Revista Gestão & Tecnologia
e-ISSN: 2177-6652
revistagt@fpl.edu.br
http://revistagt.fpl.edu.br/

Uma Ferramenta para Apoiar a Seleção de Dados no Processo de Descoberta de Conhecimento em Bancos de Dados de Produção Acadêmica

A Tool to Support Data Selection in Knowledge Discovery Process on Academic Production Databases

Ricardo Antônio Câmara da Silva
Analista judiciário - informática no Tribunal Regional Federal da 3ª Região de São Paulo. Aluno do Programa de Mestrado Profissional em Administração - Gestão de Projetos -MPA-GP da UNINOVE. Graduado em Ciência da Computação pela Universidade Federal de Pernambuco e especialização em Tecnologia da Informação pela mesma universidade,São Paulo, Brasil .
rcamara62@gmail.com

Emerson Antônio Maccari
Doutor em Administração pela Universidade de São Paulo – USP, Estágio Doutoral na University of Massachusetts Amherst – USA. Mestre em Administração pela Universidade Regional de Blumenau - FURB .Graduado em Administração e em Ciências da Computação pela FURB. Especialista em Tecnologia da Informação aplicada à Gestão de Negócios pela FURB/INPG.
Diretor do Programa de Pós-Graduação em Administração PPGA da Universidade Nove de Julho, São Paulo, Brasil
emersonmaccari@gmail.com

Luc Marie Quoniam
Professor Visitante da Universidade Nove de Julho - UNINOVE no Programa de Pós-Graduação em Administração - PPGA. Livre Docente em Ciências da Informação e da Comunicação na Université Aix Marseille III .Doutor em Ciências da Informação e da Comunicação - Université Aix Marseille III Mestre em Oceanologia - Université Aix Marseille II (1985). Graduado em Océanologie - Université Aix Marseille III .Graduado em Química Analítica e Proteção do Meio Ambiente - Université Aix Marseille III.Universidade Nove de Julho, São Paulo, Brasil
mail@quoniam.info

Editor Científico: José Edson Lara
Organização Comitê Científico
Double Blind Review pelo SEER/OJS
Recebido em 28.04.2015
Aprovado em 04.05.2015

RESUMO

A produção da pós-graduação *stricto sensu* brasileira é importante fonte geradora de conhecimento, mas há dificuldades em recuperar dados de maneira direta, para análises sobre a produção de universidades, áreas de conhecimento ou regiões geográficas específicas. O objetivo deste relato técnico é o de construir uma ferramenta que cria listas de entrada automatizadas para recuperar informações e gerar conhecimento sobre a produção científica de docentes da pós-graduação brasileira, por meio do programa ScriptLattes. A implementação da ferramenta proposta possibilita a geração de conhecimentos que podem apoiar análises de produção acadêmica e redes de colaboração de pesquisadores, projetos de pesquisa e desenvolvimento, formação de equipes multidisciplinares, elaboração de políticas e currículos, acompanhamento e avaliação de programas.

Palavras-chave: Gestão do Conhecimento, Pós-graduação, Instituições de Ensino Superior, Descoberta do Conhecimento em Bancos de Dados, ScriptLattes.

ABSTRACT

The scientific production of Brazilian *stricto sensu* post-graduation is an important source of knowledge, but there are difficulties to retrieve data in a direct way, to perform analyzes on the production of specific universities, knowledge areas or geographical regions. This technical report aims at building a tool to create automated entry lists to retrieve information and generate knowledge about the scientific work of Brazilian post graduation professors, using ScriptLattes program. The implementation of the proposed tool will help generate knowledge to support, for example, analysis on academic research and collaborative networks of researchers, research and development projects, building of multidisciplinary teams, policies and curricula development, monitoring and evaluation of programs.

Keywords: Knowledge Management, Post Graduation,Higher Education Institutions, Knowledge Database Discovery, ScriptLattes.

1 INTRODUÇÃO

O conhecimento tem sido considerado um dos ativos mais importantes das organizações, possibilitando ações inteligentes que conduzem à inovação e à criação contínua de produtos e serviços (Cardoso & Machado, 2008; Davenport & Prusak, 2013). A geração e a difusão do conhecimento são elementos indispensáveis para o avanço de uma sociedade, desempenhando papel relevante no aumento da qualificação geral e na formação permanente dos cidadãos (Moreira & Massarini, 2002). Além disso, a transformação do conhecimento em produtos e serviços, por meio dos mecanismos de inovação, é fator essencial para o crescimento das economias modernas (Sidone, 2013). Nesse contexto, é fundamental que o conhecimento, os métodos e as descobertas em todas as áreas da ciência estejam disponíveis a todos, na mais ampla escala (Sagan, 1996).

O processo de gestão do conhecimento abrange toda a forma de gerar, armazenar, distribuir e utilizar o conhecimento. Atualmente, no ambiente organizacional, a velocidade de coleta de dados é alta, dificultando a análise das grandes quantidades de dados armazenadas. Faz-se necessária a aplicação de técnicas e ferramentas, suportadas pela tecnologia da informação, para agilizar o processo de extração de informações relevantes desses grandes volumes de dados. Uma área de conhecimento emergente, que tenta solucionar este problema é a Descoberta de Conhecimento em Banco de Dados - DCBD (Cardoso & Machado, 2008).

As Instituições de Ensino Superior – IES são organizações voltadas para a criação, transmissão e disseminação do conhecimento, desempenhando papel destacado nos processos que configuram a sociedade contemporânea (Bernheim & Chaui, 2003). A produção científica e tecnológica das IES brasileiras constitui uma importante fonte geradora de conhecimento, principalmente no que se refere aos docentes que atuam em programas de pós-graduação *stricto sensu*, que englobam os cursos de Doutorado e Mestrado. (Almeida, 2010; CAPES/MEC, 2010; Moritz, Pereira, Moritz, & Maccari, 2013). O acesso a essa produção deveria ser fácil, no entanto o que se verifica é que embora as informações estejam disponíveis na internet, nem sempre o acesso a elas é simples ou rápido. Não é possível, por exemplo, recuperar de forma direta e imediata toda a produção de uma

universidade, nem comparar produções de áreas de conhecimento, programas ou regiões geográficas específicos.

