Ministério da
Ciência, Tecnologia e Inovação

sid.inpe.br/mtc-m21b/2014/03.21.14.11-TDI

DESENVOLVIMENTO E APLICAÇÕES DE FERRAMENTAS COMPUTACIONAIS PARA O MAPEAMENTO DE PRODUÇÃO CIENTÍFICA

Alexandre Donizeti Alves

Tese de Doutorado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Horacio Hideki Yanasse, e Nei Yoshihiro Soma, aprovada em 26 de fevereiro de 2014.

URL do documento original:
<http://urlib.net/8JMKD3MGP5W34M/3FUNDJ5>

INPE
São José dos Campos
2014

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GB)
Serviço de Informação e Documentação (SID)
Caixa Postal 515 - CEP 12.245-970
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/6921
Fax: (012) 3208-6919
E-mail: pubtc@sid.inpe.br

Ministério da
Ciência, Tecnologia e Inovação

sid.inpe.br/mtc-m21b/2014/03.21.14.11-TDI

DESENVOLVIMENTO E APLICAÇÕES DE FERRAMENTAS COMPUTACIONAIS PARA O MAPEAMENTO DE PRODUÇÃO CIENTÍFICA

Alexandre Donizeti Alves

Tese de Doutorado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Horacio Hideki Yanasse, e Nei Yoshihiro Soma, aprovada em 26 de fevereiro de 2014.

URL do documento original:
<http://urlib.net/8JMKD3MGP5W34M/3FUNDJ5>

INPE
São José dos Campos
2014

Dados Internacionais de Catalogação na Publicação (CIP)

Alves, Alexandre Donizeti.

Al87d Desenvolvimento e aplicações de ferramentas computacionais para o mapeamento de produção científica / Alexandre Donizeti Alves. – São José dos Campos : INPE, 2014.

xxxii + 242 p. ; (sid.inpe.br/mtc-m21b/2014/03.21.14.11-TDI)

Tese (Doutorado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2014.

Orientadores : Drs. Horacio Hideki Yanasse, e Nei Yoshihiro Soma.

1. extração de informação. 2. bases de dados científicas. 3. produção científica. 4. indicadores bibliométricos. 5. linguagem de domínio específico. I.Título.

CDU 004.738.1:167

Aprovado (a) pela Banca Examinadora em cumprimento ao requisito exigido para obtenção do Título de *Doutor(a)* em

Computação Aplicada

Dr. Solon Venâncio de Carvalho

Presidente / INPE / SJCampos - SP

Dr. Nei Yoshihiro Soma

Orientador(a) / ITA/CTA / São José dos Campos - SP

Dr. Horacio Hideki Yanasse

Orientador(a) / INPE / SJCampos - SP

Dr. Edson Luiz França Senne

Membro da Banca / UNESP/GUARA / Guaratinguetá - SP

Dra. Ana Paula Cabral Seixas Costa

Convidado(a) / UFPE / Recife - PE

Dr. Paulo Afonso de Oliveira Soviero

Convidado(a) / ITA/CTA / São José dos Campos - SP

Este trabalho foi aprovado por:

( ) maioria simples

(X) unanimidade

Aluno (a): *Alexandre Donizeti Alves*

São José dos Campos, 26 de Fevereiro de 2014

"Ando devagar porque já tive pressa

E levo esse sorriso porque já chorei demais

Hoje me sinto mais forte, mais feliz, quem sabe

Só levo a certeza de que muito pouco sei, ou nada sei

...

Cada um de nós compõe a sua história

Cada ser em si carrega o dom de ser capaz e ser feliz"

Tocando em frente (1990)

Almir Sater e Renato Teixeira

Este trabalho é dedicado a meu filho Bruno que, por muitas vezes, me fez parar este trabalho pedindo a minha atenção e um pouco do meu tempo.

E é com grande alegria que eu posso dizer que, quando eu não estava me dedicando a este trabalho, eu estava me dedicando a meu amado filho.

AGRADECIMENTOS

A Deus, por ter me dado forças e condições para concluir mais um trabalho.

Ao meu orientador Nei Yoshihiro Soma, pela oportunidade e confiança em um momento muito difícil da minha vida. Ao longo do doutorado, por toda a ajuda e incentivo. Também gostaria de agradecer muitos pelos momentos de sabedoria. Quantas vezes, na sua sala, tive o privilégio de ouvir pensamentos e histórias que me fizeram pensar e crescer como ser humano. Obrigado também por muitas vezes ser mais que um orientador, ser um amigo.

Ao meu orientador Horacio Hideki Yanasse, primeiro por ter concordado em me orientar. Também por sempre me atender e me ouvir com toda a paciência. Tive a oportunidade de aprender muito com o seu modo de ser e de pensar. Tenho muito orgulho de ter tido um orientador com a sua postura e ética. Agradeço também por todas as suas revisões que, além de melhorarem muito o texto, contribuíram também para a minha formação como doutor.

A minha esposa Marinalva, por todo o carinho e motivação, além de toda a sua enorme ajuda em banco de dados. Mais uma vez completamos mais uma etapa das nossas vidas juntos, passando por bons e maus momentos. Só nos dois sabemos o que passamos para chegarmos até aqui. Por isso, dou muito valor a tudo que conquistamos, pois foi conquistado com muito sacrifício, amor e união. Obrigado por tudo e principalmente, pelos nossos lindos filhos.

A meus filhos Bruno e Sofia, que são a razão da minha vida, por todos os momentos que passamos juntos. A Sofia veio para completar a nossa família.

A meus pais e familiares, por toda torcida e principalmente, pelo apoio em momentos difíceis.

Aos amigos Erwin e Andréa, por todo o incentivo dado desde antes do doutorado e pela ajuda, principalmente, quando chegamos a São José dos Campos.

Ao professor José Demísio Simões da Silva[t], que foi a primeira pessoa a me receber no INPE, pela paciência e disposição em me ajudar.

Ao professor Rafael Duarte Coelho dos Santos, por ter me atendido diversas vezes possibilitando esclarecer minhas dúvidas.

Ao professor Nandamudi L. Vijaykumar, por ter me ajudado em várias circunstâncias e por sempre estar disposto a me atender a qualquer momento.

À professora Margarete Oliveira Domingues, por toda a sua ajuda e incentivo em resolver um problema que tive no final do meu doutorado.

Às secretárias do LAC e da CAP, e em especial, à Cristina, que sempre esteve à disposição para nos ajudar em qualquer momento, com muita paciência e com um sorriso no rosto.

Aos colegas de disciplinas, Bruno, Márcio e Marcos, pela ajuda e companheirismo.

À CAPES, pelo apoio financeiro.

Ao INPE, pela oportunidade.

RESUMO

O crescimento explosivo e a popularidade da Web têm resultado em uma grande quantidade de fontes de informação na Internet. A Web é hoje uma grande fonte de informação, fazendo com que o processo de extração de informações relevantes de conteúdos Web seja considerado um problema importante. Cada vez mais se fazem necessárias ferramentas capazes de extrair automaticamente os dados de interesse de um usuário, facilitando o acesso e a manipulação dessas informações. Agências governamentais de fomento à pesquisa se empenham cada vez mais em deixar público dados e informações sobre o ensino e pesquisa no Brasil, desde investimentos financeiros até informações sobre os pesquisadores em geral. Este trabalho descreve um conjunto de ferramentas computacionais desenvolvidas para a extração de informações em bases públicas de dados científicas nacionais e internacionais. Com isso é possível realizar análises e estudos da produção científica de pesquisadores, instituições, áreas e até mesmo países. Esse conhecimento permite que políticas públicas sejam mais bem definidas. Além disso, esses estudos podem contextualizar a produção científica brasileira no cenário internacional. Para mostrar as potencialidades das ferramentas desenvolvidas foram realizados alguns estudos de casos. Em um dos estudos foi possível identificar inconsistências em algumas bases de dados científicas. Em outro estudo foi definida uma metodologia para identificar pesquisadores que realmente atuam em uma determinada área do conhecimento. Também foi proposto um novo índice que permite medir o nível de colaboração entre os autores de um artigo.

DEVELOPMENT AND APPLICATIONS OF COMPUTATIONAL TOOLS FOR MAPPING SCIENTIFIC PRODUCTION

ABSTRACT

The explosive growth and popularity of the Web has resulted in many sources of information on the Internet. The Web is nowadays a great source of information, hence the process of extracting relevant content is an important problem. Tools that automatically extract only the data of interest are increasingly needed, in order to facilitate the access and the manipulation of the relevant information. Government agencies for research funding increasingly strive to leave public data and information on teaching and research in Brazil, from financial investments to information about the researchers, in general. This paper describes a set of computational tools developed for the extraction of information in public databases of national and international scientific data. This makes it possible to perform analysis and studies of the scientific production of researchers, institutions, areas and even countries. This knowledge allows managers to define more clearly public policies. These studies also can contextualize the Brazilian scientific production in the international scenario. To show the potential of the tools developed some case studies were performed. In one of the studies, inconsistencies in scientific databases were identified. In another study, a methodology to identify researchers who actually work in a particular area of knowledge was defined. A new index that measures the level of collaboration between the authors of an article was also proposed.

LISTA DE FIGURAS

Pág.

LISTA DE TABELAS

Pág.

LISTA DE LISTAGENS

Pág.

SUMÁRIO

Pág.

1 INTRODUÇÃO

O crescimento explosivo e a popularidade da Web têm resultado em uma grande quantidade de fontes de informação na Internet. A Web é hoje uma grande fonte de dados, fazendo com que o processo de extração de informações relevantes de conteúdos Web seja considerado um problema importante. Cada vez mais se fazem necessárias ferramentas capazes de extrair automaticamente os dados de interesse de um usuário, facilitando o acesso e a manipulação dessas informações. Isto traz grandes desafios na elaboração de metodologias eficazes para pesquisa, acesso e integração de informação (VADREVU et al., 2007).

No Brasil, agências governamentais de fomento à pesquisa, desenvolvimento e inovação como o CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) e a CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) se empenham, cada vez mais, em deixar público dados e informações sobre o ensino e pesquisa no Brasil, desde aqueles relativos aos investimentos financeiros até informações individuais dos pesquisadores em geral.

Apesar da grande quantidade de dados públicos disponíveis atualmente nas mais diversas bases de dados científicas nacionais e internacionais, há um sério problema no que diz respeito à obtenção desses a partir dessas bases. Normalmente, somente é possível consultar esses dados via uma interface Web pré-definida disponibilizada pelas próprias bases de dados. Algumas bases impõem limitações de acesso e/ou disponibilizam os dados apenas em formatos que dificultam a extração de informações. Como consequência, não é possível analisar os dados de mais de uma base de maneira padronizada, ou seja, não é possível cruzar dados dessas bases e realizar análises mais abrangentes bem como estudos mais complexos.

A proposta deste trabalho foi investigar a possibilidade de desenvolver ferramentas computacionais para extrair informações automaticamente de

bases de dados científicas, permitindo que análises e estudos mais abrangentes possam ser realizados.

1.1. Motivação

No Brasil, a demanda por financiamento de atividades científicas faz com que seja necessária a comparação de uma quantidade grande de dados utilizados na avaliação de grupos de pesquisadores e instituições. Isso faz necessário que as informações das avaliações sejam obtidas rapidamente e se possível, automaticamente, principalmente quando há uma grande quantidade dessas. Percebe-se então que há a necessidade de ferramentas computacionais que possam auxiliar na obtenção automática de informações sobre pesquisadores, programas de pós-graduação, áreas do conhecimento e até mesmo grandes áreas do saber.

Outra necessidade é entender como ocorre a colaboração científica entre pesquisadores. Segundo Vanz e Stumpf (2010), entendê-la é fundamental para que se tenha uma ideia mais clara de como este fenômeno vem acontecendo na comunidade científica brasileira, permitindo a definição de políticas científicas mais adequadas. Com isso, também é possível verificar como o Brasil está inserido no cenário internacional.

1.2. Objetivos

Este trabalho faz parte de um projeto maior denominado “Sistema Unificado de Currículos e Programas: Identificação de Redes Acadêmicas - SUCUPIRA”. O projeto SUCUPIRA, processo CAPES 23038-029609/2008-02, cujo acrônimo traz a lembrança do sobrenome do falecido professor Emérito da Universidade Federal do Rio de Janeiro, Newton Lins Buarque Sucupira e relator do importante Parecer 977/65 sobre a Pós-Graduação, visa ser uma ferramenta computacional automatizada e de domínio público que pode eventualmente auxiliar na obtenção de indicadores de desempenho de docentes, pesquisadores e programas de pós-graduação.

Um aspecto que foi considerado neste trabalho como parte do Projeto SUCUPIRA é como obter informações em bases públicas de dados científicas nacionais e internacionais. Com essas informações é possível realizar análises e estudos de pesquisadores, instituições, áreas e até mesmo países. Esses estudos podem contextualizar a produção científica brasileira no cenário internacional e permitir que políticas públicas sejam melhor definidas.

1.2.1. Objetivo geral

O objetivo geral deste trabalho é propor de forma automática mapeamentos de produção científica de acordo com dados disponibilizados em bases de dados científicas, utilizando indicadores bibliométricos para que análises e estudos mais abrangentes possam ser realizados de maneira transparente e simples.

1.2.2. Objetivos específicos

Para alcançar o objetivo geral, foram definidos os seguintes objetivos específicos:

- Identificar e analisar as principais bases públicas de dados científicas nacionais e internacionais.
- Analisar as técnicas utilizadas para a extração de informação em documentos Web e definir quais destas são aquelas mais adequadas a serem utilizadas neste trabalho.
- Propor ferramentas computacionais que permitam extrair informações em bases de dados científicas de forma que essas possam ser utilizadas por outros usuários e com uma baixa curva de aprendizado.
- Extrair as informações das bases de dados científicas definidas utilizando as ferramentas computacionais propostas.
- Realizar análises e estudos utilizando as informações extraídas das bases de dados científicas definidas.

1.3. Justificativas

Inicialmente, a ideia deste trabalho era utilizar apenas dados da Plataforma Lattes (PL), que é uma base de dados de currículos de pesquisadores que atuam no Brasil e é mantida pelo CNPq. O currículo Lattes é um dos elementos decisivos no julgamento e avaliação de bolsas, e na captação de recursos financeiros em editais de pesquisa. Além disso, a expressiva maioria dos pesquisadores com doutorado no País possui currículo Lattes, que é necessário para solicitar qualquer tipo de auxílio. Todos os pesquisadores cadastrados em programas de Pós-Graduação possuem o currículo Lattes por exigência da avaliação dos programas realizada pela CAPES. As informações são fornecidas pelos pesquisadores, que utilizam senha para acesso e que precisam atestar formalmente a veracidade das informações prestadas, o que pode torná-las mais precisas.

A PL é hoje, sem dúvida, a principal fonte de informações sobre pesquisadores brasileiros. Porém, isso limitaria o trabalho a estudos envolvendo apenas pesquisadores brasileiros e cadastrados na PL. Portanto, foi necessário considerar outras bases de dados científicas, como, por exemplo, a Scopus e a *Web of Science* (WoS), bases reconhecidas mundialmente pela sua amplitude e tradição em estudos bibliométricos.

Assim, fez-se necessário propor ferramentas computacionais que permitam extrair informações de bases públicas de dados científicas nacionais e internacionais.

É importante destacar não só a importância do Portal de Periódicos da CAPES neste trabalho, mas principalmente para o avanço recente na ciência brasileira. Criado em 2000, o Portal é hoje, sem dúvida, um dos maiores acervos do saber no cenário mundial (ALMEIDA et al., 2010).

1.4. Resultados alcançados

Os principais resultados alcançados com este trabalho são:

- Um conjunto de ferramentas computacionais que permitem extrair informações em bases de dados científicas, possibilitando que análises e estudos mais abrangentes sejam realizados.
- Um índice que permite verificar como ocorreu a colaboração científica entre os autores de um artigo.

