cabe ressaltar, que ao aplicar o software, é possível visualização e análise tridimensional, além da extração

detalhada dos currículos de cada Competência Especial desejada e sua respectiva rede de interações.

d. Quantificação dos dados

Após o tratamento pelo ScriptLattes, foram obtidas informações essenciais no campo da Pesquisa, Desenvolvimento e Inovação para a doença dengue, haja vista, a metodologia usada, onde considerou "extrair" as competências essenciais que trabalham com dengue. Nesse sentido, identificado e extraindo seus respectivos trabalhos científicos e tecnológicos, pode-se afirmar que foi gerada uma base referencial brasileira inicial para a doença dengue com "core" bibliografias.

Sinteticamente, seguem as quantidades de referências bibliográficas essenciais para a doença dengue extraídas segundo a metodologia adotada (detalhamento pode ser visualizado na seguinte página web: http://vlab4u.info/doencas%20negligenciadas/Dengue/denguecore1/denguecore1results/index.html):

- Quanto análise científica:

I. Artigos completos publicados em periódicos: 49058;
II. Livros publicados/organizados ou edições: 1291;
III. Capítulos de livros publicados: 6363;
IV. Textos em jornais de notícias/revistas: 5403;
V. Trabalhos completos publicados em anais de congressos: 10348;
VI. Resumos expandidos publicados em anais de congressos: 3840;
VII. Resumos publicados em anais de congressos: 81051;
VIII. Artigos aceitos para publicação: 1178;
IX. Apresentações de trabalho: 25130;
X. Demais tipos de produção bibliográfica: 3115.

Total de produção bibliográfica: 186.777

- Quanto a produção técnica:

I. Produtos tecnológicos: 775;
II. Processos ou técnicas: 744;
III. Trabalhos técnicos: 8189;
IV. Demais tipos de produção técnica: 10530.

Total de produção técnica: 20.238.

5 CONSIDERAÇÕES FINAIS

Em tempos de Big data, são imprescindíveis ferramentas da tecnologia da informação para auxiliar na mineração da quantidade de dados existente e seus respectivos tratamentos. Para a base Lattes não é diferente quando precisamos analisar cerca de 2,5 milhões de currículos e identificar as competências essenciais na área em questão e suas respectivas instituições e/ou produções científicas. Portanto, a busca relacional é importante além da busca existente através do sistema disponível pelo website da Plataforma Lattes.

A base de dados Lattes permite obter informações para buscas simples, com operadores booleanos, porém; buscas relacionais mais complexas não são possíveis. Como a maioria das questões científicas envolve o pensamento relacional, os mecanismos de busca oferecidos pelas bases Lattes do CNPq ainda não proporcionam a possibilidade de extrair de forma gerencial, como oferecido nesta proposta considerando o ScriptLattes, obtendo assim toda robustez nela existente.

Cabe ressaltar que o sistema Lattes enaltece o conceito web 2.0, pois a base é curricular e não referencial. Assim, ela oferece diferentes oportunidades de análise mais holística.

O trabalho aqui apresentado está concentrado mais na metodologia de mineração dos dados, pela forma de exemplificar possibilidades de extração e análises para mapeamento de competências essenciais em determinada área da ciência, bem como identificação das produções científicas dessas competências essenciais. Portanto, futuros trabalhos poderão ser mais focados nos resultados obtidos, de maneira analisar melhor a produção (cientometria, mapeamentos, detecção de pontos fortes e fracos, oportunidade de pesquisa nas áreas, detecção de equipes de especialistas, coautorias dentre outros).

Não obstante, esta pesquisa foi exploratória no sentido de validar uma metodologia para posteriormente replicá-lo em casos mais complexos (envolvendo operadores booleanos na busca e outras Doenças Negligenciadas como, por

exemplo, malária e tuberculose) bem como outras áreas estratégicas como nanomateriais.

Cabe destacar que houve diversas dificuldades na elaboração da metodologia, como a confecção de scripts de automação necessários à pesquisa, dos upgrades constantes dos scripts (são inevitáveis). Dessa forma, mais que pretensão de esgotar o assunto, o trabalho focou na forma de usar um conjunto de programas na forma de código fonte aberto a fim de que outros especialistas pudessem replicá-lo com uma metodologia consolidada para identificar competência essencial.

Ao considerar os resultados científicos e tecnológicos das "core competencies" em dengue obtidas, nota-se que somente em referências bibliográficas e técnicas, foram elencadas 207.015 produções. Nesse sentido, seria relevante para a Ciência & Tecnologia, a criação de uma base referencial em temas específicos, ou seja, áreas e/ou temas prioritários e específicos para o Brasil.

REFERÊNCIAS

ALLARAKHIA, Minna; AJUWON, Larry. Understanding and creating value from open source drug discovery for neglected tropical diseases. **Expert opinion on drug discovery**, London, v. 7, n. 8, p. 643–657, ago. 2012.