1.1 A situação-problema

O Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq disponibiliza na Plataforma Lattes os currículos dos docentes de pós-graduação, com dados sobre atuação profissional, linhas de pesquisa e trabalhos publicados (Guedes, 2001). As consultas de pesquisadores e outros interessados, todavia, dependem de prévia liberação do acesso (Quoniam & Ferraz, 2014). A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES, também disponibiliza em seu portal, em formato PDF, os cadernos de avaliação da pós-graduação, com os dados de todos os programas e docentes (Moritz et al., 2013). Dessa maneira, tem-se disponível na internet a informação sobre os docentes da pós-graduação brasileira e de suas produções, mas armazenada em formatos e bases de dados distintos, que não foram projetados para um acesso unificado.

Para facilitar as consultas a documentos e indicadores da atividade científica é necessária a integração de dados de acesso às bases desses sistemas heterogêneos (Pacheco & Kern, 2001). Uma vez que isto ocorra, existem ferramentas independentes que conseguem extrair conhecimento da Plataforma Lattes, mapeando a produção científica de forma automatizada (Ferraz, Quoniam, & Maccari, 2014). O ScriptLattes, por exemplo, recupera os currículos a partir de listas de códigos de pesquisadores e extrai informações acadêmicas e profissionais, gera relatórios, gráficos e mapas (Mena-Chalco & Junior, 2011). Não é possível, entretanto, gerar automaticamente para esta ferramenta listas que abranjam todo o universo da pós-graduação brasileira, filtrando dados por critérios diversos como programas, áreas de conhecimento, instituições ou regiões geográficas.

A concretização dessa possibilidade tornaria mais eficiente a obtenção de dados sobre a produção de qualquer subconjunto de docentes dos programas da pós-graduação brasileira. Tais conjuntos de informações são importantes para a produção de conhecimento, com vistas a subsidiar análises de produção acadêmica e redes de colaboração de pesquisadores. Adicionalmente, essa sistemática poderia apoiar, por exemplo, projetos de pesquisa e desenvolvimento, formação de equipes

multidisciplinares, elaboração de políticas, gestão acadêmica, criação de cursos e currículos, acompanhamento e avaliação de programas de pós graduação.

Considerando a relevância do preenchimento dessa lacuna, o trabalho aqui apresentado tem o objetivo de construir uma ferramenta para a geração automatizada de listas de entrada para recuperação da produção científica de docentes da pós-graduação brasileira, por meio do programa ScriptLattes. Espera-se, desta maneira, contribuir para o processo de descoberta de conhecimento em bancos de dados da Plataforma Lattes, integrando informações dos cadernos de avaliação da CAPES e dos currículos Lattes.

Nas próximas seções, será realizada uma revisão teórica resumida, explicados os métodos de produção técnica empregados, apresentada a situação problema, descrita a intervenção realizada, analisados os resultados obtidos e, por fim, apresentadas as conclusões e considerações finais.

2 Referencial Teórico

Para fundamentar a pesquisa proposta neste relato técnico, apresenta-se uma breve revisão de informações e conceitos sobre o assunto abordado, com o objetivo de expor conhecimentos fundamentais que irão direcioná-la. Os tópicos considerados relevantes para a elaboração e apresentação deste relato técnico foram a descoberta de conhecimento em banco de dados; os cadernos de avaliação da CAPES; a Plataforma Lattes e os currículos Lattes; e as ferramentas para extração automática de informações dos currículos Lattes, com foco especial na ferramenta ScriptLattes.

2.1 Descoberta do conhecimento em bancos de dados – DCBD

Conhecimento não é dado nem é informação, embora esteja relacionado a ambos. Dados são conjuntos de fatos discretos e objetivos sobre eventos, sem poder explicativo nem preditivo, que podem ser armazenados e processados por computador, como o registro de um valor ou um nome. Informações são dados transformados, que representam algum significado para alguém. Exigem a participação humana, para atingir um consenso em relação a significados de que, por exemplo, um valor é alto ou um nome é masculino. Conhecimento é um conjunto de valores, experiências, informações conceituais e *insights* trabalhados pela mente humana para incluir novas experiências e informações, por meio de reflexão, síntese

e contexto, como em “esse valor é o dobro do meu salário” ou “todos os que conheço com esse nome são homens” (Davenport & Prusak, 2013; Valentim, 2002).

Diversas técnicas de descoberta e criação de conhecimento explícito podem ser utilizadas pelas organizações para solucionar o problema da análise de dados, em volumes que ultrapassam a habilidade e a capacidade humanas. Uma forma de solução emergente é a Descoberta de Conhecimento em Banco de Dados - DCBD, ou *Knowledge Database Discovery – KDD* (Cardoso & Machado, 2008). Está área de conhecimento pode ser formalmente definida como um processo não trivial, interativo iterativo, para identificar, em um conjunto de dados, padrões válidos, novos, potencialmente úteis e assimiláveis ao conhecimento humano (Fayyad, Piatetsky-Shapiro, Smyth, & others, 1996).

Segundo pesquisa de (Neves, 2003), são duas as metodologias de DCBD mais conhecidas, entre as várias encontradas na literatura. A primeira é a CRISP-DM, de (Chapman, Clinton, Khabaza, Reinartz, & Wirth, 1999), que compreende as etapas de (i) entendimento do negócio; (ii) compreensão dos dados; (iii) preparação dos dados; (iv) modelagem; (v) avaliação do modelo; e (vi) publicação dos resultados.

A segunda é a de (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), que propõem um sequência de passos compreendendo basicamente (i) o entendimento do domínio da aplicação; (ii) a seleção de um subconjunto dos dados; (iii) a limpeza e pré-processamento dos dados; (iv) a transformação dos dados, por redução e projeção; (v) a seleção do método de mineração; (vi) análise exploratória, para seleção do modelo, das hipóteses e dos algoritmos de mineração; (vii) a mineração dos dados; (viii) a interpretação dos padrões encontrados; e (ix) o uso do conhecimento descoberto. Este segundo processo de DCBD foi resumido por seus autores nas etapas operacionais configuradas na Figura 1 a seguir.