De maneira geral, este trabalho pode auxiliar na extração de informações relevantes sobre pesquisadores. É possível analisar a colaboração entre pesquisadores, instituições, áreas e até mesmo países. Também permite verificar o surgimento ou desaparecimento de áreas de pesquisa, possibilitando ao Governo e às agências de fomento saberem onde melhor investir. Também permite explorar diversas questões pontuais, tais como:

- Qual o pesquisador mais produtivo em um dado ano?
- Qual a trajetória e tendência de carreira para um pesquisador com base em seu currículo atual?
- Quais pesquisadores têm maior número de colaboradores?
- Qual o nível de colaboração de um determinado pesquisador?
- Pesquisadores que mais colaboram são também os que mais publicam?
- Pesquisadores que mais colaboram são também os mais citados?
- Qual é o perfil de um pesquisador produtivo analisando a sua produção ao longo do tempo?
- É possível definir qual é realmente a área de atuação de um pesquisador?

- Um grupo é produtivo porque tem vários pesquisadores medianos ou porque tem alguns poucos muito produtivos?
- O ambiente em que estou inserido pode influenciar minha carreira profissional e acadêmica?
- Se trabalho com um pesquisador que está trabalhando ativamente, a minha chance de sucesso na vida acadêmica aumenta?
- Quais os relacionamentos acadêmicos de um pesquisador?

Também podem ser realizados estudos mais abrangentes, tais como:

- Mapeamento de competências;
- Mapeamento geográfico de competências;
- Avaliação de áreas do conhecimento;
- Histórico de evolução de áreas do conhecimento;
- Comparação de grupos de pesquisa (instituições, programas de Pós-Graduação, regiões, países etc.);
- Acompanhamento de egressos de cursos de Pós-Graduação.

Essas são apenas algumas das questões que podem ser exploradas e alguns dos estudos que podem ser realizados.

1.5. Organização

Esta tese está organizada da seguinte maneira: no Capítulo 2 é apresentada uma breve descrição das principais bases de dados com foco em produções científicas do País e também de bases internacionais do saber. No Capítulo 3 é apresentada uma revisão da literatura com os principais tópicos abordados neste trabalho. No Capítulo 4 é apresentada uma visão geral das ferramentas

computacionais desenvolvidas ao longo deste trabalho, destacando suas principais características e ilustrando as suas funcionalidades. No Capítulo 5 é apresentado um novo índice proposto que permite medir a colaboração entre os autores de um artigo. No Capítulo 6 são apresentados alguns estudos de casos, ilustrando como as ferramentas desenvolvidas podem ser utilizadas para realizar análises nas bases de dados consideradas neste trabalho. Finalmente, no Capítulo 7 são apresentadas as conclusões e trabalhos futuros.

2 BASES DE DADOS CIENTÍFICAS

Neste Capítulo são apresentadas as principais bases de dados com foco em produções científicas do País e também de bases internacionais do saber e que são utilizadas neste trabalho. Essas bases permitem acessar informações sobre toda a produção científica indexada aumentando o alcance das pesquisas que estão acontecendo nos países bem como realizar estudos e análises, contribuindo para o melhor entendimento e tendências das diversas áreas do conhecimento. Apresentam-se também as duas ferramentas, do melhor do nosso conhecimento, que permitem extrair informações automaticamente de uma das bases de dados nacionais.

No Brasil, agências de fomento como o CNPq e a CAPES se esforçam para oferecer acesso as principais bases de dados científicas do mundo. Nesse sentido, o Portal de Periódicos da CAPES desempenha papel fundamental à ciência nacional, posto que otimizou a política de acesso atualizado ao conhecimento científico do País. Sua implantação reduziu custos e promoveu acesso universal a um acervo amplo e atualizado de artigos publicados em periódicos internacionais e a bases de dados científicas, a qualquer momento, e sem limitações geográficas. Além disso, preencheu as enormes lacunas nos acervos das bibliotecas (ALMEIDA et al., 2010).

As bases de dados científicas consideradas neste trabalho foram divididas em nacionais e internacionais. Nas nacionais, foram utilizadas a PL e as Bolsas em curso, ambas mantidas pelo CNPq. Também foram utilizados o Qualis Periódicos da CAPES e os Cursos de Pós-Graduação recomendados e reconhecidos pela CAPES. Nas internacionais, foram utilizadas a WoS e o JCR® (*Journal Citation Reports*) ambas mantidas pela Thomson Reuters. Também foram utilizadas a Scopus, mantida pela Elsevier, e o *SCImago Journal & Country Rank*, que é elaborado a partir de dados da própria Scopus. Essas bases de dados foram escolhidas visando análises mais elaboradas, e principalmente, mais abrangentes, tomando-se o cuidado de escolher bases reconhecidas por toda a comunidade científica.

Há outras bases que podem futuramente ser consideradas, porém não foram utilizadas neste trabalho. Uma delas é a Scielo (acessível em http://www.scielo.org/), que é uma biblioteca eletrônica que abrange um conjunto selecionado de periódicos brasileiros. A Scielo teve sua origem no Brasil e já se expandiu para diversos outros países, alcançando uma abrangência internacional. Outra base é o *Google Scholar* (acessível em http://scholar.google.com.br/), que é uma ferramenta do Google que permite pesquisar a literatura acadêmica de forma abrangente. Entretanto, é importante mencionar que o *Google Scholar* ainda não tem um nível de controle que julgamos necessário sobre a qualidade de seus dados para a sua utilização. A maior cobertura que oferece consiste em dados não compatíveis com os fornecidos por outras bases de dados (NORUZI, 2005; AGUILLO, 2012). Além disso, a Google não disponibiliza uma biblioteca para acessar o *Google Scholar* e impede que programas realizem buscas automáticas.

2.1. Bases nacionais

No contexto nacional, quatro bases de dados foram consideradas neste trabalho. Uma delas é a PL, que é uma base de dados de currículos de pesquisadores que atuam no Brasil, mantida pelo CNPq. A PL é hoje, sem dúvida, a principal fonte de informações sobre pesquisadores brasileiros.

Entre as formas de financiamento do CNPq, inclui-se a bolsa de Produtividade em Pesquisa (PQ), que é concedida através de julgamento por pares a pesquisadores que se destacam em suas áreas no Brasil. As bolsas em curso na modalidade PQ oferecem informações básicas sobre os pesquisadores ativos de acordo com a sua área de atuação ou instituição de ensino. Essas informações compõem a outra base considerada neste trabalho.

Outra base de dados considerada neste trabalho é o Qualis Periódicos da CAPES. O Qualis Periódicos é uma classificação dos periódicos científicos em que os docentes de cursos de pós-graduação no País publicaram seus artigos. Esta classificação é baseada em um conjunto de procedimentos estabelecidos

pelas 48 áreas de avaliação da CAPES e utilizados para comparar a produção científica mais relevante dos programas de pós-graduação.

Também foram considerados neste trabalho os Cursos de Pós-Graduação recomendados e reconhecidos pela CAPES. Além de informações básicas sobre os cursos e programas, a CAPES também disponibiliza dados relativos às avaliações que esta realiza.

Todas essas bases de dados são públicas e são mantidas por órgãos do Governo Brasileiro. Embora algumas não sejam propriamente bases de dados, neste trabalho foram tratadas como tal, uma vez que as informações extraídas foram armazenadas em bancos de dados, o que permitiu a realização de diversas análises.

2.1.1. CNPq

Uma das principais e mais antigas agências de fomento à ciência, tecnologia e inovação do Governo Federal no Brasil é o CNPq, sendo uma agência do Ministério da Ciência, Tecnologia e Inovação (MCTI) que tem como principais atribuições fomentar a pesquisa científica e tecnológica e apoiar, ainda que em menor escala que a CAPES, a formação de recursos humanos qualificados no Brasil. Criado em 1951, desempenha papel importante na formulação e condução das políticas de ciência, tecnologia e inovação (CNPq, 2013).

O CNPq concede bolsas com diversas finalidades para todas as áreas do conhecimento, desde a iniciação científica até o mais alto nível, como a PQ e a Bolsa de Desenvolvimento Tecnológico, que valoriza o pesquisador, tendo em consideração a sua produção científica e tecnológica, e o seu projeto a ser desenvolvido. O CNPq oferece ainda várias modalidades de bolsas aos alunos do ensino médio, graduação, pós-graduação, recém-doutores e pesquisadores experientes, tanto para desenvolver atividades no País quanto no exterior.

Nesta Subseção é apresentado um breve resumo da PL e das Bolsas em curso, ambas as bases mantidas pelo CNPq.

2.1.1.1. Plataforma Lattes

A PL é um sistema de informação desenvolvido e implantado pelo CNPq para gerenciar informações relacionadas a pesquisadores e instituições no Brasil (CNPq, 2013). Recentemente, a PL foi citada como exemplo de banco de dados completo e altamente qualificado em um artigo publicado na Nature (LANE, 2010). A PL é mantida pelo Governo Brasileiro e inclui sistemas de informação, bancos de dados e portais Web. O principal componente da plataforma é o sistema **Currículo Lattes**, que é um sistema de informação curricular.

O sistema Currículo Lattes armazena mais de 3.000.000 (em Janeiro de 2014) de currículos de pesquisadores, docentes, estudantes e profissionais das diversas áreas do conhecimento que atuam em ciência, tecnologia e inovação, principalmente no Brasil. Este sistema é hoje, sem dúvida, a principal fonte de informações individuais sobre pesquisadores brasileiros.

O “currículo Lattes” é um documento criado pelo CNPq com o objetivo de padronizar e centralizar informações pessoais, profissionais e acadêmicas da comunidade científica brasileira. Através do sistema Currículo Lattes é possível consultar essas informações a qualquer momento e de maneira muito simples via Web.

O “currículo Lattes” se tornou um padrão nacional no registro da vida acadêmica pregressa e atual dos pesquisadores, e é hoje adotado e exigido o seu preenchimento e atualização pela maioria das agências de fomento, universidades e institutos de pesquisas do país. Por sua riqueza de informações e sua crescente confiabilidade e abrangência, se tornou elemento indispensável para suporte à análise de mérito e competência dos pleitos de financiamentos na área de ciência e tecnologia (CNPq, 2013).

A disponibilização pública dos dados da PL na Web dão maior transparência e mais confiabilidade às atividades de fomento do CNPq e de outras agências que a utilizam, fortaleceu o intercâmbio entre pesquisadores e instituições e é

uma fonte de informações com muito grande potencial para diversos estudos e pesquisas. E na medida em que suas informações são recorrentes e cumulativas, têm também o importante papel de preservar a memória da atividade de pesquisa no país (CNPq, 2013).

Apesar dos dados dos currículos serem preenchidos pelo próprio pesquisador, a comunidade científica monitora a qualidade e a fidelidade das informações contidas no sistema, posto que a transparência pública aliada à depuração dos dados por pares fazem com que haja a necessária estabilidade da PL. Percebe-se, portanto, que esse sistema tem um elevado potencial para extração de informação confiável, embora isso não seja uma tarefa simples e imediata, notadamente quando uma grande quantidade de pesquisadores necessita ser considerada.

Nos últimos anos, muitos trabalhos foram realizados utilizando dados disponíveis na PL e por pesquisadores das mais diversas áreas do conhecimento. Alguns trabalhos analisaram o perfil de pesquisadores bolsistas de produtividade em pesquisa do CNPq em áreas como Saúde Coletiva (BARATA; GOLDBAUM, 2003; SANTOS et al., 2009), Odontologia (SCARPELLI et al., 2008; CAVALCANTE et al., 2008; CAVALCANTI; PEREIRA, 2008; POPOFF et al., 2012), Fisioterapia (COURY; VILELLA, 2009; FREIRE et al., 2013), Medicina (MENDES et al., 2010; MARTELLI JÚNIOR et al., 2010; OLIVEIRA et al., 2011a; OLIVEIRA et al., 2011b; OLIVEIRA et al., 2012; ROMANO-SILVA et al., 2013; OLIVEIRA et al., 2013), Química (SANTOS et al., 2010), Matemática (SILVA, 2011), Educação Física (LEITE et al., 2012), Enfermagem (SOUTO et al., 2012), Medicina Veterinária (SPILKI, 2013) e Psicologia (WENDT et al., 2013); outros mapearam o sexo e a região dos pesquisadores (ARRUDA et al., 2009) ou a correlação estatística entre a produtividade dos pesquisadores e sua proficiência no inglês escrito (VASCONCELOS et al., 2009). Diversas teses (SILVA, 2007; VASCONCELOS, 2008; MOREIRA, 2009; VANZ, 2009; ROSA, 2010; MELO, 2011; ALMEIDA, 2013), dissertações (BALANCIERI, 2004; BOVO, 2004; PAULA, 2004; CARDOSO, 2005; ALMEIDA, 2006; MARINHO, 2007; CASTAÑO, 2008;

CIVIDANES, 2010) e trabalhos de conclusão de curso (KALIL, 2008; NASCIMENTO-JÚNIOR, 2008) também fizeram uso de dados da PL. Além desses, diversos outros trabalhos foram realizados (BORGES et al., 2004; PACHECO et al., 2007; CARDOSO; MACHADO, 2008, entre outros).

Uma tarefa (e, possivelmente, um problema) comum apresentado em quase todos esses trabalhos é que os currículos e as informações extraídas foram obtidas manualmente. Cavalcante et al. (2008) descreveram que levaram quase 3 anos para analisar 132 currículos, o que desencoraja a repetição do processo.

No trabalho pioneiro de Barata e Goldbaum (2003) os autores apontaram a existência de problemas referentes às informações obtidas nos currículos Lattes, principalmente no que diz respeito a diferenças de interpretação no preenchimento das informações pelos pesquisadores. O CNPq vem continuamente realizando esforços para melhorar a entrada de dados no sistema Currículo Lattes buscando uma uniformização e, principalmente, um maior controle na inserção dos dados.

2.1.1.2. Bolsas em curso

Uma das modalidades de concessão de bolsas considerada dentre as mais importantes pela comunidade científica brasileira entre as oferecidas pelo CNPq é a de PQ. Esta bolsa é outorgada aos pesquisadores que a solicitam ao CNPq e se destacam entre seus pares que atuam no Brasil, comparando-se os seus projetos de pesquisa ou desenvolvimento e sua produção científica qualificada, segundo critérios normativos estabelecidos pelo CNPq e os especificados pelos Comitês de Assessoramento do CNPq.

A bolsa PQ é dividida e hierarquizada em 3 categorias: SR, 1 (dividido em quatro níveis: 1A, 1B, 1C e 1D) e 2. Trata-se de um processo baseado em mérito técnico e científico tendo avaliação por pares. Mais da metade dos pesquisadores bolsistas são da categoria 2 e, por ordem léxico-decrescente

para aqueles do nível 1, tem-se a indicação do quão sênior o mesmo é em relação aos seus demais pares.

O valor da bolsa é definido de acordo com a categoria e o nível de cada bolsista, sendo que os bolsistas da categoria 1 também recebem um adicional para apoio às pesquisas ou desenvolvimentos denominado “taxa de bancada”. Existem editais para o financiamento de projetos que exigem que o pesquisador responsável seja da categoria 1; apenas pesquisadores da categoria 1 podem ser membros dos Comitês de Assessoramento do CNPq, e apenas pesquisadores da categoria 1 podem participar das consultas do conselho deliberativo à comunidade científica.

O CNPq disponibiliza no seu sítio uma relação das bolsas em curso dos bolsistas PQ (http://plsql1.cnpq.br/divulg/RESULTADO_PQ_102003.curso) que pode ser consultada pelo nome do pesquisador, pela área do conhecimento, ou pela instituição de origem. Com isso é possível obter uma relação completa de todos os pesquisadores com bolsas PQ ativas no País. Neste sítio, entretanto, poucas informações complementares são fornecidas e não há ligação para o currículo Lattes desses pesquisadores bolsistas.

Outra forma de obter os pesquisadores bolsistas PQ é consultando a PL. Todo bolsista PQ tem essa informação em destaque no início do seu currículo Lattes. Entretanto, existem dificuldades que precisam ser consideradas, por exemplo, como é possível obter todos os pesquisadores bolsistas PQ manualmente? Isso seria extremamente trabalhoso, pois o sistema Currículo Lattes permite apenas a consulta pelo nome do pesquisador ou por assunto, o que dificulta obter uma relação de pesquisadores bolsistas PQ. Além disso, como saber se a relação está completa?

Cabe ressaltar que há problemas em buscas nestas bases de dados que merecem destaque. Na relação de Bolsas em curso constam algumas áreas do conhecimento que não possuem pesquisadores como bolsistas PQ. Por exemplo, na área de “Planejamento Energético” nenhum pesquisador é

retornado, conforme ilustra a Figura 2.1. O mesmo ocorre com a área do conhecimento “Geociências: Geologia e Geografia Física”.