ANTUNES, Adelaide; MAGALHAES, Jorge Lima de. **Oportunidades em medicamentos genéricos**: a indústria farmacêutica brasileira. Rio de Janeiro: Interciencia, 2008.

ARASU, Arvind; GARCIA-MOLINA, Hector. **Extracting structured data from web pages**. 2003. Disponível em:<http://doi.acm.org/10.1145/872757.872799>. Acesso em: 7 mar. 2014

BALES, Michael E. et al. Evolution of coauthorship in public health services and systems research. **American Journal of Preventive Medicine**, New York, v. 41, n. 1, p. 112–117, jul. 2011.

CNPQ. **Plataforma lattes**. Disponível em:<http://lattes.cnpq.br/>. Acesso em: 2 ago. 2013.

COSTA, Laís Silveira L. S.; GADELHA, Carlos Augusto Gadelhal; MALDONADO, José. A perspectiva territorial da inovação em saúde: a necessidade de um novo enfoque. **Revista de Saúde Pública**, Rio de Janeiro, n.46, supl., p. 59-67, 2012.

DE MEO, Pasquale et al. Enhancing community detection using a network weighting strategy. **Information Sciences**, New York, v. 222, p. 648-668, fev. 2013.

DE MEO, Pasquale et al. Mixing local and global information for community detection in large networks. **Journal of Computer and System Sciences**, New York, v. 80, n. 1, p. 72–87, fev. 2014.

DHALL, Shivangi; BHARAT, Bharat Bhushan; CHAUDHARY, Swarna. Web link structure mining of the world wide web using hyperlink induced topic search. **MIT International Journal of Computer Science & Information Technology**, Cambridge, v. 2, n. 1, p. 12–15, 2012.

DOSI, Giovanni; LLERENA, Patrick; LABINI, Mauro Sylos The relationships between science, technologies and their industrial exploitation: an illustration through the myths and realities of the so-called “European Paradox”. **Research Policy**, Amsterdam, v. 35, n. 10, p. 1450-1464, dez. 2006.

DYE, Christopher et al. WHO and the future of disease control programmes. **The Lancet**, London, v. 381, n. 9864, p. 413-418, fev. 2013.

GRAY, Jonathan; CHAMBERS, Lucy; BOUNEGRU, Liliana. **The Data Journalism Handbook**.Sebastopol: O’Reilly Media, 2012.

LAENDER, Alberto H. F.; RIBEIRO-NETO, Berthier; DA SILVA, Altigran S. DEByE – Data Extraction By Example. **Data & Knowledge Engineering**, Amsterdam, v. 40, n. 2, p. 121–154, fev. 2002.

LIU, Tantan; WANG, Fan; AGRAWAL, Gagan. Stratified sampling for data mining on the deep web. **Frontiers of Computer Science**, Beijing, v. 6, n. 2, p. 179–196, 1 abr. 2012.

LYNCH, Clifford. Big data: How do your data grow? **Nature**, London, v. 455, n. 7209, p. 28–29, 4 set. 2008.

MAGALHÃES, Jorge Lima; QUONIAM, Luc; BOECHAT, Nubia. Web 2.0 tools for network management and patent analysis for health public. **Revista de Gestão em Sistemas de Saúde**, São Paulo, v. 2, n. 1, p. 26 – 41, 2013.

MAGALHAES, Jorge et al. Neglected Disease In Social Network? A Blueprint of Dengue In Twitter as a contribution of Information Science for Public Health. **International Journal of Management, IT and Engineering (IJMIE)**, Moorebank, v. 3, n. 10, p. 194-204, 2013.

MASSARANI, Laisa. Brazil’s science investment reaches record high. **Nature**, New York, ago. 2013.

MCAFEE, Andrew; BRYNJOLFSSON, Erik. **Big Data**: The management revolution. 2012. Disponível em: <http://hbr.org/2012/10/big-data-the-management-revolution/ar/1>. Acesso em: 7 mar. 2013.

MENA-CHALCO, Jesus Pascual et al. **Brazilian bibliometric co-authorship networks**. Disponível em: <http://professor.ufabc.edu.br/~jesus.mena/publications/pdf/bbcn-jasist-preprint.pdf>. Acesso em: 7 mar. 2013.

MENA-CHALCO, Jesus Pascual; CESAR JUNIOR, Roberto M. Scriptlattes: an open-source knowledge extraction system from the Lattes platform. **Journal of the Brazilian Computer Society**, Rio de Janeiro, v. 15, n. 4, p. 31–39, 1 dez. 2009.

MENDONÇA, Manoel, SANTOS, Celso A. Saibel. Uma abordagem para geração e visualização de mapas de conhecimento. **International Conference on Information Systems and Technology Management**, v. 1, n. 1, 23 jun. 2004.

BRASIL. Ministério da Saúde. **Agenda Nacional de Prioridades de Pesquisa em Saúde**. Brasília, 2008. Disponível em: <http://bvsms.saude.gov.br/bvs/publicacoes/AGENDA_PORTUGUES_MONTADO.pdf>. Acesso em: 23 jul. 2013.