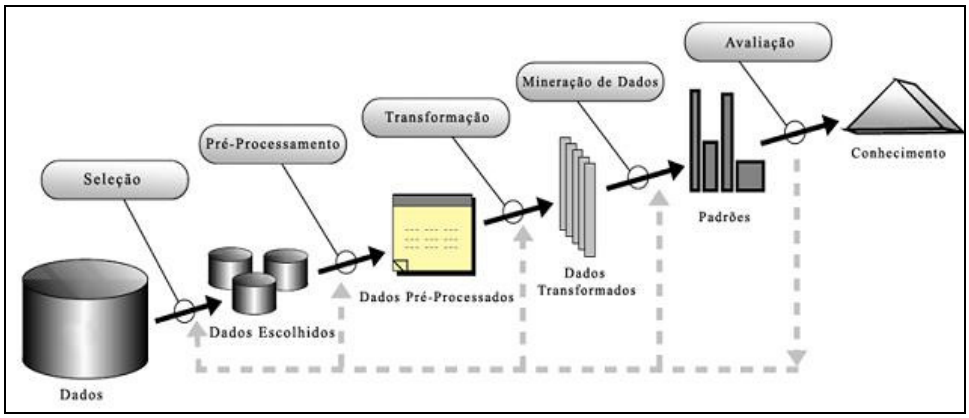

Figura 1 – Configuração resumida das etapas operacionais do processo de DBCB, segundo(Fayyad, Piatetsky-Shapiro, Smyth, et al., 1996).
Fonte: recuperada de www.devmedia.com.br/mineracao-e-analise-de-dados-em-sql/2933, em 21 de janeiro de 2015.

A observação e comparação das etapas do processo das duas metodologias mais utilizadas e ainda de outras existentes, deixa claro que há falta de consenso quanto ao processo de implementação ideal do ciclo de um processo de DCDB, ou seja, não existe a melhor abordagem (Schiessl, 2007).

2.2 Os cadernos de avaliação da CAPES

A CAPES foi criada pelo Decreto nº 29.741, em 11 de julho de 1951, durante o governo de Getúlio Vargas. Hoje, uma das atividades sob sua responsabilidade é a avaliação dos programas nacionais da pós-graduação *stricto sensu* (CAPES/MEC, 2008). O sistema de avaliação mantido pela instituição compreende o acompanhamento anual e a avaliação periódica dos programas. Foi implantado em 1976 e desde então tem desempenhado um papel importante para a elevação da qualidade, o aprimoramento e a regulação dos cursos de mestrado e doutorado, (Moritz et al., 2013).

No processo de avaliação, os programas informam os dados sobre as atividades desenvolvidas por meio de um sistema informatizado denominado "Coleta Capes" (Akim, Mergulhão, & Borrás, 2013), desenvolvido pela Fundação CAPES com o objetivo específico de coletar essas informações, que são usadas principalmente para subsidiar a avaliação dos programas de pós-graduação no país, mas também para a constituição do acervo de informações consolidadas sobre o Sistema Nacional de Pós-Graduação – SNPG. A partir de 2014 o sistema foi reformulado e passou a ser um módulo integrante da Plataforma Sucupira, uma ferramenta mais moderna, que permite a disponibilização em tempo real das informações, processos e procedimentos. (CAPES/MEC, 2014).

As informações relativas a cada programa são agrupadas pelo sistema em um conjunto de 11 documentos temáticos, tratados e organizados para permitir a emissão dos Cadernos de Indicadores, que são os relatórios utilizados no processo de avaliação propriamente dito. A relação dos 11 cadernos, com as respectivas siglas, e a descrição resumida de seus conteúdos, está apresentada na Tabela 1.

Tabela 1 - Conteúdo dos cadernos de avaliação da CAPES

Caderno	Sigla	Conteúdo
Teses e Dissertações	TE	Teses e dissertações defendidas no programa, no ano avaliado
Produção Bibliográfica	PB	Relação da produção bibliográfica de docentes, discentes e egressos do programa, destacando as cinco melhores produções bibliográficas, selecionadas pelo próprio programa.
Produção Técnica	PT	Relação da produção técnica de docentes, discentes e egressos do programa, destacando as cinco melhores produções técnicas, selecionadas pelo próprio programa.
Produção Artística	PA	Relação da produção artística de docentes, discentes e egressos do programa, destacando as cinco melhores produções artísticas, selecionadas pelo próprio programa.
Corpo Docente, Vínculo e Formação	CD	Relação dos docentes permanentes e colaboradores do programa, seus dados de vinculação com o programa, dados da titulação e situação em outros programas de pós-graduação.
Disciplinas	DI	Relação das disciplinas que integram o currículo dos cursos oferecidos no programa, com informações sobre carga horária, número de créditos, ementa, bibliografia, semestre de oferta da disciplina no ano avaliado e dados dos docentes responsáveis pela oferta.
Linhas de Pesquisa	LP	Relação das linhas de pesquisa do programa, informando a descrição, a área de concentração e os projetos de pesquisa vinculados à linha, sua situação e sua natureza.
Projetos de Pesquisa	PP	Relação dos projetos de pesquisa em andamento no ano avaliado, informando a descrição, a área de concentração, a quantidade de alunos envolvidos, dados da sua equipe e do seu financiamento.
Proposta do Programa	PO	Visão geral, evolução e tendências do programa, a integração com a graduação, a infraestrutura, intercâmbios, pontos fortes e fracos, ensino à distância, trabalhos em preparação, atividades complementares, solidariedade, nucleação e visibilidade.
Atuação do Corpo Docente	DA	Indicadores de atuação de cada docente, informando sobre oferta de disciplina(s) na graduação e na pós-graduação, orientações , participação em projeto de pesquisa e em bancas examinadoras.
Produção do Corpo Docente	DP	Indicadores de produção bibliográfica, técnica e artística dos docentes que orientaram e/ou ministraram disciplinas no ano avaliado.

Nota. Fonte Akim, É. K., Mergulhão, R. C., & Borrás, M. Á. A. (2013). Proposta de análise de falhas na coleta de informações para a avaliação de programas de pós-graduação baseada no FMEA. *Revista Eletrônica Produção & Engenharia*, *5*(1), 500–518.

Com o propósito de disseminar amplamente o conhecimento sobre a forma como é realizada a avaliação da pós-graduação, a CAPES procura garantir o pleno acesso de todos os interessados a esse conjunto de relatórios. Os documentos ficam disponíveis de forma transparente na internet, para consulta e para *download*, em arquivos no formato PDF.

2.3 A Plataforma Lattes e os currículos Lattes

A Plataforma Lattes é um conjunto de sistemas computacionais desenvolvido e mantido pelo CNPq, para gerenciar suas atividades de fomento e integrar num mesmo ambiente informações referentes a pesquisadores e instituições brasileiros e estrangeiros (Guedes, 2001). Utilizando alguns conceitos da *web* 2.0, o acesso a ela se dá diretamente pela internet, possibilitando o preenchimento dos formulários eletrônicos, a efetivação dos cadastramentos e a consulta às informações a partir de qualquer lugar, equipamento e hora (Ferraz et al., 2014).