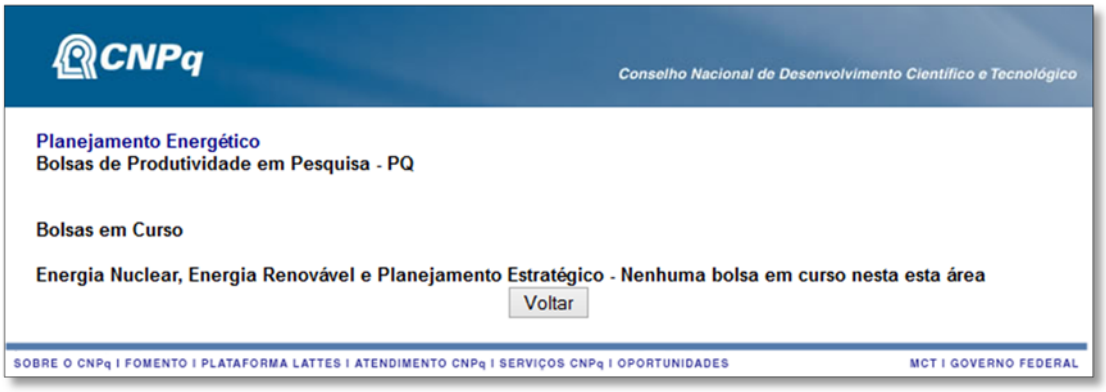


Figura 2.1 - Consulta por área do conhecimento nas Bolsas em curso do CNPq.

Entretanto, se for realizada uma consulta pelo nome do pesquisador “Luiz Pinguelli Rosa” é possível verificar que o mesmo consta como bolsista PQ da área de “Planejamento Energético”, conforme ilustra a Figura 2.2. O mesmo ocorre com o pesquisador “Igor Ivory Gil Pacca”, pois quando realizada uma consulta pelo seu nome ele aparece como bolsista da área de “Geociências”.

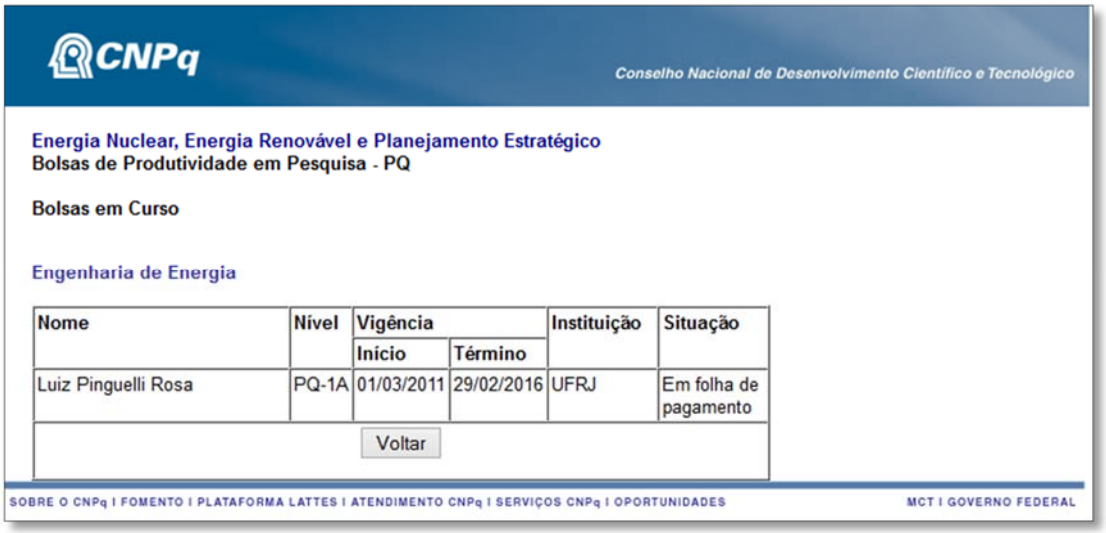


Figura 2.2 - Consulta pelo nome do pesquisador nas Bolsas em curso do CNPq.

Outro problema observado é nomes de pesquisadores constando como bolsistas PQ em consultas em Bolsas em curso e, nos currículos Lattes desses

pesquisadores, não constar que eles são bolsistas. Esse problema pode causar erro na avaliação de um pesquisador e até mesmo em avaliações de programas de pós-graduação.

2.1.2. CAPES

A CAPES é uma agência de fomento do Ministério da Educação à pesquisa brasileira que atua, entre outras coisas, na expansão e consolidação da pós-graduação stricto sensu (mestrado e doutorado) no País e a partir de 2007 da Educação Básica também. Foi criada por Anísio Teixeira em 1951 com o objetivo de "assegurar a existência de pessoal especializado em quantidade e qualidade suficientes para atender às necessidades dos empreendimentos públicos e privados que visam ao desenvolvimento do País" (CAPES, 2013).

Atualmente, as atividades da CAPES podem ser agrupadas nas seguintes linhas de ação (CAPES, 2013): avaliação da pós-graduação stricto sensu; acesso e divulgação da produção científica; promoção da cooperação científica internacional; indução e fomento da formação inicial e continuada de professores para a educação básica nos formatos presencial e a distância. Cada linha de ação é desenvolvida por um conjunto de programas.

Nesta Subseção é apresentado um breve resumo do Qualis Periódicos da CAPES e dos Cursos de Pós-Graduação por ela recomendados e reconhecidos.

2.1.2.1. Qualis Periódicos

Qualis é o conjunto de procedimentos utilizados pela CAPES para estratificação da produção intelectual dos programas de pós-graduação. Esse processo foi concebido para atender as necessidades específicas do sistema de avaliação da CAPES e é baseado nas informações fornecidas por meio do aplicativo Coleta de Dados. Como resultado, disponibiliza uma lista com a classificação dos veículos utilizados pelos programas de pós-graduação para a divulgação da sua produção (QUALIS, 2013). Isso significa que somente

constam no QUALIS os periódicos em que docentes que atuam em programas de pós-graduação credenciados pela CAPES já publicaram.

A estratificação dessa produção é realizada de forma indireta. O Qualis classifica os artigos e outros tipos de produção a partir da estratificação dos periódicos científicos em que esta produção foi veiculada. A classificação de periódicos é realizada pelas áreas de avaliação e passa por processo anual de atualização. Esses veículos são enquadrados em estratos indicativos - A1, o mais valorizado; A2; B1; B2; B3; B4; B5; C - com peso zero (QUALIS, 2013).

Um mesmo periódico pode ser classificado por duas ou mais áreas distintas, podendo receber diferentes avaliações nessas. Segundo a CAPES, isto não constitui inconsistência, mas expressa o valor atribuído, em cada área, à pertinência da política editorial do periódico à área de avaliação (QUALIS, 2013).

O uso do Qualis Periódicos não é adequado para a avaliação individual de pesquisadores. Ele foi concebido para a análise de programas de pós-graduação, e não para a avaliação de pesquisadores.

2.1.2.2. Cursos de Pós-Graduação recomendados e reconhecidos

A CAPES mantém uma lista dos cursos e programas de Pós-Graduação recomendados e reconhecidos, podendo ser consultada de acordo com a área de avaliação, conceito, região ou instituição. Os cursos recomendados são aqueles que já foram oficializados pelo Ministério da Educação e os cursos reconhecidos são aqueles que foram aprovados pela CAPES e encaminhados ao Conselho Nacional de Educação para a instrução de seus processos de reconhecimento (CAPES, 2013).

A classificação dos cursos de pós-graduação é realizada por conceitos que podem variar de 1 a 7. Os conceitos mais baixos, 1 e 2 (insuficiente), são eliminatórios, não sendo credenciado pela CAPES o funcionamento de cursos com esses conceitos; os conceitos 3, 4 e 5 são considerados cursos regulares, bons e muito bons, respectivamente. O conceito 5 é a nota máxima atribuída a

programas que possuam apenas curso de mestrado. Os programas com conceitos mais elevados, 6 e 7, são os reconhecidos pela CAPES como de desempenho equiparados a cursos internacionais de excelência, na mesma área. Atualmente, há no Brasil 320 (8,56%) cursos de Pós-Graduação recomendados e reconhecidos pela CAPES com os conceitos 6 e 7.

2.2. Bases internacionais

No contexto internacional, foram consideradas quatro bases de dados neste trabalho: WoS, Scopus, JCR® e o *SCImago Journal & Country Rank*.

As bases de dados WoS e Scopus, das editoras Thomson Reuters e Elsevier, contêm informações sobre a produção científica em nível mundial. Os dados das citações da WoS são integrados ao JCR® para cálculo do fator de impacto e de outras métricas, e os dados das citações da Scopus são integrados ao *SCImago Journal & Country Rank*. A WoS e a Scopus são as duas principais bases de dados de citações que são frequentemente utilizadas para classificar a relevância dos periódicos, bem como o total de citações recebidas, de modo a indicar o impacto, a influência ou o prestígio dos periódicos (ABRIZAH et al., 2013). Índices como o *SCImago Journal Rank* (SJR), da base de dados Scopus da Elsevier e o JCR®, da base de dados WoS da Thomson Reuters, apresentam indicadores de impacto dos periódicos por meio de estatísticas baseadas em dados de citações.

As bases de dados WoS, Scopus e JCR® podem ser acessadas por meio do Portal de Periódicos da CAPES. O *SCImago Journal & Country Rank* pode ser acessado livremente.

2.2.1. Thomson Reuters

A Thomson Reuters é a maior agência internacional de notícias e multimídia do mundo, fruto da fusão da canadense Thomson Corporation com a britânica Reuters. É uma empresa especializada em informações para empresas e profissionais, que combina a experiência no mercado com a tecnologia para fornecer informação crítica que contribua nas tomadas de decisões nos

mercados: financeiro, jurídico, fiscal e contábil, científico e de saúde, com o respaldo da organização internacionalmente reconhecida, a Reuters.

A WoS é um dos produtos da Thomson Reuters e serve de base para o JCR®. Ambas as bases de dados são apresentadas nas subseções seguintes.

2.2.1.1. Web of Science

A WoS é uma base de dados multidisciplinar mantida pela Thomson Reuters com informações sobre artigos publicados, a partir de 1945, em mais de 12.000 periódicos em todas as áreas do conhecimento (WOS, 2013). Permite a recuperação de artigos publicados em periódicos internacionais, apresentando as referências bibliográficas e informando sobre os documentos que os citaram, com referências a outros documentos. Além disso, a WoS oferece registros bibliográficos padronizados, dando a possibilidade de utilizar esses dados em outras ferramentas. Também oferece informações sobre o impacto e a visibilidade das publicações nela indexadas.

A WoS é uma das mais antigas bases de dados científicas. A base foi criada em 1958, na Filadélfia (Estados Unidos), por Eugene Garfield (um dos pioneiros da bibliometria) com o objetivo de proporcionar acesso à informação de relevância e conteúdo de qualidade para pesquisadores em todo o mundo (WOS, 2013).

A WoS consiste de três bases distintas que podem ser consultadas individualmente ou combinadas: 1) o "*Science Citation Index Expanded*" (SCIE), editado desde 1961, indexa mais de 8.500 títulos de periódicos internacionais das áreas de Ciências Exatas e Biológicas; 2) o "*Social Sciences Citation Index*" (SSCI), lançado em 1972, indexa mais de 3.000 periódicos das áreas de Ciências Sociais; e 3) o "*Arts & Humanities Citation Index*" (AHCI), criado em 1978, indexa mais de 1.700 periódicos da área de Artes e Humanidades (WOS, 2013).

2.2.1.2. Journal Citation Reports®

O JCR® oferece um modo sistemático e objetivo de avaliar, em termos de citações, os principais periódicos de pesquisa do mundo. Com recursos que permitem analisar e comparar o desempenho de periódicos por meio da informação estatística baseada em dados de citação, o JCR® divulga todos os anos (normalmente, a primeira versão é divulgada em meados do ano e a segunda, com eventuais correções, alguns meses depois) o fator de impacto e outros indicadores bibliométricos para todos os periódicos indexados na base WoS (JCR, 2013).

O JCR® também permite verificar os periódicos mais citados de uma determinada área. Atualmente, o JCR® cobre mais de 10.800 periódicos de mais de 2.550 editoras em aproximadamente 232 categorias de 83 países e permite acesso à estatística de citações desde 2007 até o presente. O JCR® é fornecido em duas edições: "*Science Edition*", com dados de mais de 8.400 periódicos e "*Social Science Edition*", com dados de mais de 3.000 periódicos (JCR, 2013). É válido lembrar que um mesmo periódico pode estar indexado em ambas as edições.

O JCR® serve também como ferramenta auxiliar ao pesquisador na determinação de títulos de periódicos para publicação de seus trabalhos e às bibliotecas. Observa-se que o Brasil contava com 114 periódicos indexados no JCR® de 2012.

2.2.2. Elsevier

A Elsevier é uma das mais antigas e conceituadas editoras do mundo nas áreas de ciência, tecnologia e saúde. Criada em 1880, em Amsterdã na Holanda, a Elsevier provém da editora familiar "*House of Elzevir*", criada em 1580. Evoluiu de uma pequena editora dedicada à publicação de estudos acadêmicos a uma editora multimídia internacional com mais de 20 mil produtos voltados à comunidade científica e médica mundial (ELSEVIER, 2013).

A editora Elsevier está presente em 24 países e atende a uma comunidade de 30 milhões de cientistas, estudantes e profissionais de informação e saúde em todo o mundo. Anualmente, a Elsevier publica mais de 2.000 periódicos e 1.900 livros. Um dos seus principais produtos é a base de dados Scopus, cujas informações servem de base para o SCImago.

2.2.2.1. Scopus

A Scopus é uma base de dados multidisciplinar, mantida pela Elsevier, com cobertura desde 1960, que contém cerca de 50 milhões de registros com resumos, citações e textos completos de aproximadamente 21.000 periódicos de mais de 5.000 editoras internacionais. A Scopus abrange todas as áreas do conhecimento, com mais de 6.800 periódicos de Ciências da Saúde, mais de 7.200 periódicos de Ciências Físicas, mais de 5.300 periódicos de Ciências Sociais e mais de 4.300 periódicos de Ciências da Vida (SCOPUS, 2013).

Apesar de ter sido criada em 2004, a Scopus possui uma cobertura muito maior do que a WoS e a exemplo desta última é considerada atualmente uma das maiores bases de dados científicas do mundo. Além disso, a sua atualização também ocorre de forma bem mais rápida, uma vez que a Scopus atualiza sua base diariamente (SCOPUS, 2013) enquanto que na WoS a atualização é feita semanalmente (WOS, 2013). A Scopus também é uma ferramenta para estudos bibliométricos e avaliações de produção científica em nível mundial.

2.2.2.2. SCImago Journal & Country Rank

O *SCImago Journal & Country Rank* é uma plataforma que inclui indicadores de periódicos e países, obtidos a partir de informações da base de dados Scopus. Esses indicadores podem ser utilizados para avaliar e analisar o impacto da produção científica em todo o mundo. Essa plataforma foi criada em 2007 na Espanha e tem seu nome a partir do índice SJR (SCIMAGO, 2013).

Além de ser uma ferramenta com acesso totalmente aberto, outra vantagem é a quantidade de indicadores oferecidos, o que permite realizar diversas

análises, como por exemplo, o índice H agregado de um país. Também é possível exportar facilmente os dados para planilhas.

2.3. Ferramentas para Extração de Informações

Do melhor do nosso conhecimento, há duas ferramentas que permitem extrair informações do sistema Currículo Lattes de forma automática: Lattes Extrator e scriptLattes.

2.3.1. Lattes Extrator

Lattes Extrator é uma ferramenta acessível via Web (http://lattesextrator.cnpq.br/lattesextrator/) que foi desenvolvida pelo próprio CNPq e é umas das ferramentas que compõe a PL. O acesso é restrito a instituições licenciadas que podem extrair informações somente de seus próprios pesquisadores, docentes, estudantes e colaboradores (CNPq, 2013). As informações são extraídas diretamente do banco de dados do sistema Currículo Lattes e disponibilizadas em arquivos no formato XML definido pela comunidade LMPL (Linguagem de Marcação da Plataforma Lattes) (PACHECO; KERN, 2001). Dessa forma, as instituições precisam desenvolver rotinas para a importação dessas informações para as suas próprias bases. As extrações são feitas em lote e podem ser configuradas de acordo com o interesse e as permissões de cada usuário. A Figura 2.3 ilustra a página inicial da ferramenta Lattes Extrator.