MOREL, Carlos Medicis et al. Co-authorship Network Analysis: a powerful tool for strategic planning of research, development and capacity building programs on neglected diseases. **PLoS Neglected Tropical Diseases**, San Francisco, v. 3, n. 8, p. e501, 18 ago. 2009.

PRAHALAD, C. K. Globalization: the intellectual and managerial challenges. **Human Resource Management**, Belo Horizonte, v. 29, n. 1, p. 27–37, 1990.

PRAHALAD, C.K.; HAMEL, Gary. The core competence of the corporation. In: FOSS, Nicolai (Ed.). **Resources Firms and strategies**: a reader in the resource-Based perspective. New York: Oxford University Press, 1997. p. 235-256.

QUONIAM, Luc; LUCIEN, Armaud. **Intelligence compétitive 2.0** : organisation, innovation et territoire. France: Librairie Lavoisier, 2010.

WALLACE, Matthew L.; LARIVIÈRE, Vincent; GINGRAS, Yves. A small world of citations? the influence of collaboration networks on citation practices. **PLoS ONE**, v. 7, n. 3, p. e33339, 7 mar. 2012.

WORLD HEALTH ORGANIZATION - WHO. Control of Neglected Tropical Diseases. **Sustaining the drive to overcome the global impact of neglected tropical diseases**. Switzerland: WHO, 2013a. Disponível em: <http://www.who.int/neglected_diseases/en/index.html>. Acesso em: 19 jan. 2013.

WORLD HEALTH ORGANIZATION - WHO. Control of Neglected Tropical Diseases. **Sustaining the drive to overcome the global impact of neglected tropical diseases**. Switzerland: WHO, 2013b. Disponível em: <http://www.who.int/neglected_diseases/en/index.html>. Acesso em: 19 jan. 2013.

ZHANG, Chichen et al. Research collaboration in health management research communities. **BMC Medical Informatics and Decision Making**, London, v. 13, n. 1, p. 52, abr. 2013.

Title

Identification of core competencies in dengue through research in lattes database

Abstract

Introduction: The environment of global competitiveness requires the increasingly bold practices of Competitive Intelligence for organizations in order to obtain, analyze, process and disseminate information to assist in decision making. Big Data in the Web provide caution to the rescue, analysis and transform the data into essential information for the managers.
Objective: To identify and extract the scientific production, technological products, institutions, networks of scientists working on the Dengue disease.
Methodology: Use the bibliometric techniques to measure the production and dissemination of scientific knowledge. Analyze the 2.5 million curriculum in Lattes Brazilian database, extract and process all with the word "dengue" using software ScriptLattes through Web 2.0 concept.
Results: Identification of 15465 scientists with Dengue in curriculum database. These were extracted 424 renowned scientists in the field, as well as have over 971 collaboration in national and international level. The method allows to extract the geolocation of experts, publications, advisors, etc.
Conclusions: The results showed the relationship of multidisciplinary scientists in Dengue. It is possible to replicate anywhere area of science. Bibliometrics as a source of aid in the treatment of Big Data was effective through ScriptLattes tool.

Keywords: Metrics studies of information. Bibliometry. Analysis of scientific Production. Scriptlattes. Lattes database. Competitive intelligence. Dengue.

Título

Identificación de las competencias fundamentales en dengue a través de investigación en la base de datos lattes

Resumen

Introducción: El entorno de la competitividad global requiere de las prácticas cada vez más audaces de las organizaciones de inteligencia competitiva con el fin de obtener, analizar, procesar y difundir la información para ayudar en la toma de decisiones. Big Data en el entorno de la Web, es urgente la precaución al rescate y análisis de datos con el fin de transformarlos en información esencial para el gerente.
Objetivo: identificar y extraer los aspectos científicos de los productos tecnológicos, instituciones, redes de científicos que trabajan en la enfermedad del dengue.
Metodología: Se utiliza con las técnicas bibliométrica para medir la producción y difusión del conocimiento científico. Analizar los 2,5 millones de currículos base Brasileña Lattes, luego fueron extraídos y tratados las personas con la palabra "dengue" con el software ScriptLattes con concepto Web 2.0.
Resultados: La identificación de 15.465 currículos con Dengue en el curriculum. Fueron identificados 424 científicos de reconocido prestigio en el campo de ciencia, así como más de 971 colaboraciones de nivel nacional e internacional en dengue. El método hace posible la extracción de la geo localización de expertos, sus publicaciones, supervisores, etc.

Conclusiones: Los resultados muestran la relación de científicos multidisciplinarios en Dengue. El método se puede replicar en cualquier área de la ciencia. La Bibliometria fue efectiva con una fuente de ayuda en el tratamiento de grandes volúmenes de datos a través de la herramienta ScriptLattes.

Palabras clave: Estudios métricos de la información. Bibliometria. Análisis de la producción científica. ScriptLattes. Base de datos Lattes. Inteligencia Competitiva. El Dengue.

Recebido em: 22.04.2014
Aceito em: 16.12.2014