A Plataforma Lattes é de grande importância para a ciência brasileira. É uma experiência reconhecida internacionalmente como exemplo de boas práticas relacionadas a bases de dados de produção acadêmica, pelo fato de ter sido criada e desenvolvida por uma "comunidade virtual", pelo forte incentivo para que pesquisadores e instituições a utilizem e pela identificação única estabelecida para cada pesquisador, resolvendo problemas de homonímia (Lane, 2010).

2.4 O ScriptLattes e outras ferramentas de extração para os currículos Lattes

O *ScriptLattes* é uma ferramenta de *software* livre, projetada para a extração e compilação automáticas de produções bibliográficas, técnicas e artísticas, orientações, projetos de pesquisa, prêmios e títulos, grafos de colaborações e mapas de geolocalização de um conjunto de pesquisadores da plataforma Lattes, disponibilizando-os em uma página na *web,* facilmente acessada e consultada (Mena-Chalco & Junior, 2011). Tem sido amplamente utilizado por diversos pesquisadores no Brasil e os resultados obtidos até agora têm sido de grande valia (Ferraz et al., 2014; Martins, Maccari, Storopoli, & Andrade, 2013; Mena-Chalco & Cesar Junior, 2009; Quoniam & Ferraz, 2014; Sidone, 2013; Sidone, Haddad, & Mena-Chalco, 2014).

Possui recursos para baixar em formato HTML os currículos de um grupo de interesse, de forma simples e automática, compilar as listas de produções e tratar as ocorrências de produções duplicadas e similares. Gera relatórios, em formato HTML, com listas de produções e orientações concluídas e em andamento. Permite a criação automática de gráficos de redes de coautoria e de um mapa de geolocalização dos membros e alunos de pós-graduação com orientação concluída (Mena-Chalco & Junior, 2011).

O ScriptLattes recupera os dados dos currículos que o usuário deseja analisar por meio de um identificador de 16 algarismos, o idLattes16. Além deste, cada currículo tem um outro código de acesso, composto por 10 caracteres alfanuméricos, iniciados pela letra "K", que chamaremos aqui de idLattes10. A entrada do sistema é composta por uma lista de IdLattes16, em conjunto com o intervalo de tempo, em anos, em que os dados de cada membro do grupo vai ser recuperado e analisado (Ferraz et al., 2014).

3 MÉTODO DA PRODUÇÃO TÉCNICA

Em conformidade com o protocolo proposto por (Biancolino, Kniess, Maccari, & Rabechini Jr., 2012) este relato técnico foi elaborado com propósitos profissionais, sem deixar de também buscar o rigor científico e metodológico. Pretende-se compartilhar aqui uma experiência de natureza técnica e aplicação prática, com vistas à resolução de um problema específico. Seguindo, portanto, as recomendações do protocolo, será detalhada nesta seção a maneira como o trabalho foi executado para atingir o objetivo do relato.

O problema foi abordado por meio do método indutivo, em que a aproximação dos fenômenos ocorre de forma ascendente e cada vez mais abrangente, a partir da observação do mais específico para o mais geral, conforme definição de (Marconi & Lakatos, 2010). A pesquisa também se caracterizou por seu aspecto qualitativo, já que não intencionava medir eventos nem fazer análises estatísticas. Envolveu a obtenção de dados descritivos sobre processos interativos, o contato direto com a situação estudada, o reconhecimento e a análise de diferentes perspectivas e as reflexões do pesquisador como parte do processo de produção do conhecimento (Flick, 2009; Godoy, 1995).

A primeira etapa do estudo foi a investigação realizada em diversas fontes, para aprofundar o conhecimento do assunto (Theóphilo & Martins, 2009), uma vez que o momento inicial situava-se no ponto em que ainda não se sabia, mas se desejava descobrir. Além da pesquisa bibliográfica, foram acessados os portais da CAPES e do CNPq, para a busca de informações e conteúdo.

No intuito de ampliar o conhecimento, durante todo o processo de investigação e análise foram realizadas entrevistas com dois pesquisadores que trabalham no desenvolvimento de novas funcionalidades para a ferramenta ScriptLattes e também a utilizam em estudos sobre produção acadêmica e redes de colaboração científica. As entrevistas foram abertas e não foram gravadas. Foram feitas anotações em papel e o material digital disponibilizado foi gravado em um *pendrive.* As perguntas foram direcionadas para o conhecimento a respeito do problema, as fontes de informação relevantes, o que já havia sido feito e o que poderia ser utilizado nesta pesquisa a partir do material já existente.

Com base no conhecimento adquirido sobre o problema, o processo de exploração prosseguiu, com a análise e seleção dos elementos relevantes para o estudo. Algumas questões e focos de interesse foram definidas durante o desenvolvimento do trabalho, como prevê (Godoy, 1995) para uma pesquisa exploratória. Nesta etapa buscou-se entender o relacionamento entre os dois bancos de dados e identificar as características de uma ferramenta computacional capaz de possibilitar a integração entre eles. Também foram analisadas ferramentas e elementos já existentes, que pudessem apoiar o processo de integração, ou mesmo ser incorporadas ao resultado deste trabalho.

Por fim, a última etapa consistiu na elaboração de uma proposta de ferramenta computacional para gerar automaticamente listas de entrada para o programa ScriptLattes, conforme está previsto no objetivo do estudo.

4 TIPO DE INTERVENÇÃO E MECANISMOS ADOTADOS

Nesta seção, serão descritas as atividades desenvolvidas no estudo, abrangendo as entrevistas com os pesquisadores do ScriptLattes, que complementaram os conhecimentos adquiridos na fase de conhecimento do assunto e possibilitaram que fosse construída e apresentada uma solução para o problema.

Nas entrevistas realizadas, as opiniões sobre a relevância do objetivo deste trabalho foram positivas. Durante as conversas, os pesquisadores relataram sua experiência com o uso da ferramenta Scriptlattes e repassaram informações úteis sobre páginas relevantes, navegação e acesso a dados nos portais da CAPES e do CNPq.