Figura 2.3 - Página inicial da ferramenta Lattes Extrator.

2.3.2. scriptLattes

scriptLattes é um script desenvolvido em Python para extração e compilação de produções bibliográficas, produções técnicas, produções artísticas, orientações, projetos de pesquisa, prêmios e títulos, grafos de colaborações, mapa de geolocalização e coautoria, e internacionalização de um grupo de pesquisadores cadastrados na PL (MENA-CHALCO; CESAR-JUNIOR, 2009).

A primeira versão, lançada em 2005, foi desenvolvida para auxiliar a secretaria do Programa de Pós-Graduação do IME-USP na elaboração de relatórios sobre a produção bibliográfica dos docentes do Departamento de Ciência da Computação. Esses relatórios foram baseados nas informações cadastrados nos currículos Lattes desses docentes. Atualmente, os relatórios podem ser gerados em português, inglês e espanhol.

Para executar o script é necessário criar um arquivo no formato texto contendo os números associados aos pesquisadores os quais são gerados pela PL. Esse número contém 16 dígitos e é utilizado como um identificador (ID) para cada currículo Lattes. Opcionalmente, também podem ser informados o nome do

pesquisador, o período que se deseja considerar e um rótulo que é utilizado para identificar o pesquisador na visualização do grafo de colaborações, em que cada rótulo é representado por uma cor diferente. A versão 8.01 do scriptLattes suporta até 21 rótulos diferentes.

Em seguida, é necessário definir os parâmetros no arquivo de configurações do scriptLattes. Esses parâmetros permitem a geração de relatórios e grafos de colaborações. Feito isso, o script baixa automaticamente os currículos em formato HTML, compila as listas de publicações e orientações e gera páginas Web contendo essas informações separadas por tipo e colocadas em ordem cronológica invertida, um grafo de colaborações entre os pesquisadores e um mapa de geolocalização.

A Figura 2.4 ilustra a página inicial gerada após a compilação dos currículos do Grupo de Visão e Processamento de Imagens (IME-USP) pelo script. A partir dessa página é possível acessar outras páginas clicando nos links disponíveis.

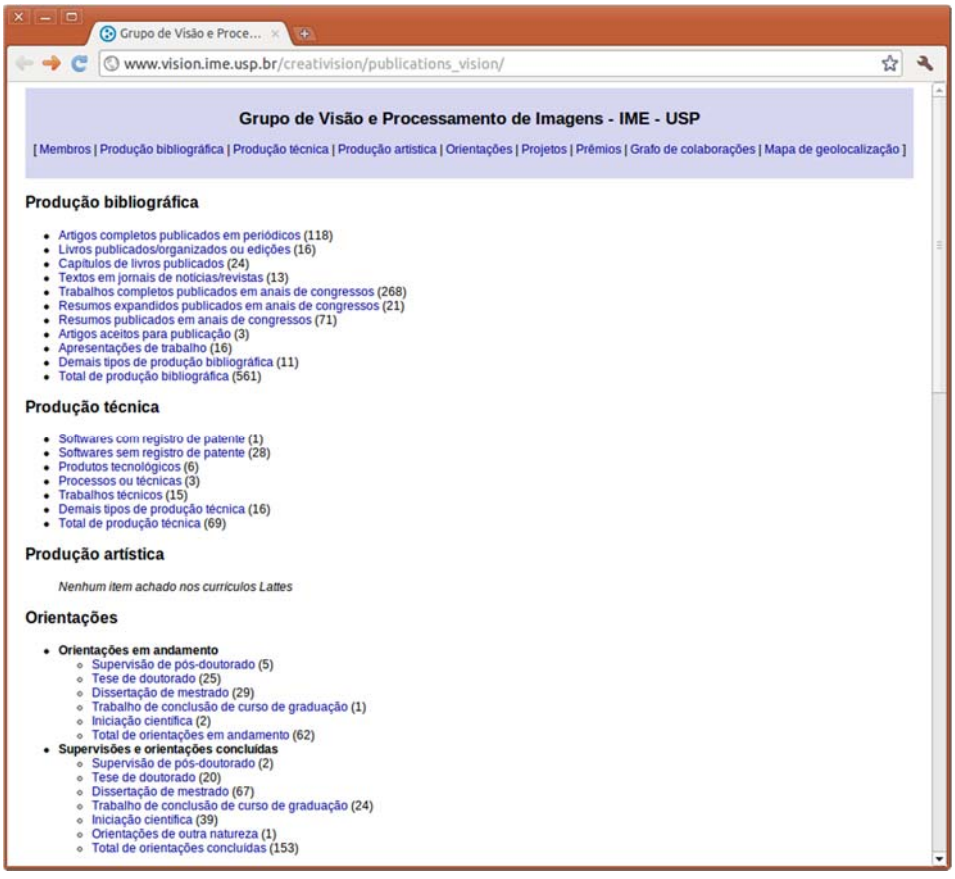


Figura 2.4 - Página inicial gerada pela ferramenta scriptLattes.

A Figura 2.5 ilustra uma das páginas geradas pela ferramenta scriptLattes que contém um relatório de publicações de acordo com os parâmetros definidos no arquivo de configurações. Um gráfico com o número de publicações por ano também é apresentado.

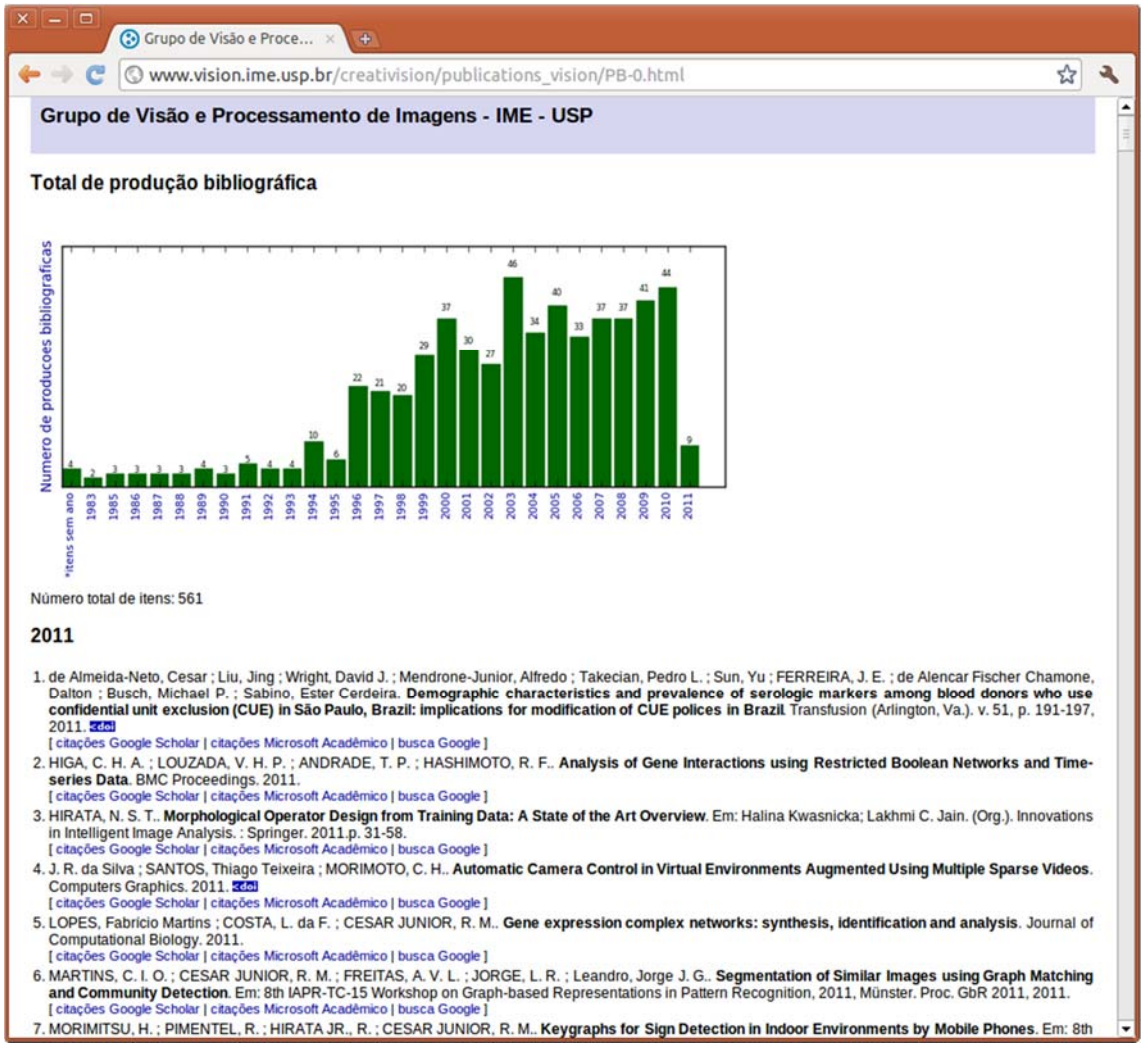


Figura 2.5 - Página contendo um tipo de relatório de publicações gerado pela ferramenta scriptLattes.

A Figura 2.6 ilustra uma página contendo um grafo de colaborações obtido a partir de relações entre os pesquisadores e um mapa de geolocalização. O grafo foi gerado considerando publicações com títulos iguais ou similares (dentro do mesmo tipo e ano de publicação) e o número de relações encontradas entre os pesquisadores pode ser exibido nas arestas. O grafo é estático, ou seja, não permite qualquer tipo de interação com o usuário desta

ferramenta; permitindo apenas clicar nos nomes dos pesquisadores. Esta ação abre uma página contendo o currículo Lattes do pesquisador em que o nome foi clicado.

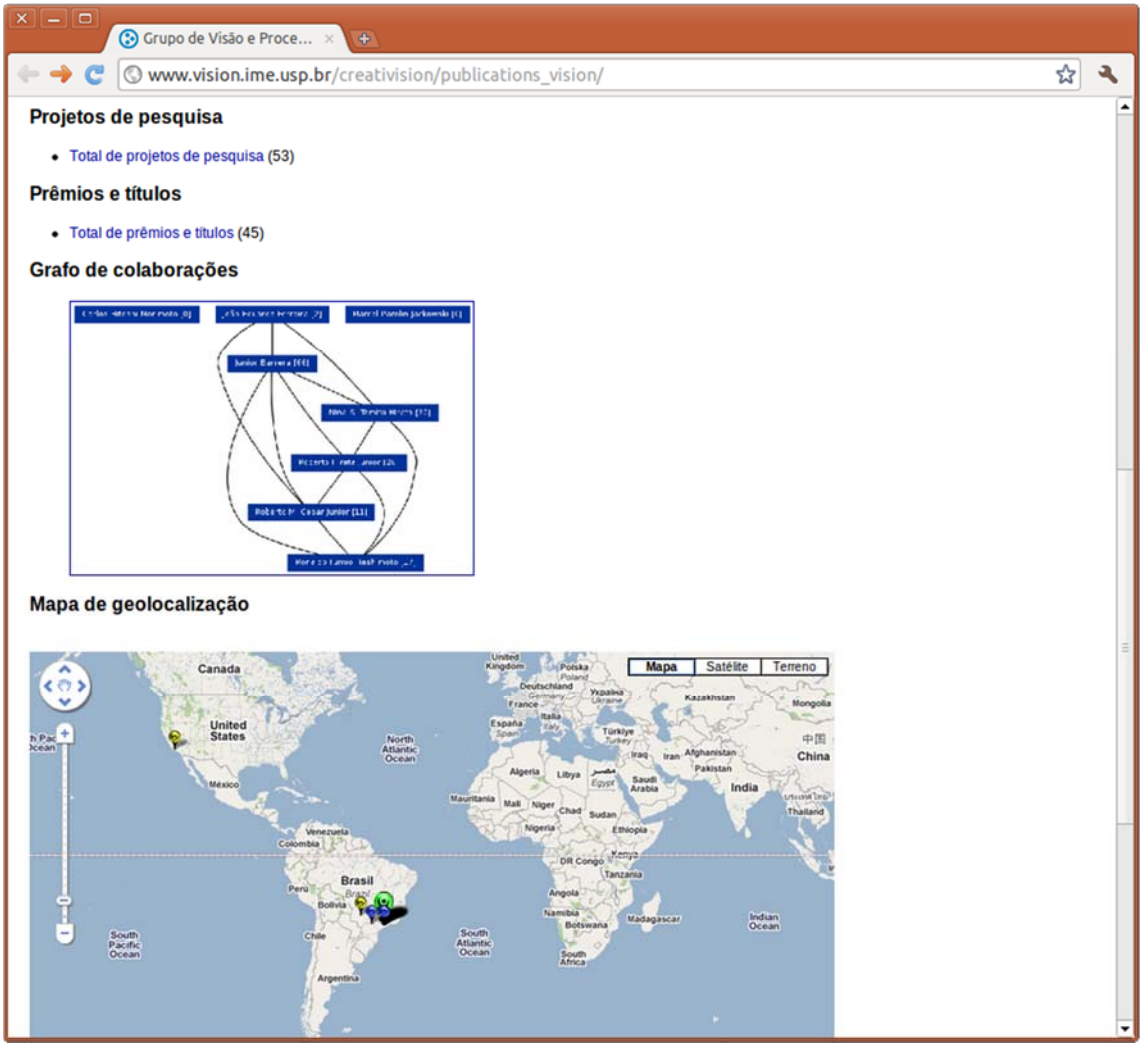


Figura 2.6 - Página contendo um grafo de colaborações e um mapa de geolocalização gerados pela ferramenta scriptLattes.

O mapa de geolocalização é gerado baseando-se nos CEPs (Código de Endereçamento Postal) cadastrados nos currículos Lattes dos pesquisadores, calculando a latitude e a longitude de cada endereço. O mapa é exibido utilizando a API (*Application Programming Interface*) do *Google Maps*, sendo necessário criar uma chave (*google-map-key*) para acessar as páginas do *Google Maps*.

A licença da ferramenta scriptLattes é GPL (*General Public License*) e a ferramenta é executada apenas no sistema operacional Linux. É necessário ter um compilador Python configurado e alguns módulos instalados para a geração de gráficos. Para utilizar o scriptLattes em outros sistemas operacionais é necessário compilar o código-fonte e configurar o ambiente.

2.4. Considerações finais

Este Capítulo apresentou um breve resumo das principais bases de dados científicas. Essas bases permitem que diversas análises sejam realizadas, englobando inclusive toda a produção científica mundial indexada. É importante destacar que essas bases se complementam, aumentando a abrangência dos possíveis estudos. As informações contidas nessas bases possibilitam entender como uma determinada área do conhecimento varia ao longo dos anos. Com isso, é possível identificar tendências e até mesmo permitir a definição de políticas de alocação de recursos para o financiamento de pesquisas técnico-científicas e a formação de recursos humanos qualificados. Também foram apresentadas duas ferramentas computacionais que permitem extrair informações de forma automática da PL. No próximo Capítulo é apresentada uma revisão da literatura com os principais tópicos abordados neste trabalho.

3 REVISÃO DE LITERATURA

Neste Capítulo é apresentada uma visão geral sobre Extração de Informação, destacando algumas técnicas básicas utilizadas para extraí-las a partir de documentos da Web. Também são introduzidos os conceitos necessários para este trabalho sobre Linguagem de Domínio Específico, destacando suas vantagens e desvantagens. Também são apresentados alguns conceitos básicos sobre Análise de Redes Sociais, visando o entendimento de alguns termos utilizados neste trabalho. Também é apresentado um resumo sobre a Lei de Benford que mostra que os algarismos mais significativos não aparecem uniformemente distribuídos. Por fim, são apresentados os principais indicadores bibliométricos utilizados para analisar a produção científica. Todos esses tópicos foram importantes para o desenvolvimento deste trabalho.