Também informaram que já foram desenvolvidas rotinas para recuperação dos cadernos de avaliação e para a recuperação do idLattes16. Para esta última, foi preciso conhecer o idLattes10 do currículo, usado para formatar e submeter uma consulta *web,* que devolve o número do idLattes16 dentro do código interno HTML da página de resposta.

As rotinas foram desenvolvidas na linguagem de programação de código aberto Phyton. Todos os códigos-fontes dos programas já desenvolvidos foram colocados à disposição, para análise, aperfeiçoamento e uso em uma possível futura implementação da ferramenta aqui proposta. Ainda foram repassados aos autores deste trabalho os modelos de configurações necessárias para a geração de listas de pesquisadores para entrada na ferramenta ScriptLattes.

Em consonância com o que se esperava no início do trabalho, o conhecimento mais profundo gerou familiaridade suficiente com o problema, a ponto de ser possível desenvolver uma proposta para sua solução (Gil, 2008). A experiência dos pesquisadores também subsidiou reflexões e decisões tomadas no processo de produção do conhecimento (Flick, 2009). Após a análise de todas as informações obtidas durante as atividades de pesquisa realizadas, procedeu-se à elaboração da solução, que resultou na proposta de construção de um sistema de informação, estruturado em dois módulos.

5 APRESENTAÇÃO E ANÁLISE DOS RESULTADOS OBTIDOS

A solução proposta compreende uma base de dados e uma aplicação. Na base de dados ficam armazenados os dados referentes às IESs, aos cursos de pós-graduação e aos professores que neles atuam, recuperados dos cadernos de avaliação da CAPES, da Plataforma Lattes e do sítio da GEOCAPES, em (*geocapes.capes.gov.br/geocapesds/#)*. A aplicação proposta deve disponibilizar uma interface para que nela os usuários selecionem os critérios para a formação das listas de pesquisadores, de acordo com os dados armazenados no banco. O

processamento acontece após a confirmação pelo usuário dos critérios selecionados, por meio de uma consulta aos dados solicitados e subsequentes geração e disponibilização das listas de pesquisadores, que servirão como entradas para o ScriptLattes. Um esquema desta solução é apresentado na Figura 2.

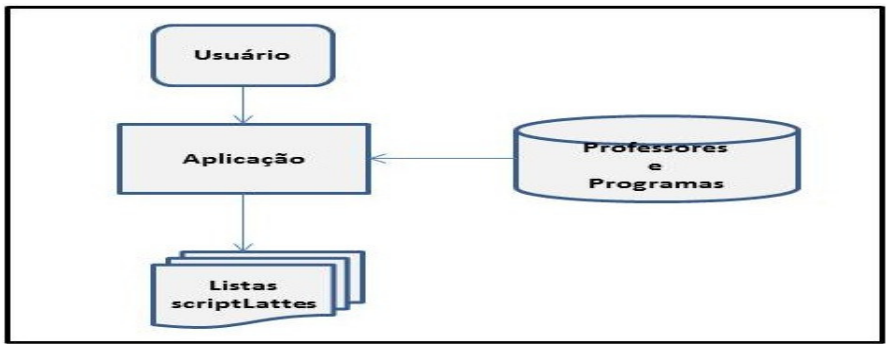

Figura 2 – Estrutura da solução proposta para a ferramenta: base de dados e aplicação.
Fonte: Elaborada pelos autores.

5.1 Descrição da base de dados proposta

A base de dados contempla as informações de professores, programas e IES. Dos professores, foram armazenados os dados pessoais e profissionais, dados de acesso à Plataforma Lattes e dos programas a que estão vinculados. Dos programas e IESs, foram armazenados os dados pelos quais possam ser filtrados e agrupados para a geração de listas para o ScriptLattes. Na tabela 2 estão descritos os principais componentes identificados para compor a base de dados, bem como seus principais atributos, sua origem e o domínio de valores que podem assumir.

Tabela 2 - Componentes da base de dados e seus atributos

Entidade	Atributos	Origem/Domínio
Docente	idLattes 16, idLattes 10, nome, formação, titulação	Cadernos de Avaliação da CAPES e currículos Lattes
Programa	código, nome, IES, grande área, área de avaliação, área de conhecimento, conceito, URL do caderno, tipo de programa	Portal da CAPES - Geocapes
IES	código, nome, UF, região, município, status jurídico	Portal da CAPES - Geocapes
Grande Área	código, nome	Portal da CAPES - Tabela de Áreas
Área de Avaliação	código, nome, grande área	Portal da CAPES - Tabela de Áreas
Área de Conhecimento	código, nome, grande área, área de avaliação	Portal da CAPES - Tabela de Áreas
Tipo de Programa	sigla, descrição	Mestrado, Doutorado, Mestrado Profissional, Mestrado/Doutorado
Tipo de Vinculação	sigla, descrição	P - Permanente, V - Visitante, C - Colaborador
Status Jurídico	sigla, descrição	Federal, Estadual, Municipal, Particular
Região Geográfica	sigla, nome	Norte, Nordeste, Sul, Sudeste, Centro-Oeste
UF	sigla, nome, região	Tabela IBGE
Município	código, nome, UF	Tabela IBGE

Nota. Fonte: Elaborado pelos autores.

No que se refere à forma como os dados serão armazenados, a decisão deverá ser tomada pela equipe responsável pelo projeto de implementação. A escolha vai depender dos recursos existentes e das características que forem definidas para a implementação do sistema de informação.

5.2 A aplicação para gerar listas ScriptLattes

A aplicação proposta para a geração de listas a partir dos dados do banco é de baixa complexidade. Deve permitir ao usuário gerar listas de entrada para o programa ScriptLattes, a partir de aplicação de filtros aos diversos critérios armazenados na base de dados. A aplicação deve ainda oferecer recursos para aplicar os critérios na seleção e na classificação hierárquica dos dados e também para gerar as listas no padrão adotado pelo ScriptLattes. Recomenda-se que seja uma aplicação *web*, mas este requisito não é obrigatório, ficando sua definição a critério da equipe de implementação, também dependendo dos recursos existentes e das características que forem definidas para o serviço a ser implementado.

Na figura 3, apresenta-se uma proposta para a interface da aplicação onde o usuário poderia selecionar os critérios para confecção das listas para o ScriptLattes e comandar o processamento para sua geração e disponibilização para processamento

Figura 3 Proposta para a interface da aplicação.
Fonte: Elaborada pelos autores.