3.1. Extração de Informação

Extração de Informação (EI) é o processo de obtenção de informação a partir de documentos: não estruturados (ou livres), semiestruturados e estruturados (VADREVU et al., 2007). As informações extraídas podem ser exibidas diretamente aos usuários ou podem ser armazenadas em bases de dados ou em planilhas, para uso posterior em outras aplicações.

O desafio tecnológico associado à EI está fortemente correlacionado às características do tipo de documento do qual a informação é extraída. Por exemplo, em documentos estruturados, como em XML, a EI pode ser realizada de maneira direta utilizando técnicas básicas de *parser*. Porém, quando o documento não possui nenhuma estrutura, por exemplo, com marcadores não completos, a EI é feita utilizando técnicas de Processamento de Linguagem Natural (PLN) (XIAO et al., 2004).

De acordo com Silva et al. (2005), um texto estruturado segue um formato rígido, o que possibilita que a informação seja extraída utilizando regras baseadas em delimitadores e/ou na ocorrência de termos. Nos textos livres têm-se, basicamente, sentenças de alguma língua natural, o que inviabiliza a

extração com base apenas na formatação. Textos semiestruturados, por sua vez, apresentam estruturação (*e.g.*, referências bibliográficas), juntamente com irregularidades, como campos ausentes ou com valor nulo, variações na ordem dos dados e ausências de delimitadores entre as informações a serem extraídas.

Técnicas de PLN são comumente utilizadas para tratar textos livres e aliadas às técnicas de Inteligência Artificial (IA) para textos estruturados e semiestruturados. As técnicas de PLN podem, eventualmente, lidar com as irregularidades de estrutura dos documentos das línguas naturais. No caso das técnicas de IA, podem ser citados os sistemas baseados em regras de extração definidas manualmente através de engenharia do conhecimento. Embora esses sistemas apresentem bons resultados, é preciso trabalho manual e a existência de bons especialistas, o que dificulta sua aplicação a novos domínios. Para minimizar essas dificuldades, algoritmos de aprendizado de máquina são utilizados para tentar obter regras de extração de forma automática (SILVA et al., 2005).

Embora a EI exista há vários anos, seu estudo se concentra principalmente em textos não estruturados (CHANG et al., 2003). Por outro lado, a informação na Web, em sua maioria, está organizada de forma semiestruturada, ou seja, em tabelas, listas enumeradas e itemizadas. Segundo Chang et al. (2003), uma diferença importante entre páginas Web semiestruturadas e estruturadas é que os formatos de layout das páginas semiestruturadas são em princípio exclusivos para cada sítio Web. Dessa forma, nenhuma gramática genérica (no sentido de Linguagens Formais) pode descrever todos os possíveis formatos de layout de forma que se possa ter um único extrator para qualquer página Web semiestruturada. Isso resulta na necessidade de extratores específicos para formatos diferentes, o que faz com que a sua programação manual seja impraticável.

A tarefa de EI de conteúdos Web difere da tarefa de EI tradicional porque o objetivo desta última é extrair dados de textos não estruturados que são escritos livremente em língua natural, enquanto que a EI de conteúdos Web

processa documentos que são semiestruturados e normalmente são gerados automaticamente. Como resultado a EI tradicional utiliza técnicas de PLN tais como gramáticas, enquanto que EI de conteúdos Web aplica técnicas de aprendizagem de máquina e mineração de dados para explorar os padrões sintáticos ou estruturas de layout dos documentos baseados em *templates* (KAYED; SHAALAN, 2006).

As dificuldades encontradas em uma tarefa de EI normalmente estão relacionadas com a entrada de dados, que pode ser estruturada, semiestruturada ou texto livre (não estruturada). Porém, a definição de como o dado está estruturado ou não, varia de acordo com o domínio de pesquisa e com o ponto de vista dos pesquisadores (KAYED; SHAALAN, 2006).

Por exemplo, de acordo com Kayed e Shaalan (2006) alguns pesquisadores defendem que: a informação armazenada em uma base de dados é estruturada; XML são dados semi-estruturados, pois os dados aparecem junto ao esquema da informação; páginas Web são dados não estruturados, pois não existe nenhuma indicação do tipo de dado. Já do ponto de vista de Kayed e Shaalan (2006), documentos XML são considerados documentos estruturados, pois existe um DTD (*Document Type Definition*) ou XML *Schema* disponível que descreve os dados; textos livres são não-estruturados, pois requerem PLN e páginas Web são dados semiestruturados, pois os dados nelas embutidos são regularmente definidos através de *tags* HTML.

Com relação ao nível de interação com o usuário, os sistemas de EI podem ser divididos em quatro classes: construídos manualmente, supervisionados, semissupervisionados e não supervisionados (KAYED; SHAALAN, 2006).

Programas que realizam a tarefa de EI são conhecidos como extratores ou *wrappers*. Um *wrapper* é definido como um componente em um sistema de integração de informação cujo objetivo é fornecer uma interface de consulta única e uniforme para acessar múltiplas fontes de informação (KAYED; SHAALAN, 2006).

Sistemas supervisionados recebem como entrada um conjunto de páginas Web rotuladas com exemplos dos dados a serem extraídos e submete a um *wrapper*. O usuário fornece um conjunto inicial de exemplos rotulados e o sistema pode sugerir páginas adicionais para o usuário rotular através de uma interface gráfica. Para tais sistemas, os usuários que não são programadores podem ser treinados para utilizar a interface gráfica de rotulação, reduzindo assim o custo de geração do *wrapper*.

Ao contrário dos sistemas supervisionados, os sistemas semissupervisionados aceitam exemplos incompletos e não exatos dos usuários para a geração das regras de extração. Os usuários têm que especificar as regras de extração depois da fase de aprendizagem através de uma interface gráfica. Já os sistemas não supervisionados não utilizam exemplos de treinamento rotulados e não têm interação com o usuário para gerar um *wrapper*.

As técnicas básicas de EI são: extração de texto completo, extração com similaridade e expressões regulares (XIAO et al., 2004). A forma simples é a busca por texto completo, de forma que o texto ou a palavra exata seja encontrada. Essa técnica é rápida, simples e pode se adequar aos tipos de informação que podem ser representados como listas de palavras (*strings*) pré-definidas, listas de nomes ou constantes. Entretanto, um problema dessa técnica é que, frequentemente, existem variações para uma mesma palavra. Uma solução é encontrar *strings* utilizando uma medida de similaridade para aceitar variações nessas.

Expressão regular é um padrão genérico que descreve um conjunto de instâncias de *strings*. Elas são adequadas para encontrar conteúdo com propriedades sintáticas significativas (tais como um número, data, hora, preço etc.) e são definidas utilizando vários operadores para combinar expressões menores. Em geral, o processamento de expressões regulares é muito rápido, pois a expressão pode ser compilada para uma rede de transição de estados finitos e nenhum conhecimento anterior ou léxico é necessário. Entretanto, expressões regulares altamente aninhadas e expressões contendo grande número de opções podem ser computacionalmente caras.

Segundo Kayed e Shaalan (2006), a maioria dos sistemas utiliza regras de extração que são representadas como gramáticas regulares para identificar o início e o fim de cada dado relevante. Além disso, regras de expressão regular são utilizadas para entradas semiestruturadas, especialmente páginas Web baseadas em *templates*.

Um problema é que páginas Web estão em constante atualização, ou seja, o código HTML que gera as páginas, frequentemente, é alterado. Isso pode impedir que informações sejam automaticamente extraídas por ferramentas, principalmente, se fazem uso de expressões regulares. É importante lembrar que uma página Web é projetada para ser vista pelo usuário e não para a extração de informações.

3.2. Linguagem de Domínio Específico

Uma LDE ou DSL (*Domain-Specific Language*) é uma linguagem de programação ou uma linguagem de especificação executável que oferece, por meio de notações e abstrações, poder de expressão focado a um único domínio (DEURSEN et al., 2000). Uma LDE deve ter somente um conjunto restrito de notações e abstrações, sendo usualmente declarativa, podendo ser vista como linguagem de especificação bem como linguagem de programação. Exemplos de LDEs incluem: SQL, HTML, LaTeX e até mesmo visuais como UML.

Uma LDE tem como objetivo resolver um problema em particular, tornando-a mais acessível ao público comparada às linguagens de programação tradicionais (TAHA, 2008). O processo de aprendizagem poderia ser bem mais rápido e intuitivo, uma vez que a linguagem poderia ser mais próxima da língua dos usuários não exigindo especialistas em programação.

O processo de aprendizagem de uma linguagem de programação é uma tarefa difícil e demorada. A maioria das linguagens deixa de lado aspectos de comunicação, ou seja, os usuários têm que aprender uma linguagem de comunicação com a máquina totalmente diferente da que utilizam no dia a dia.

Além disso, as linguagens de programação atuais estão se tornando cada vez mais difíceis de serem usadas, devido ao grande número de bibliotecas que são acrescentadas a cada nova versão. Tudo isso é ainda mais agravado quando os usuários não têm nenhum tipo de experiência em programação, sendo apenas especialistas no domínio (SILVA; PINHEIRO, 2004).

Uma LDE apresenta as seguintes vantagens (DEURSEN et al., 2000):

- Permite que as soluções sejam expressas no idioma e no nível de abstração do domínio do problema. Consequentemente, os próprios especialistas do domínio podem entender, validar, modificar e até mesmo, desenvolver programas em uma LDE.
- Programas são concisos, auto-documentados e podem ser reusados para diferentes propósitos.
- Aumenta a produtividade, confiabilidade, manutenibilidade e portabilidade.
- Incorpora conhecimento do domínio, permitindo a conservação e reutilização deste conhecimento.
- Permite validação e otimização dos programas em seu próprio domínio.

As desvantagens do uso de uma LDE são (DEURSEN et al., 2000):

- Os custos de projeto, implementação e manutenção.
- Os custos de adaptação dos usuários.
- A disponibilidade limitada.
- A dificuldade na definição do próprio escopo.
- Possível perda de desempenho comparado com uma linguagem de propósito geral.

Normalmente, o desenvolvimento de uma LDE envolve as seguintes etapas (DEURSEN; KLINT, 1998; DEURSEN et al., 2000):

- **Análise**

  (1) Identificar o domínio do problema.

  (2) Obter todo o conhecimento relevante sobre o domínio.

  (3) Incorporar nesse conhecimento noções semânticas e operações.

  (4) Projetar uma LDE que descreva concisamente aplicações no domínio.

- **Implementação**

  (5) Construir uma biblioteca que implemente noções semânticas.

  (6) Projetar e implementar um compilador que traduz programas na LDE para uma sequência de chamadas a biblioteca.

- **Uso**

  (7) Escrever programas na LDE para todas as aplicações desejadas e compilá-los.

Há duas abordagens para a definição de uma LDE: externa e interna (FOWLER, 2009). Uma LDE externa é uma linguagem completamente separada, para a qual é preciso criar um compilador ou interpretador para implementar sua semântica de execução. Ao contrário, uma LDE interna utiliza como ponto de partida a sintaxe de uma linguagem de propósito geral. Dessa forma, a LDE fica limitada a sintaxe da linguagem utilizada. Também é conhecida como LDE embutida ou embarcada (*embedded DSL*) (FREEMAN; PRYCE, 2006). Uma vantagem dessa abordagem é que é utilizado o compilador ou interpretador da linguagem em questão. A principal limitação é a perda de expressividade devido à sintaxe da linguagem utilizada (KOSAR et al., 2008).

3.3. Análise de Redes Sociais

Uma rede social é um conjunto de atores e das relações definidas entre eles (WASSERMAN; FAUST, 1994). Um ator pode representar um indivíduo (*e.g.*, um pesquisador ou uma instituição) ou pode representar um grupo (*e.g.*, todas as instituições federais do País). As relações são geralmente definidas por vínculos (*e.g.*, profissional ou familiar) e podem ser direcionais ou não direcionais. No primeiro caso, um ator funciona como transmissor e o outro como receptor (*e.g.*, orientação acadêmica). No segundo, a relação é recíproca (*e.g.*, publicação).

Redes sociais podem ser representadas graficamente por sociogramas. Um sociograma é um grafo em que os vértices representam atores e as arestas representam as relações entre os atores. O estudo de redes na forma de grafos é um dos pilares da Matemática Discreta e teve o seu início por volta de 1736, quando Leonard Euler propôs uma solução para o problema das pontes de Königsberg, originando a “teoria dos grafos” (NEWMAN et al., 2006).

O estudo de redes sociais no contexto deste trabalho foi iniciado em meados de 1930 por Moreno (1934), quando sociólogos utilizavam essas redes com a finalidade de estudar o comportamento da sociedade e a relação entre as pessoas. Um estudo importante nesta área foi o de Milgram (MILGRAN, 1967) através de suas experiências que levaram ao conceito de “mundo pequeno” (*small world*). Milgram enviou uma correspondência a um grupo de voluntários que tinham que fazer com que esta chegasse às mãos do destinatário especificado no envelope. As regras determinavam que os voluntários fizessem a correspondência chegar ao destinatário através de quem o conhecia pessoalmente, não sendo permitido utilizar o correio. E para saber quantas pessoas tinhas sido necessárias, Milgram determinou que cada um escrevesse seu nome na correspondência, possibilitando o monitoramento do caminho percorrido.

Inicialmente, Milgram acreditava que as correspondências chegariam ao seu destinatário passando por 100 pessoas aproximadamente. Entretanto, à

medida que as correspondências foram chegando ao destinatário especificado, a maioria havia passado apenas por 6 pessoas, em média. Dessa forma, surgiu o conceito de “seis graus de separação”. Recentemente, Watts e Strogatz (1998) tentaram comprovar que as pessoas estão separadas por seis graus (no máximo), ou seja, pessoas aparentemente sem relação alguma têm grande probabilidade de possuírem, em algum grau, amigos em comum que as aproximem.

A partir desses trabalhos pioneiros, desenvolveu-se um novo método de análise denominado “Análise de Redes Sociais - ARS” (*Social Network Analysis - SNA*). A ARS é uma abordagem oriunda da sociologia, da psicologia social e da antropologia, podendo ser aplicada no estudo de diferentes situações e tendo como foco as relações entre os atores (WASSERMAN; FAUST, 1994).

O entendimento da estrutura e da interação em uma rede social pode ser obtido através de métricas, também, chamadas de propriedades de redes sociais (WASSERMAN; FAUST, 1994). Estas propriedades se dividem em dois grupos: as relacionadas a um ator e as relacionadas à própria rede em sua totalidade.

As propriedades relacionadas a um ator se baseiam nas ligações existentes entre os atores e suas relações. Dessa forma, cada ator possuirá uma ponderação própria - valor - na rede que será considerada ao analisá-lo em relação aos demais. Essas propriedades dizem respeito à centralidade dos atores em relação à rede, isto é, a intensidade com a qual este ator está envolvido em relacionamentos com outros atores, tornando-o mais visível aos outros atores da rede.

Segundo Wasserman e Faust (1994), há quatro tipos de propriedades de centralidade de um ator:

- Grau do ator na rede: o ator mais central é o que possui maior grau, ou seja, o que possui maior quantidade de conexões dentro da rede em relação aos outros.

- Centralidade de intermediação: é a capacidade que um ator tem em intermediar as comunicações entre os demais. Corresponde às interações entre dois atores não vizinhos que dependem dos atores que se localizam entre eles. Os atores que estão entre os atores não adjacentes, possuem controle sobre as interações entre os dois atores não vizinhos.

- Para ter uma alta centralidade de intermediação, um ator deve estar no caminho entre diversos outros atores.

- Centralidade de proximidade: é baseada na distância e representa o quão próximo um ator está de todos os outros, resultando em eficácia na comunicação e pouco esforço em se comunicar com toda a rede. É calculada por meio do menor caminho existente entre dois atores. Indica a capacidade de um ator alcançar todos os demais atores na rede.

- Centralidade da informação: generaliza a noção de centralidade de intermediação em todos os caminhos entre os atores, dando valores às relações dependendo do tamanho de cada caminho. Dessa forma, se um ator origem possui um alto grau de centralidade da informação, pode-se dizer que a soma das relações percorridas para chegar do ator origem ao ator destino é baixa.

Segundo Wasserman e Faust (1994), as propriedades relacionadas à rede como um todo são:

- Densidade da rede: esta propriedade está relacionada ao número de relações que mantêm os atores interligados na rede. Quanto mais relações possuir esta rede, mais densa ela será. O limite máximo de densidade é alcançado quando todos os atores estabelecem relações com os demais.

- Transitividade da rede: a transitividade é o quão difundida é a conexão de um ator em relação à média dos demais atores da rede. Esta propriedade tem como exemplo as relações conhecidas como “amigo do amigo”: se A se relaciona com B e B se relaciona com C, então A se relaciona com C.

- Reflexividade da rede: esta propriedade é caracterizada pela ocorrência de grupos altamente interconectados dentro da rede.

Outro conceito importante em ARS é o de clique, pois permite analisar a coesão de um grupo ou subgrupo. Uma clique é um subgrafo completo que não está contido em qualquer outro subgrafo completo distinto do grafo original. Entre todas as cliques, a de maior cardinalidade é a clique máxima e a maximal é aquela que não se pode adicionar mais vértices (WASSERMAN; FAUST, 1994).

De uma maneira geral, a visualização pode auxiliar a análise de redes sociais, pois incorpora a percepção humana para a criação de hipóteses sobre os dados. Visualização de redes sociais é uma subárea de visualização de informações. A diferença fundamental é que na visualização de redes sociais o foco está nas pessoas, nos grupos que se formam, seus padrões, suas interações e como os grupos se relacionam com as comunidades (KARAHALIOS; VIÉGAS, 2006).

Conforme já mencionado anteriormente, os dados a serem visualizados são representados por um grafo, com vértices representando entidades sociais e arestas representando os relacionamentos existentes. O princípio fundamental na visualização de redes sociais é facilitar a compreensão dos dados. As técnicas existentes direcionam-se a solucionar subgrupos de informações com o objetivo de simplificar a visualização (FREITAS et al., 2008).

Outro conceito que merece destaque é o de rede social semântica. Segundo Lim et al. (2009), é uma rede multimodal que contém atores representando diferentes tipos de pessoas ou entidades, e as arestas representando as

relações entre eles. Ao contrário das redes sociais tradicionais, nas redes sociais semânticas os atores podem ter diferentes caracterizações a eles associados e tipos de relações semânticas, ou seja, instâncias de cada tipo de ator ou relação podem compartilhar um conjunto comum de atributos. Segundo Singh et al. (2007), propriedades estruturais e atributos descritivos são necessários para uma análise mais completa de redes sociais assim como, para o suporte das tarefas de mineração visual.

3.4. Lei de Benford

A Lei de Benford (Benford 1938), também conhecida como a "Lei do Primeiro Dígito", é uma função de distribuição de probabilidade logarítmica para os primeiros dígitos significativos, e pode ser escrita como

$$P(d) = \log_{10}\left(1 + \frac{1}{d}\right), \quad d = 1, 2, \ldots, 9 \qquad (1)$$

em que *P* é a probabilidade e *d* é o primeiro dígito significativo em questão. O primeiro dígito significativo de um número é o primeiro dígito diferente de zero em sua extrema esquerda, como 7 para 725 e 2 para 0,0239. De acordo com a Equação 1, em um determinado conjunto de dados a probabilidade de ocorrência de um certo dígito como primeiro dígito significativo diminui logaritmicamente quando o valor do dígito aumenta de 1 para 9. As proporções esperadas para os primeiros dígitos são apresentados na Tabela 3.1.

Isso foi observado pela primeira vez em 1881 pelo astrônomo e matemático norte-americano Simon Newcomb (NEWCOMB, 1881), que observou que as primeiras páginas do seu livro de tabelas logarítmicas eram mais desgastadas do que as últimas páginas, o que indicava que as tabelas de logaritmos não foram utilizadas de maneira uniforme. A partir disso ele deduziu que os outros cientistas que utilizavam as tabelas de logaritmos procuraram números começando com 1 com mais frequência do que números começando com 2, números com primeiro dígito 2 com mais frequência do que 3, e assim por diante.

Tabela 3.1 - Proporções esperadas da Lei de Benford para os primeiros dígitos.

Primeiro Dígito *(d)*	*P(d)*
0	-
1	0,3010
2	0,1761
3	0,1249
4	0,0969
5	0,0792
6	0,0669
7	0,0580
8	0,0512
9	0,0458
Total	1

Esta lei foi redescoberta em 1938 por Frank Benford e atualmente é conhecida como “Lei de Benford”. Benford analisou 20 listas de grandes conjuntos de dados, com um total de 20.229 observações e 10 listas de pequenos conjuntos de dados, com um total de 2.968 observações. Estas listas incluíam áreas de superfícies de rios, tamanho de populações, constantes físicas, pesos moleculares, entradas de um manual de matemática, números contidos em uma edição de uma revista, taxas de mortalidade etc. Ele constatou que o dígito 1 tende a ocorrer com uma probabilidade de cerca de 30%, muito maior do que o esperado de 11,1% aproximadamente (isto é, uma distribuição de 1 para 9).

A Lei de Benford é utilizada em diferentes cenários em que há grandes conjuntos naturais ou artificiais numéricos de dados, como na distribuição aderente em termos de países das principais religiões do mundo (MIR, 2012), dados financeiros de comunidades religiosas (CLIPPE; AUSLOOS, 2012), detecção de fraudes em publicações científicas (HEIN et al., 2012), detecção de fraude eleitoral (BEBER; SCACCO, 2012) etc.

Do melhor do nosso conhecimento, o trabalho de Campanario e Coslado (2011) foi a primeira aplicação da Lei de Benford para dados cientométricos. Naquele estudo realizou-se uma amostra do número de artigos publicados,

citações recebidas de periódicos e o fator de impacto de periódicos indexados no *Science Citation Index*® (acessado via WoS) de 1998 a 2007. Para tanto, os autores utilizaram dados publicados na base de dados JCR® disponível na Web, para universidades espanholas. Eles identificaram o primeiro dígito significativo de cada uma dessas variáveis para cada ano separadamente e compararam com o número previsto pela Lei de Benford. Dados de citações seguiram a Lei de Benford em todos os anos estudados. Entretanto, para os dados sobre o número de artigos publicados, não houve conformidade com a Lei de Benford em qualquer dos anos considerados. O mesmo ocorreu com os dados de fator de impacto em quase todos os anos estudados.

3.5. Índices Bibliométricos

Índices ou indicadores bibliométricos são utilizados para analisar a produção científica dos pesquisadores e podem revelar diversas características de uma comunidade científica. Esses índices também são utilizados para avaliar periódicos e instituições. Entretanto, esses índices devem ser utilizados de forma criteriosa para que equívocos não sejam gerados. Essa questão tem gerado discussões na comunidade científica e ainda sem solução consensual.

Nesta Subseção são apresentados alguns índices bibliométricos, destacando principalmente aqueles que são utilizados pelas bases de dados científicas que foram consideradas neste trabalho.

3.5.1. Fator de impacto

O Fator de Impacto (FI) talvez seja o mais conhecido e debatido índice bibliométrico. Foi criado por Eugene Garfield em 1955, o fundador do *Institute for Scientific Information* (ISI), hoje parte da Thomson Reuters.

O FI é calculado anualmente pela Thomson Reuters para todos os periódicos indexados na WoS e publicado no JCR®. Para o cálculo do FI considera-se o número de citações recebidas pelos artigos publicados em um periódico em determinado ano, dividido pelo número de artigos publicados neste mesmo periódico nos dois anos anteriores (GARFIELD, 1955). Entram no cálculo

apenas artigos publicados em periódicos indexados na WoS. Para entender melhor como é calculado, um exemplo é apresentado na Tabela 3.2.

Tabela 3.2 - Exemplo de cálculo do FI de um periódico em 2012.

Ano	Artigos	Citações
2010	96	65
2011	120	87
Total	**216**	**152**
$\mathbf{FI_{2012}} = \frac{\mathbf{65 + 87}}{\mathbf{96 + 120}} = \frac{\mathbf{152}}{\mathbf{216}} = \mathbf{0,704}$		

No Brasil, o JCR® também é utilizado por vários comitês de área da CAPES para avaliar a produção intelectual dos programas de pós-graduação nestas áreas. Também é utilizado na PL nos currículos Lattes dos pesquisadores.

Entretanto, há diversas características que devem ser observadas decorrentes do uso indiscriminado do FI (SEGLEN, 1997). Uma delas é a autocitação (ARNOLD, 2009). Em 2013, o JCR® suspendeu a publicação de índices de mais de 60 periódicos (6 deles brasileiros), sendo que parte desses pelo excesso de citações cruzadas (*citation stacking*). O FI também é influenciado pela área de publicação do periódico, ou seja, áreas com um número maior de pesquisadores tendem a naturalmente receber mais citações. O FI também privilegia áreas que têm artigos com vida média curta de citações. Há também diversos outros problemas que surgem com o seu uso, tais como: apenas poucos artigos de um periódico é que são muito citados; artigos podem ser citados sem terem sido lidos; um artigo pode ser muito citado devido a um erro nos métodos empregados ou na interpretação dos resultados experimentais; artigos de revisão recebem grande quantidade de citações (ALMEIDA; GUIMARÃES, 2013), sabendo disso, alguns editores podem tender a privilegiar este tipo de artigo em seus periódicos etc.

Recentemente, os cientistas da *American Society for Cell Biology* promoveram uma iniciativa denominada "*San Francisco Declaration on Research*

Assessment" (DORA) com o objetivo de se parar a utilização do FI para avaliação da pesquisa científica (DORA, 2013). A declaração recomenda que o FI não deve ser utilizado em avaliações relativas a financiamento, promoções na carreira e contratações de pesquisadores. O documento foi assinado por mais de 150 cientistas proeminentes e 75 organizações acadêmicas. Segundo o documento, o mérito de um trabalho (e do cientista que o realizou) deve ser avaliado com base na qualidade do próprio trabalho, e não do periódico no qual ele foi publicado.

3.5.2. Índice H

O índice H pode ser utilizado para medir o impacto de um determinado pesquisador individualmente. Este índice é baseado na distribuição das citações recebidas pelas publicações de um pesquisador. Foi proposto em 2005 pelo físico argentino Jorge Eduardo Hirsch da Universidade da Califórnia, em San Diego (Estados Unidos). Ele é dado pelo número de artigos com citações maiores ou iguais a esse número (HIRSCH, 2005). Assim, um pesquisador com índice H igual a 10, tem pelo menos 10 artigos publicados que receberam, cada um deles, pelo menos 10 citações. Segundo Hirsch (2005), é um índice muito simples de ser calculado e integra a produtividade científica e o impacto das publicações.

O índice H é influenciado também pelas autocitações e não leva em conta o número de coautores. O próprio Hirsch afirma que o índice não deve ser utilizado para se comparar pesquisadores de áreas diferentes (HIRSCH, 2005). Ele também afirma que o índice apresenta limitações técnicas, tais como a dificuldade em se obter o número total de artigos e citações de pesquisadores com nomes comuns. Pesquisadores com tempo de carreira diferentes também não devem ser comparados (ALONSO et al., 2009). Outra limitação é que artigos muito citados são importantes para o cálculo do índice H, mas o número de citações que um determinado artigo recebe que superam o índice H não é mais importante (COSTAS; BORDONS, 2007).

Há várias outras variações do índice H (BORNMANN et al., 2008; ALONSO et al., 2009). Um deles é o índice g, que é definido como o maior número *g* de artigos que juntos receberam g^2 ou mais citações (EGGHE, 2006). Em contraste ao índice H, o índice g dá mais peso aos artigos altamente citados, o que é considerado uma limitação do índice H. Assim como o índice g, o índice h(2) também dá mais peso aos artigos altamente citados. O índice h(2) é definido como o maior número tal que os h(2) artigos mais citados receberam cada um pelo menos $h(2)^2$ citações (KOSMULSKI, 2006). Por exemplo, um índice h(2) de 20 significa que um pesquisador publicou pelo menos 20 artigos e que cada um deles foi citado pelo menos 400 vezes.

As bases de dados WoS e Scopus permitem verificar o índice H dos autores citados nos documentos publicados em seus periódicos indexados. O índice também é publicado no SCImago.

3.5.3. Outros índices

Além do FI, o JCR® também publica outros índices, tais como o índice de citação imediata (ou de imediatez) e o índice meia-vida das citações. O índice de citação imediata corresponde ao número de vezes que um artigo de um periódico específico é citado pelos periódicos indexados na WoS durante o ano de sua publicação. Ele indica a rapidez com que a ideia de um trabalho se dissemina na comunidade. O índice meia-vida das citações de um periódico corresponde ao tempo (em anos) necessário para que metade das citações recebidas por um periódico apareçam na literatura científica.

Há outros índices utilizados para a avaliação de um pesquisador. Os dois mais conhecidos na comunidade científica são o número total de artigos publicados e o número total de citações recebidas. O primeiro representa a produtividade do pesquisador e o segundo, o impacto de suas publicações. Também foram propostos índices que agrupam a produtividade científica e o impacto das publicações. Por exemplo, o número de citações por artigo e o número de publicações significativas, definido como o número de artigos com mais de X

citações. Esses índices podem considerar toda a vida acadêmica de um pesquisador ou apenas parte dela (BATISTA, 2010).

A Scopus também disponibiliza alguns índices, entre eles o SJR e o SNIP (*Source Normalized Impact per Paper*). O índice SJR também é divulgado no SCImago. O SJR é uma medida do "prestígio" de um periódico que considera tanto o número de citações recebidas por um periódico quanto a importância dos periódicos de onde tais citações vêm (GONZÁLEZ-PEREIRA et al., 2010). Com o SJR, a área, a qualidade e a reputação de um periódico têm um impacto direto sobre o valor de uma citação. O SJR é obtido através de um processo iterativo e sua determinação não é imediata como no FI, mas se apresenta como uma alternativa ao FI. Já o SNIP mede o impacto da citação pesando as citações com base no número total de citações em uma determinada área do conhecimento (MOED, 2010).

3.6. Considerações finais

Neste Capítulo foi apresentado um resumo sobre os tópicos que são relevantes para este trabalho. Além de contribuírem para um melhor entendimento do trabalho realizado, esses foram fundamentais no desenvolvimento das ferramentas computacionais e contribuíram para a realização dos estudos de casos. O levantamento bibliográfico desses tópicos, embora não exaustivo, faz a cobertura temática necessária para realizar este trabalho. No próximo Capítulo é apresentada uma descrição das ferramentas computacionais desenvolvidas neste trabalho.

4 FERRAMENTAS COMPUTACIONAIS

Neste Capítulo é apresentada uma visão geral das ferramentas computacionais desenvolvidas ao longo deste trabalho, destacando suas principais características e funcionalidades. Todas essas ferramentas têm como principal objetivo extrair informações de alguma das bases de dados científicas consideradas neste trabalho. Dessa forma, essas ferramentas foram desenvolvidas a partir da necessidade de realizar análises mais abrangentes.

As duas principais ferramentas foram desenvolvidas como linguagens voltadas para domínios específicos. A ideia foi facilitar a utilização por outros usuários desenvolvendo linguagens com alto nível de abstração e poder de expressão, incorporando o conhecimento do domínio. Também é apresentado o sistema SUCUPIRA, destacando a sua arquitetura e as suas funcionalidades. Algumas ferramentas foram desenvolvidas como extratores, permitindo acessar as bases de dados e obter automaticamente as informações necessárias. Algumas bases de dados permitem exportar os seus dados para planilhas. Por isso, houve a necessidade de criar um mecanismo que permita converter planilhas em um banco de dados. Esse mecanismo foi utilizado para duas bases de dados.

4.1. Linguagens de Domínio Específico

Foram desenvolvidas duas linguagens: LattesMiner e ScopusMiner. A primeira é voltada para a PL e a segunda para a base de dados Scopus. A seguir são apresentadas essas linguagens, destacando detalhes de implementação e ilustrando com um exemplo de uso.

4.1.1. LattesMiner

A PL é hoje, sem dúvida, a principal fonte de informações sobre os pesquisadores brasileiros e tem elevado potencial para extração de informação. Entretanto, não existem mecanismos que permitam que isso seja feito de maneira simples e rápida, e sem o auxílio de desenvolvedores experientes. Dessa forma, há a necessidade de encontrar mecanismos que

permitam realizar essa tarefa com um maior nível de abstração, por um número maior de usuários e também de forma mais eficiente.