A interface disponibilizada para que os usuários selecionem os critérios para geração das listas deve possibilitar a prevenção de erros e garantir a integridade das opções feitas, oferecendo para seleção apenas os dados compatíveis com o que já foi previamente selecionado. Quando é selecionada uma região geográfica por exemplo, apenas as IES daquela região devem ser disponibilizadas para as escolhas subsequentes.

Quanto às demais características da aplicação, como a linguagem de programação, o padrão de projeto, o padrão de interface com o usuário e recursos de usabilidade, além de outros requisitos não funcionais, como padrões de desempenho e segurança das informações, todas deverão ser definidas pela equipe responsável pela implementação do projeto.

5.3 Os passos para o desenvolvimento da ferramenta

A Figura 4 apresenta uma visão resumida de uma sequência de etapas para uma implementação da ferramenta aqui proposta. O processo de implantação contempla a criação inicial da base de dados seguida pelo desenvolvimento da aplicação. Este processo tem características de DCBD, apresentando aspectos iterativos, com repetições de etapas, e interativos, controlados pelo usuário, com a possibilidade de participação de especialistas no domínio do problema (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

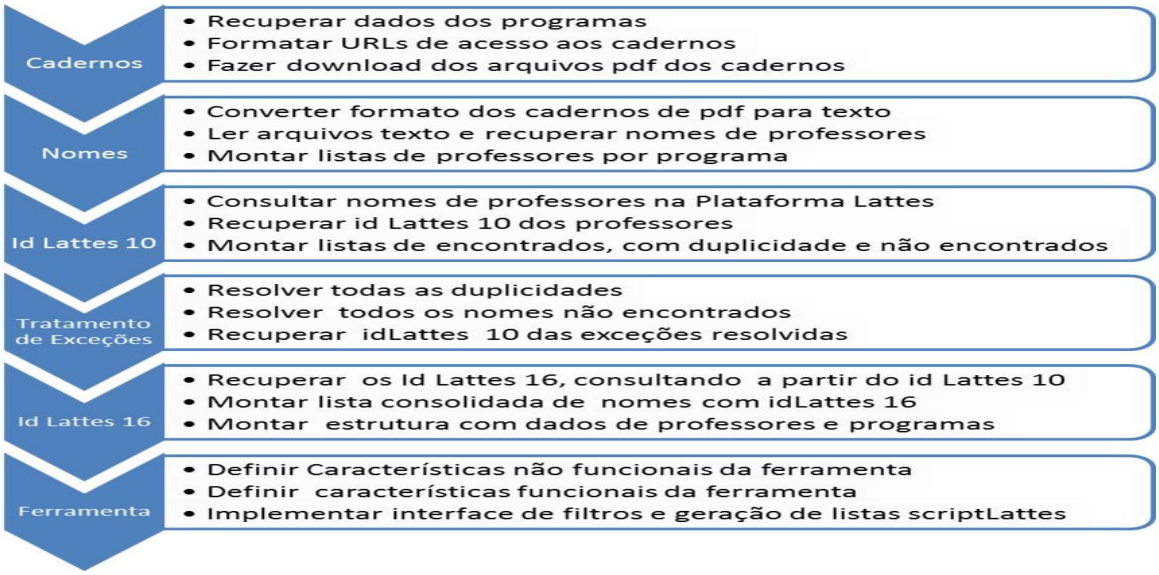

Figura 4 - Uma sequência possível de procedimentos, sugerida para a implementação da ferramenta.
Fonte: Elaborado pelos autores.

A sequência de procedimentos sugerida começa com a geração dos dados para a criação da base. O passo inicial é recuperar os dados dos programas de pós-graduação do sítio da GEOCAPES e com eles formatar as URLs de acesso e fazer o *download* de todos os cadernos de avaliação do ´triênio 2010-2012, cerca de três mil.

Os cadernos devem ser convertidos do formato PDF para o formato texto e em seguida processados para extração dos dados referentes aos professores nos programas a que estejam vinculados. Cada nome da lista de professores resultante será submetido à consulta na ferramenta de Busca Simples de currículos na Plataforma Lattes, para identificação de seu idLattes10. Espera-se que grande parte dos nomes será encontrada diretamente, outro conjunto, menor, apresentará nomes com homônimos e outro conjunto, ainda menor, de nomes de pesquisadores, não será encontrado.

O passo seguinte será o tratamento das exceções, quando serão resolvidas as ocorrências de duplicidades e nomes não encontrados. Ao seu final, todos os nomes de professores extraídos dos cadernos de avaliação deverão estar associados ao seu respectivo idLattes10. A Figura 5 mostra uma representação esquemática dessas atividades.

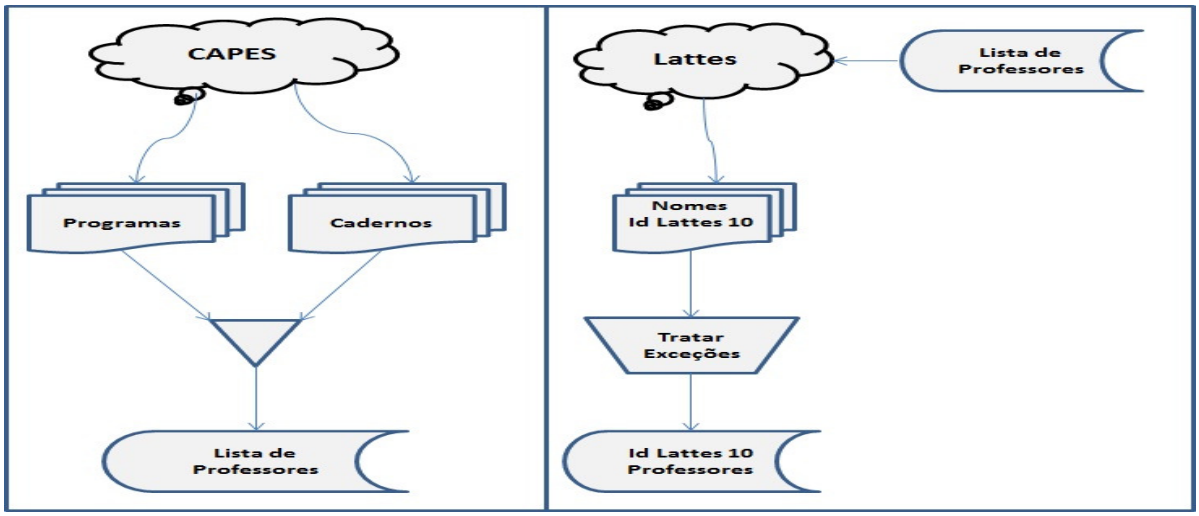

Figura 5 - Geração da lista de professores e recuperação do idLattes10.
Fonte: Elaborado pelos autores.

Ressalta-se o caráter iterativo e interativo desta fase, características inerentes aos processos de DCBD, segundo (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), que deve ser controlado e no qual pode ser preciso intervir para retornar e avançar,

repetindo operações com técnicas diferentes, até se obter o resultado desejado. Este trabalho deverá ser feito com apoio de programas de computador.

Na sequência, os idLattes10 serão submetidos à consulta na plataforma Lattes, para recuperação dos idLattes16 correspondentes, que comporão a base de dados proposta para a ferramenta, em conjunto com os dados já existentes de professores e programas. Também nessas etapas será necessário o suporte de programas de computador, devendo ser verificada a possibilidade de utilização das rotinas cedidas pelos pesquisadores entrevistados na etapa inicial desta pesquisa. A Figura 6 mostra uma representação esquemática dessas atividades.

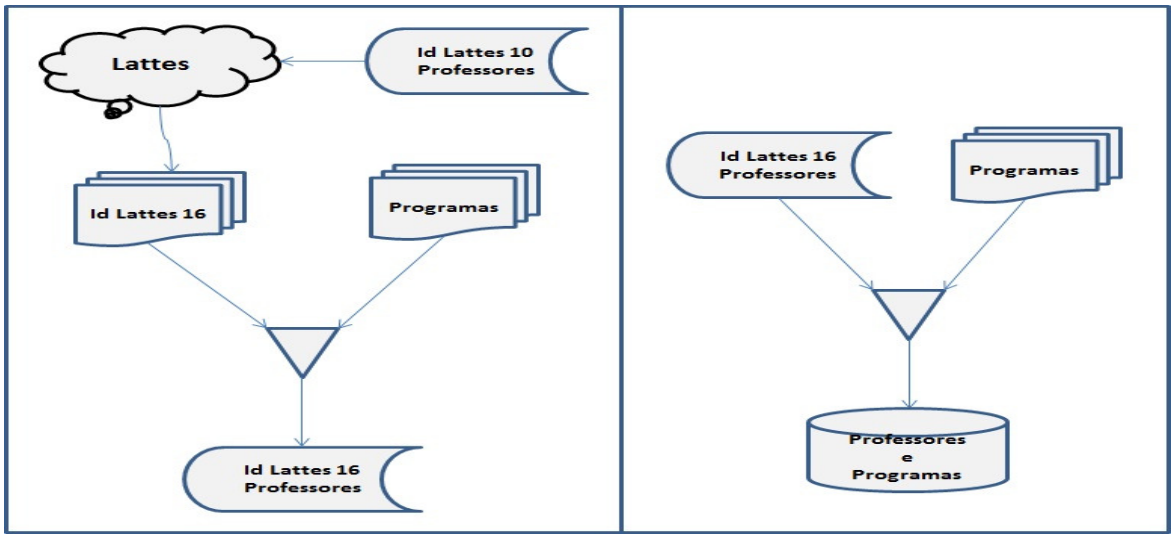

Figura 6– Recuperação do idLattes16 e geração do banco de dados.
Fonte: Elaborado pelos autores.

Depois de criada a base de dados, ou ainda durante os procedimentos para sua criação, procede-se ao desenvolvimento da aplicação. Por se tratar de um software de baixa complexidade, não há recomendações de características específicas para ele, apenas para a atenção nas fases de definição de requisitos não funcionais e funcionais. Isso possibilitará que a implementação da interface esteja aderente aos padrões da ferramenta ScriptLattes e que tenha uma boa usabilidade e sejam de fácil manutenção.

6 CONCLUSÕES

Este trabalho propõe uma solução para o problema de integrar informações sobre produção acadêmica da pós-graduação brasileira existentes nas bases de dados da CAPES e do CNPq, para delas extrair conhecimento por meio da utilização do programa ScriptLattes.

Verificou-se, com a pesquisa, que este processo de extração de dados dos currículos da Plataforma Lattes é de natureza semelhante aos estudados pela área da gestão de conhecimento denominada Descoberta de Conhecimento em Bancos de Dados (DCBD). Verificou-se ainda que a atividade específica de integração dos dados para geração de listas para o ScriptLattes pode ser considerada como parte da etapa de seleção de dados, no ciclo de vida típico de um processo de descoberta de conhecimento em bancos de dados realizado com a aplicação de uma das metodologias existentes mais utilizadas, proposta por (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

O objetivo do trabalho foi alcançado com a apresentação da proposta da ferramenta, um sistema de informação composto por dois módulos, um banco de dados e uma aplicação. O banco de dados armazena os dados de docentes, IES e programas; já a aplicação permite a criação de filtros para gerar as listas que podem ser processadas pelo programa ScriptLattes. Com base no conhecimento adquirido durante as atividades desenvolvidas, foi ainda apresentada uma sequência de etapas recomendadas para a criação do banco de dados proposto, em uma futura implementação, bem como recomendações para o desenvolvimento da aplicação.

Apesar do trabalho desenvolvido ter ficado restrito à elaboração de uma proposta conceitual da ferramenta, espera-se que a contribuição trazida por esta pesquisa seja parte de um projeto mais amplo. Recomenda-se, portanto, a continuação das atividades deste trabalho, com a implementação do modelo aqui apresentado.

Caso esta implementação aconteça, a utilização da ferramenta desenvolvida tem potencial para gerar conhecimentos importantes, relativos à produção acadêmica de toda a pós-graduação brasileira, a partir de critérios diversificados e com abrangência nacional. Este conhecimento pode subsidiar análises de produção acadêmica e redes de colaboração de pesquisadores, bem como apoiar projetos de pesquisa e desenvolvimento, formação de equipes multidisciplinares, elaboração de políticas, gestão acadêmica, criação de cursos, acompanhamento e avaliação de programas de pós graduação, entre outros trabalhos.

CITAÇÕES E REFERÊNCIAS BIBLIOGRÁFICAS

Akim, É. K., Mergulhão, R. C., & Borrás, M. Á. A. (2013). Proposta de análise de falhas na coleta de informações para a avaliação de programas de pós-graduação baseada no FMEA. *Revista Eletrônica Produção & Engenharia*, *5*(1), 500–518.