Nesse sentido, foi desenvolvida a linguagem **LattesMiner**, que é uma LDE interna e multilíngue para extração automática de informações de currículos Lattes. É composta por um conjunto de classes escritas em Java que permite que outros desenvolvedores implementem suas próprias aplicações com alto nível de abstração e poder de expressão.

A linguagem LattesMiner permite extrair informações de um pesquisador individual ou de um grupo de pesquisadores (até todo o conjunto deles) utilizando o nome ou ID do pesquisador. As informações extraídas permitem identificar redes sociais acadêmicas, competências regionais, perfil de grupos de diferentes áreas de pesquisa etc.

4.1.1.1. Domínio do problema

A primeira tarefa no projeto de uma LDE é definir os termos do problema (DEURSEN et al., 2000). Vale a pena mencionar que, embora o currículo Lattes esteja disponível em Português, também é possível disponibilizá-lo em Inglês. Além disso, o sistema Currículo Lattes já está sendo utilizado em outros idiomas e países, como Argentina, Chile, Colômbia, Cuba, Equador, México, Panamá, Paraguai, Peru, Portugal e Venezuela. O Brasil e esses países são membros da rede ScienTI, que é uma rede pública de fontes de informação e conhecimento com o objetivo de contribuir à gestão da atividade científica, tecnológica e de inovação desses países (SCIENTI, 2013).

Assim, quando a linguagem LattesMiner foi projetada estes fatos foram considerados. A definição dos termos do problema é muito importante, pois os mesmos são considerados e utilizados para projetar a LDE, que deve descrever concisamente aplicações de um domínio particular (nesse caso, do currículo Lattes), permitindo uma solução no idioma e no nível de abstração do domínio do problema.

Para definir os termos do problema foram verificados os termos utilizados no menu de opções de acesso rápido em currículos Lattes de pesquisadores. O menu de opções de um currículo Lattes é definido de acordo com os dados informados pelo pesquisador. A Figura 4.1 ilustra um exemplo de menu de opções de um currículo Lattes em português. Também foram verificados os termos utilizados em menus de opções de currículos Lattes em inglês.

Figura 4.1 - Menu de opções de acesso rápido de um currículo Lattes.

4.1.1.2. Componentes

A linguagem LattesMiner é composta por seis componentes: Descoberta de Dados, Aquisição de Dados, Extração de Dados, Estruturação de Dados, Visualização de Dados e Análise de Dados. A saída de um componente é utilizada como entrada para outro componente. A Figura 4.2 ilustra os componentes da linguagem LattesMiner.

Figura 4.2 - Componentes da linguagem LattesMiner.

O componente “Descoberta de Dados” é opcional, ou seja, é necessário somente se o ID dos pesquisadores não estiver disponível. O componente “Aquisição de Dados” também é opcional, uma vez que o currículo Lattes de um pesquisador pode ser baixado diretamente do sítio do CNPq, sendo necessário apenas que o currículo seja armazenado como arquivo HTML. Uma visão geral da arquitetura de componentes da linguagem LattesMiner é ilustrada na Figura 4.3.

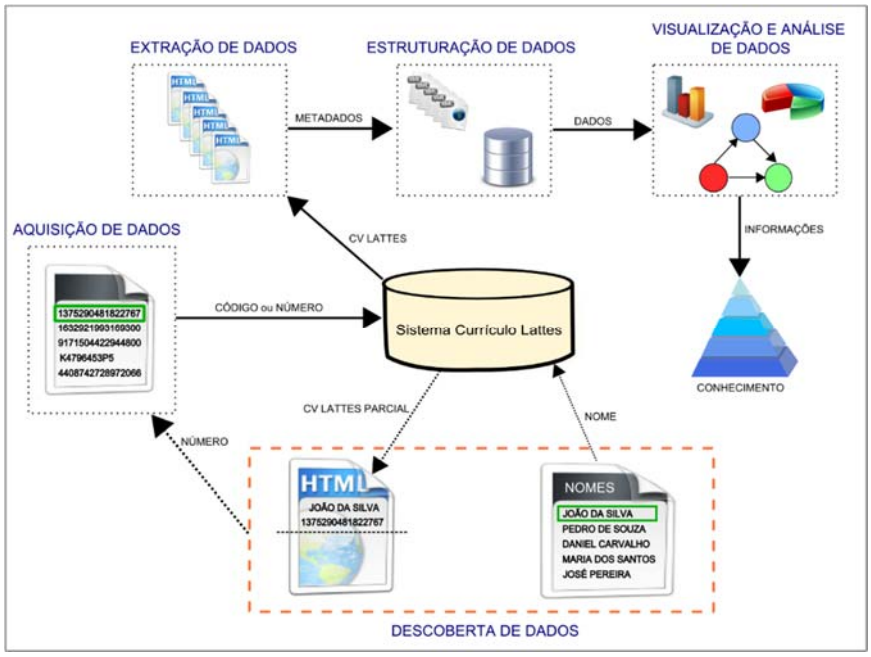


Figura 4.3 - Arquitetura de componentes da linguagem LattesMiner.

Os componentes “Descoberta de Dados” e “Aquisição de Dados” acessam o sistema Currículo Lattes através do código ou ID do pesquisador. Para o componente “Descoberta de Dados” é retornado apenas a parte inicial do currículo Lattes para verificar se o nome contido no currículo é igual ao nome procurado. Para o componente “Aquisição de Dados” é retornada uma cópia do currículo Lattes que é armazenado como arquivo HTML.

Os componentes “Visualização de Dados” e “Análise de Dados” dependem do componente “Estruturação de Dados” que armazena os dados extraídos em XML ou em um banco de dados. Isso é necessário para que o desempenho não seja tão comprometido, uma vez que os arquivos XML são bem menores que os arquivos HTML, pois guardam apenas os dados de interesse e o acesso a um banco de dados é ainda mais eficiente pois não há a necessidade de carregar para a memória o arquivo HTML armazenado em disco. Esses dois componentes extraem informações que permitem a descoberta de conhecimento. A seguir todos os componentes da linguagem LattesMiner são detalhados.

Descoberta de Dados

O componente “Descoberta de Dados” é utilizado para descobrir o número (ID) dos pesquisadores. Cada currículo Lattes tem uma URL que permite acessá-lo diretamente. Usualmente, apenas o nome do pesquisador está disponível e o sistema Currículo Lattes não permite realizar a busca automática por nome. A URL é composta por um número de 16 dígitos (por exemplo, http://lattes.cnpq.br/6187221670775160). Com esse número (ID), é possível acessar um determinado currículo automaticamente quantas vezes forem necessárias e, portanto, pode ser acessada por mecanismos de busca automáticos.

Outra forma de acessar um currículo Lattes é utilizando um outro identificador do pesquisador (código) que é composto por letras e números (por exemplo, http://buscatextual.cnpq.br/buscatextual/visualizacv.do?metodo=apresentar&id=

K4787637J9). A linguagem LattesMiner permite acessar um currículo Lattes utilizando qualquer uma das formas.

Uma questão importante a ser respondida é como obter esses identificadores. Normalmente, a única informação disponível é o nome do pesquisador. Isso configura outros problemas, pois em uma lista simples de nomes pode conter homônimos, os nomes podem ser informados incorretamente ou parcialmente, e até mesmo uma mudança no nome do pesquisador, como por exemplo, quando uma pessoa se casa e o sobrenome do cônjuge é acrescentado.

Aquisição de Dados

O resultado do componente “Descoberta de Dados” pode ser utilizado como entrada para o componente “Aquisição de Dados”. Este componente é responsável por baixar os currículos Lattes dos pesquisadores a partir do sítio do CNPq. Na implementação da linguagem LattesMiner optou-se por baixar os currículos Lattes como arquivos HTML por estarem acessíveis na Web, ao contrário dos arquivos XML que têm acesso restrito. Atualmente, para baixar um currículo Lattes como arquivo XML é necessário informar um código alfanumérico cuja finalidade é evitar que currículos sejam baixados automaticamente por *scripts*. Este componente também permite baixar um currículo Lattes utilizando o identificador do pesquisador de 16 dígitos ou o identificador composto por letras e números (código). Esse identificador é utilizado como nome do arquivo HTML que armazena o currículo Lattes do pesquisador.

Extração de Dados

O componente “Extração de Dados” é o principal componente da linguagem LattesMiner. Este componente é responsável pela extração de informações dos arquivos HTML que armazenam os currículos Lattes dos pesquisadores. Atualmente, os dados que são extraídos estão indicados na Tabela 4.1.

Tabela 4.1 - Dados extraídos pela linguagem LattesMiner.

Dados Pessoais	Código, número (ID), nome, categoria de bolsista PQ (se for o caso), data da última atualização, data da morte (se for o caso), data e hora do arquivo, e resumo
Endereço Profissional	Instituição, cidade, estado, país, CEP e homepage
Formação Acadêmica	Nível, orientador, ID do orientador (se for o caso), instituição, título, ano de início, ano de conclusão, ano de obtenção, agência financiadora da bolsa, área, curso, código do curso na CAPES, conceito CAPES e palavras-chave
Formação Complementar	Curso, instituição, carga horária, ano de início e ano de conclusão
Áreas de Atuação	Grande área, área, subárea e especialidade
Linhas de Pesquisa	Título, objetivo e palavras-chave
Projetos de Pesquisa	Título, descrição, ano de início, ano de conclusão e agência financiadora
Artigos completos publicados em periódicos	Autores, título, periódico, volume, série, páginas, DOI, ISSN, ano e se é um dos 5 trabalhos mais relevantes
Trabalhos completos publicados em anais de congressos	Autores, título, evento, páginas, ano e se é um dos 5 trabalhos mais relevantes
Resumos expandidos	Autores, título, evento, páginas, ano e se é um dos 5 trabalhos mais relevantes
Resumos publicados em anais de congressos	Autores, título, evento, páginas, ano e se é um dos 5 trabalhos mais relevantes
Livros publicados	Autores, título, ISBN, ano e se é um dos 5 trabalhos mais relevantes
Capítulos de livros publicados	Autores, título, ISBN, ano e se é um dos 5 trabalhos mais relevantes
Programas de computador sem registro	Autores, título, ano e se é um dos 5 trabalhos mais relevantes
Produtos Tecnológicos	Autores, título, ano e se é um dos 5 trabalhos mais relevantes
Processos ou Técnicas	Autores, título, ano e se é um dos 5 trabalhos mais relevantes
Outras produções bibliográficas	Autores, título, natureza, ano e se é um dos 5 trabalhos mais relevantes
Trabalhos técnicos	Autores, título, ano e se é um dos 5 trabalhos mais relevantes
Demais tipos de produção técnica	Autores, título, ano e se é um dos 5 trabalhos mais relevantes
Patentes	Inventores, título, data de depósito, instituições financiadoras, país, número do registro e ano
Participação em bancas	Tipo, aluno, título, instituição, área, curso, palavras-chave e ano
Orientações concluídas	Tipo, nível, aluno, título, instituição, área, curso, palavras-chave e ano.
Membro de Corpo Editorial	Periódico, ano de início e ano de término
Revisor de Periódico	Periódico, ano de início e ano de término
Participação em eventos	Tipo, título e ano
Organização de eventos	Tipo, título, autores e ano
Idiomas	Idioma, leitura, escrita, fala e compreensão
Prêmios e títulos	Ano e título
Citações	Todas as formas de citação de um pesquisador
Contatos	Links para outros currículos Lattes

Estruturação de Dados

Os dados extraídos podem ser armazenados em arquivos no formato XML ou em um banco de dados qualquer utilizando o componente “Estruturação de Dados”. No caso do banco de dados, qualquer um pode ser utilizado, uma vez que a linguagem LattesMiner possui um arquivo de propriedades que permite tal configuração, podendo ser alterado facilmente a qualquer instante.

Visualização de Dados

O componente “Visualização de Dados” é responsável pela identificação e visualização de redes sociais acadêmicas. A identificação dessas redes sociais é feita verificando os relacionamentos entre os pesquisadores obtidos a partir dos currículos Lattes. E como essa identificação considera apenas as informações acadêmicas dos pesquisadores, essas redes são chamadas de redes sociais acadêmicas.

Análise de Dados

O componente “Análise de Dados” é responsável pela análise dos dados extraídos e também pela análise dos relacionamentos identificados. No momento, a linguagem LattesMiner permite apenas análises simples das relações identificadas, como a identificação de cliques e da clique máxima. Este componente também permitirá a análise de dados utilizando técnicas de estatística descritiva.

4.1.1.3. Implementação

A linguagem LattesMiner é composta por um conjunto de classes escritas em Java e sua classe principal fornece a maioria das funcionalidades da LDE. A Figura 4.4 ilustra um diagrama de classes UML que descreve parte da linguagem LattesMiner.

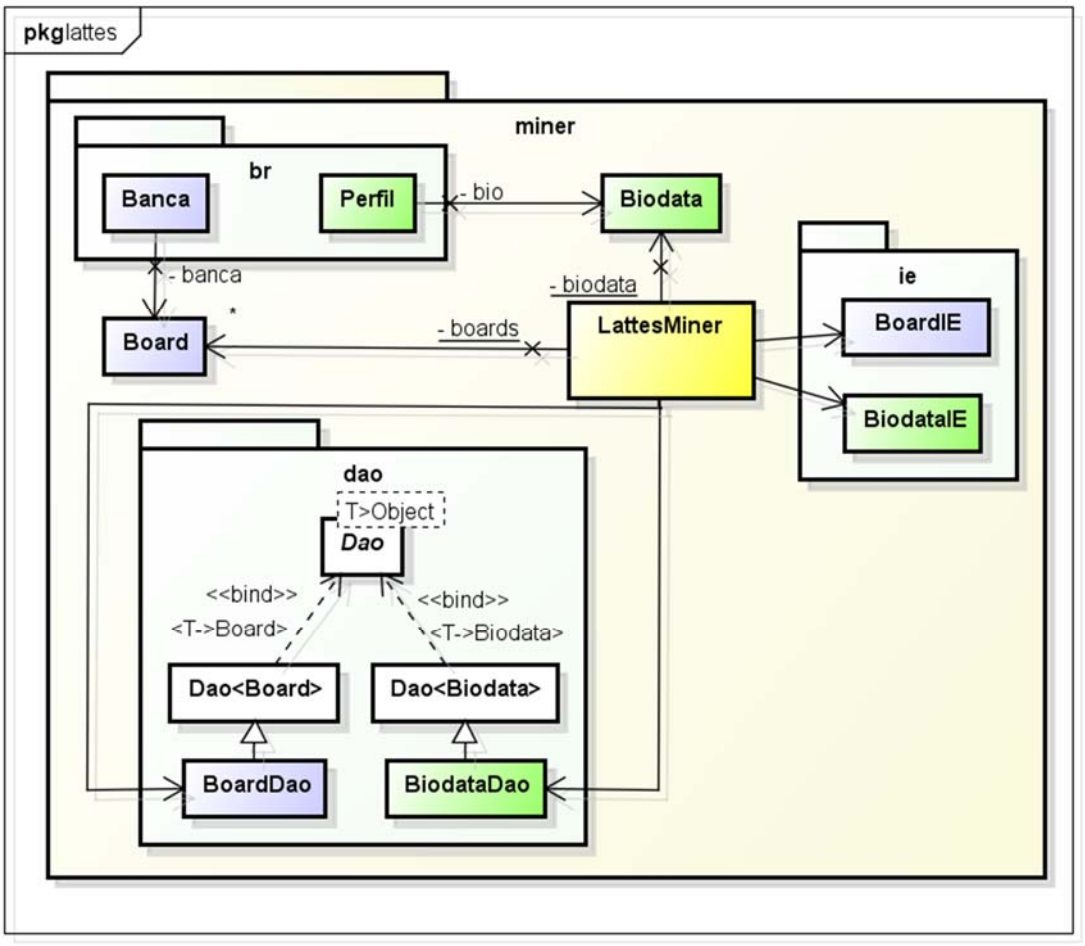


Figura 4.4 - Diagrama de Classes UML parcial da linguagem LattesMiner.