Almeida, M. H. T. de. (2010, dezembro). A Pós-Graduação no Brasil: onde Está e para onde Poderia Ir, em PLANO NACIONAL DE PÓS-GRADUAÇÃO (PNPG) 2011-2020 Documentos Setoriais Volume II. DTI/CGD/CAPES.

Bernheim, C. T., & Chaui, M. de S. (2003). Challenges of the university in the knowledge society, five years after the World Conference on Higher Education. In *Paper produced for the UNESCO Forum Regional Scientific Committee for Latin America and the Caribbean (UNESCO Forum Occasional Paper Series N 4)*.

Biancolino, C. A., Kniess, C. T., Maccari, E. A., & Rabechini Jr., R. (2012). Protocolo para elaboração de relatos de produção técnica. *Revista de Gestão e Projetos - GeP*, *3*(2), 294–307.

CAPES/MEC. (2008, junho 17). CAPES - História e missão [Institucional]. Recuperado 24 de setembro de 2014, de http://www.capes.gov.br/historia-e-missao

CAPES/MEC. (2010, dezembro). Plano Nacional de Pós-Graduação (PNPG) 2011-2020 Volume I. DTI/CGD/CAPES.

CAPES/MEC. (2014, julho 25). Coleta de Dados - Conceitos e orientações - Versão 1.5.

Cardoso, O. N. P., & Machado, R. T. M. (2008). Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras. *Revista de AdministraçãoPública*, *42*, 495 – 528.

Chapman, P., Clinton, J., Khabaza, T., Reinartz, T., & Wirth, R. (1999). The CRISP-DM process model. *The CRIP–DM Consortium*, *310*.

Davenport, T. H., & Prusak, L. (2013). *Working Knowledge: How Organizations Manage What They Know*. Harvard Business Press.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & others. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. In *KDD* (Vol. 96, p. 82–88).

Ferraz, R. R. N., Quoniam, L. M., & Maccari, E. A. (2014). A Utilização Da Ferramenta Scriptlattes Para Extração E Disponibilização On Line Da Produção Acadêmica De Um Programa De Stricto Sensu Em Administração. *Revista Brasileira de Pós-Graduação*, *11*(24). http://doi.org/10.5748/9788599693100-11.

Flick, U. (2009). *An Introduction to Qualitative Research*. SAGE Publications. Recuperado de http://books.google.com.br/books?id=sFv1oWX2DoEC.

Gil, A. C. (2008). *Como elaborar projetos de pesquisa* (4º ed). São Paulo: Atlas.

Godoy, A. S. (1995). Introdução à pesquisa qualitativa e suas possibilidades. *Revista de administração de empresas*, *35*(2), 57–63.

Guedes, C. A. (2001). CURRÍCULO LATTES Perguntas e Respostas. Recuperado 24 de setembro de 2014, de http://www.pucrs.campus2.br/manuais/dicas_lattes.pdf
Lane, J. (2010). Let's make science metrics more scientific. *Nature*, *464*(7288), 488–489.

Marconi, M. de A., & Lakatos, E. M. (2010). Fundamentos de metodologia científica. In *Fundamentos de metodologia científica*. Atlas.

Martins, C., Maccari, E. A., Storopoli, J. E., & Andrade, R. O. B. (2013). Influência das estratégias e recursos para o desenvolvimento dos programas de pós-graduação da área de Administração, Ciências Contábeis e Turismo no período de 2001 a 2009. *Revista Gestão Universitária na América Latina-GUAL*, *6*(3), 146–168.

Mena-Chalco, J. P., & Cesar Junior, R. M. (2009). ScriptLattes: an open-source knowledge extraction system from the Lattes platform. *Journal of the Brazilian Computer Society*, *15*, 31 – 39.

Mena-Chalco, J. P., & Junior, R. M. C. (2011). Prospecção de dados acadêmicos de currículos Lattes através de scriptLattes. *Capítulo do livro Bibliometria e Cientometria: reflexões teóricas e interfaces (in press). São Carlos: Pedro & João*, 1–20.

Moreira, I., & Massarini, L. (2002). Ciência e Público. *UFRJ*.

Moritz, G. de O., Pereira, M. F., Moritz, M. O., & Maccari, E. A. (2013). A Pós-Graduação brasileira: evolução e principais desafios no ambiente de cenários prospectivos. *Future Studies Research Journal: Trends and Strategies*, *5*(2), 03–34.

Neves, R. de C. D. das. (2003). *Pré-processamento no processo de descoberta de conhecimento em banco de dados*. UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL.

Pacheco, R. C. dos S., & Kern, V. M. (2001). Uma ontologia comum para a integração de bases de informações e conhecimento sobre ciência e tecnologia. *Ci. Inf*, *30*(3), 56–63.

Quoniam, L., & Ferraz, R. R. N. (2014). Extração e Disponibilização On-Line de Indicadores de Desempenho e Prospecção dos Resultados das Pesquisas em Dengue Realizadas pela Comunidade Científica Brasileira por meio da Utilização da Ferramenta Computacional Scriptlattes. Apresentado em XXXVIII Encontro da ANPAD.

Sagan, C. (1996). O mundo assombrado por demônios. *A ciência vista como uma vela na escuridão. São Paulo: Editora Cia das Letras*.

Schiessl, J. M. (2007). *Descoberta de conhecimento em texto aplicada a um sistema de atendimento ao consumidor. 2007. 106p*. Dissertação (Mestrado em Administração)–Faculdade de Economia, Administração, Contabilidade e Ciência da Informação e Documentação, Universidade de Brasília, Brasília.

Sidone, O. J. G. (2013). *Análise espacial da produção e das redes de colaboração científica no Brasil: 1990-2010*. Universidade de São Paulo.

Sidone, O. J. G., Haddad, E. A., & Mena-Chalco, J. (2014). Padrões de Colaboração Científica no Brasil: O Espaço Importa? In *Anais do XLI Encontro Nacional de Economia*. ANPEC-Associação Nacional dos Centros de Pósgraduação em Economia.

Theóphilo, C. R., & Martins, G. A. (2009). *Metodologia da investigação científica para ciências sociais aplicadas* (2º ed). São Paulo: Atlas.

Valentim, M. L. P. (2002, agosto). Inteligência Competitiva em Organizações: dado, informação e conhecimento. *DataGramaZero - Revista de Ciência da Informação*, *3*(4). Recuperado de http://www.dgz.org.br http://www.dgzero.org