A linguagem LattesMiner é multilíngue e está disponível em Português e Inglês, permitindo utilizar os métodos da linguagem no idioma desejado. Originalmente, as classes e métodos da linguagem foram escritas em inglês. As classes e métodos em português foram criados a partir dos originais. Para adicionar um novo idioma, é necessário apenas criar uma nova classe Java e instanciar um objeto dessa classe no idioma original, permitindo que os novos métodos acessem os métodos da classe original. O mecanismo de herança não foi utilizado porque quando uma nova classe fosse criada em outro idioma qualquer, ela também herdaria os métodos da classe original. Assim, quando o usuário fosse utilizar as classes do novo idioma, ele também teria acesso aos métodos da classe original. Na linguagem LattesMiner esse problema não

ocorre e o usuário apenas tem acesso aos métodos das classes no idioma que estiver utilizando no momento.

A classe **LattesMiner** é composta por instâncias das classes **Biodata** e **Board**, além de outras classes aqui não apresentadas. A classe **Biodata**, por exemplo, contém os dados do perfil de um pesquisador e a sua classe correspondente em português é a classe **Perfil**, que é associada à classe **Biodata**, ou seja, uma instância da classe **Perfil** tem uma associação com uma instância da classe **Biodata**. A classe **BiodataIE** é responsável pela extração de informações de um currículo Lattes e a classe **BiodataDao** é responsável pela persistência dessas informações em um banco de dados.

A linguagem LattesMiner foi implementada utilizando uma interface fluente, que fornece uma representação compacta e fácil de ler do domínio do problema. Interfaces fluentes foram implementadas utilizando o método de encadeamento (*method chaining*). É importante lembrar que o método de encadeamento por si só não é suficiente para criar uma LDE. Por exemplo, a classe “StringBuilder” da linguagem Java tem um método “append()” que sempre retorna uma instância da própria classe. Porém, ela não resolve o problema de um domínio específico e, portanto, não é uma LDE.

Também é interessante notar que utilizando o método de encadeamento, qualquer método da linguagem LattesMiner pode ser utilizado em qualquer ordem e várias vezes. Além do método de encadeamento, a linguagem LattesMiner também faz uso de métodos estáticos que permitem criar códigos mais compactos e ainda sim legíveis.

Uma questão que mereceu uma atenção especial na implementação da linguagem LattesMiner foi a extração de informações em currículos Lattes. Inicialmente, foi constatado que o currículo Lattes baixado como arquivo HTML não era balanceado e portanto, não era possível utilizar um *parser*. Porém, foi observado que trechos de código no arquivo HTML do currículo Lattes têm uma estrutura de repetição, ou seja, têm a mesma formatação HTML (NANNO et al.,

2003). Por essas razões a técnica de extração de informações baseado em expressões regulares foi utilizada.

Conforme já mencionado anteriormente, o currículo Lattes atualmente é disponibilizado como uma página HTML. Isso torna a linguagem LattesMiner dependente desse formato, pois se o CNPq fizer alguma modificação na geração dessa página, a linguagem LattesMiner pode não conseguir extrair as informações. Contudo, esse é um problema que é independente do formato do currículo Lattes, pois qualquer que seja o formato o mesmo pode ser modificado a qualquer momento. Para amenizar esse problema, a linguagem LattesMiner possui um módulo para extração de informações, em que cada classe é responsável por um tópico que é extraído do currículo Lattes. Além disso, as expressões regulares utilizadas em cada uma das classes são armazenadas separadamente em um arquivo de propriedades que pode ser modificado sem a necessidade de alterar o código fonte da linguagem.

A linguagem LattesMiner também possui um módulo para persistência dos dados, tendo uma classe para cada tópico extraído. A Figura 4.5 ilustra o diagrama com as tabelas geradas pela linguagem LattesMiner. Tanto o banco de dados quanto cada uma das tabelas são criados automaticamente. Para isso foi necessário definir um arquivo de propriedades com o código de criação do banco de dados e de todas as suas tabelas. Então, caso a classe responsável pela persistência dos dados de um determinado tópico não encontre a tabela correspondente para armazenar os dados, essa tabela será criada automaticamente e os dados armazenados em seguida.

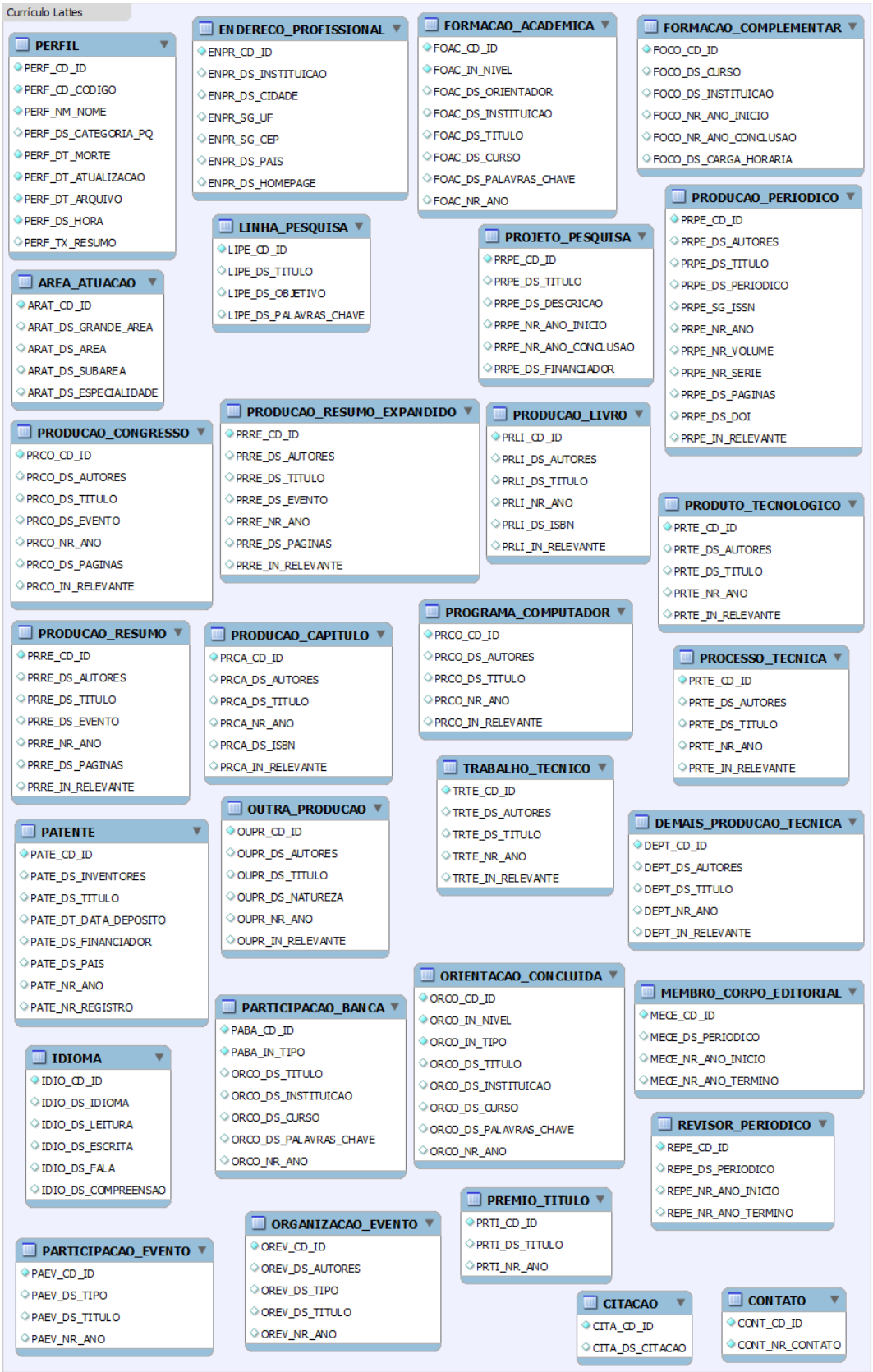


Figura 4.5 - Diagrama das tabelas que armazenam os dados extraídos utilizando a linguagem LattesMiner.

4.1.1.4. Comparação

Nesta seção é apresentada uma comparação entre as três ferramentas existentes atualmente para a extração de informações de currículos Lattes. Lattes Extrator tem a vantagem de extrair os currículos diretamente do banco de dados da PL e são extraídos como arquivos XML. As ferramentas scriptLattes e LattesMiner extraem os currículos a partir do sistema Currículo Lattes acessível na Web. Nessas ferramentas, os currículos são obtidos como páginas HTML, o que dificulta a extração de informações.

Uma vantagem da linguagem LattesMiner é o fato de ser uma LDE que permite aos desenvolvedores programarem suas próprias aplicações com alto nível de abstração e poder de expressão. Lattes Extrator e scriptLattes realizam busca apenas pelo número (ID) do pesquisador, enquanto a linguagem LattesMiner também permite a busca pelo nome do pesquisador. Dessa forma, é possível, por exemplo, buscar por qualquer nome citado em um currículo Lattes, aumentando assim o número de relacionamentos identificáveis.

Outra vantagem da linguagem LattesMiner é o fato de permitir extrair informações de grupos com muitos pesquisadores. Isso é um problema para a ferramenta scriptLattes, uma vez que é necessário criar um arquivo texto contendo o número (ID) e o nome dos pesquisadores, além de outras informações opcionais que auxiliam no processo de extração. Por exemplo, se for desejado analisar um grupo de 100 mil pesquisadores, é necessário obter manualmente o número (ID) de todos esses pesquisadores. É importante lembrar que para obter o número (ID) de um pesquisador, o nome completo do mesmo deve ser pesquisado na PL. Se o nome for informado parcialmente, é retornada uma lista de possíveis nomes e o usuário deve procurar pelo nome correto. Em seguida, o currículo Lattes do pesquisador deve ser acessado e o número (ID) e o nome do pesquisador devem ser copiados para o arquivo texto de configuração do scriptLattes. Hipoteticamente, consideremos que sejam gastos 20 segundos para se obter o número (ID) de cada pesquisador. Assim, serão necessários 23 dias ininterruptos ou 69 dias considerando uma carga

horária de 8 horas diárias para se montar o arquivo texto de configuração, o que é desencorajador.

Outra dificuldade do scriptLattes é o tempo gasto para obter um grafo de colaborações, uma vez que os dados são extraídos diretamente de arquivos HTML. Na linguagem LattesMiner, esta dificuldade é menor pois os dados podem ser armazenados em um banco de dados, o que torna o processo de obter os grafos muito mais rápido. As informações extraídas também podem ser importadas por outras ferramentas, uma vez que essas informações podem ser armazenadas em arquivos XML. No caso do scriptLattes, as informações extraídas são apresentadas em relatórios gerados como páginas HTML, o que impede o seu uso por outras aplicações.

A principal vantagem da linguagem LattesMiner é o fato da interface da linguagem não mudar, ou seja, mesmo que o processo de extração seja todo ele modificado, a interface continua a mesma. Por exemplo, se o currículo Lattes passar a ser disponibilizado em XML ao invés de HTML, isso é indiferente para a linguagem LattesMiner. Internamente, o processo de extração precisa ser modificado. Porém, para o usuário a interface continua a mesma. O mesmo não pode ser afirmar do scriptLattes, que é dependente do processo de extração.

A Tabela 4.2 apresenta uma comparação entre as ferramentas de extração, destacando as suas principais características.

Tabela 4.2 - Quadro comparativo entre as ferramentas de extração de informações de currículos Lattes.

Tópicos	Lattes Extrator	scriptLattes	LattesMiner
Linguagem de desenvolvimento	JSP	Python/JSP	Java
Local de desenvolvimento	CNPq	IME-USP	INPE/ITA
Formato dos currículos extraídos	XML	HTML	HTML
Restrição de sistema operacional	-	Linux	-
Busca pelo nome do pesquisador	-	-	Sim
Multilíngue	-	-	Sim
Biblioteca programável	-	-	Sim
Exportação de dados para XML	Sim	-	Sim
Exportação para banco de dados	-	-	Sim
Relatórios e gráficos	-	Sim	Não
Identificação de redes sociais	-	Sim	Sim
Visualização de redes sociais	-	Sim	Sim
Análise de redes sociais	-	-	Sim

4.1.1.5. Exemplo de uso

Nesta Subseção é apresentado um exemplo de uso da linguagem LattesMiner, mostrando passo a passo como utilizar as principais funcionalidades da linguagem.

O primeiro passo é criar um arquivo texto contendo o nome dos pesquisadores. Para esse exemplo, foram utilizados os nomes dos 5 pesquisadores que receberam o Prêmio Anísio Teixeira em 2011. Então foi criado o arquivo “nomes.txt” contendo cada nome em uma linha separada, conforme mostra a Listagem 4.1.

nomes.txt

```
Nelson Maculan Filho
Luiz Bevilacqua
Fernando Galembeck
Alvaro Toubes Prata
João Fernando Gomes de Oliveira
```

Listagem 4.1 - Exemplo de arquivo texto contendo o nome de pesquisadores.

O próximo passo é obter o número (ID) dos pesquisadores. A Listagem 4.2 mostra o código-fonte de uma aplicação Java para descobrir o número (ID) dos pesquisadores utilizando a linguagem LattesMiner.

ExemploLattes01.java

```
import java.util.*;
import lattes.util.Util;
import static lattes.miner.LattesMiner.*;

public class ExemploLattes01
{
 public static void main(String[] args)
 {
  List<String> list = new ArrayList<String>();

  for (String nome : Util.getList("nomes.txt"))
   list.add(search(nome));

  Util.setList(list, "ids.txt");
 }
}
```

Listagem 4.2 - Exemplo de uma aplicação Java para identificação do número (ID) de pesquisadores utilizando a linguagem LattesMiner.

O método "**search()**" realiza a busca pelo nome do pesquisador no sistema Currículo Lattes. Se for encontrado, é retornado o número (ID) do pesquisador. Caso contrário, é retornado o nome do pesquisador. Nos casos em que mais de um currículo Lattes com o mesmo nome é encontrado, são retornados todos os números (ID) concatenados e separados por vírgula. Assim, é possível verificar se algum problema ocorreu. Nesse caso, o resultado foi armazenado

em um arquivo texto denominado “ids.txt”. Todos os números (ID) dos pesquisadores foram encontrados, conforme mostra a Listagem 4.3.

ids.txt

```
K4783153E3
K4787137U2
K4787937A7
K4781599Z8
K4787011P6
```

Listagem 4.3 - Exemplo de arquivo texto contendo o número (ID) de pesquisadores.

Em seguida, a lista de números (ID) dos pesquisadores gerada anteriormente é lida e o currículo Lattes correspondente é baixado. O código-fonte para realizar essa tarefa é mostrado na Listagem 4.4.

ExemploLattes02.java

```
import lattes.util.Util;
import static lattes.miner.LattesMiner.*;

public class ExemploLattes02
{
 public static void main(String[] args)
 {
  dir("cvs");
  for (String id : Util.getList("ids.txt"))
   download(id).save();
 }
}
```

Listagem 4.4 - Exemplo de uma aplicação Java para baixar currículos Lattes de pesquisadores utilizando a linguagem LattesMiner.

Para baixar o currículo Lattes de um pesquisador é usado o método “**download()**”. O método “**save()**” armazena o currículo Lattes baixado como arquivo HTML e o número (ID) do pesquisador é usado como nome de arquivo. O método “**dir()**” é opcional e permite definir um diretório no qual o currículo baixado será armazenado. Se o diretório não existir, ele é criado automaticamente.

Após executar esses passos, é possível extrair as informações dos currículos Lattes baixados. A lista de números (ID) dos pesquisadores é lida novamente, conforme mostra o código-fonte da Listagem 4.5.

ExemploLattes03.java

```
import lattes.util.Util;
import static lattes.miner.LattesMiner.*;

public class Exemplo03
{
 public static void main(String[] args)
 {
  props("mysql");
  for (String id : Util.getList("ids.txt"))
  {
   load(id).address().boards().save();
  }
 }
}
```

Listagem 4.5 - Exemplo de uma aplicação Java para extrair informações de currículos Lattes de pesquisadores utilizando a linguagem LattesMiner.

O método "**load()**" é utilizado para carregar o arquivo HTML do currículo Lattes do pesquisador na memória como uma *string*. Dessa forma, é possível utilizar qualquer um dos métodos disponíveis na linguagem LattesMiner para extração de informações. É importante destacar que a ordem dos métodos é indiferente, pois cada um deles retorna uma instância da própria classe principal (**LattesMiner**), permitindo o encadeamento de métodos.

O método "**address()**" permite extrair o endereço profissional informado no currículo Lattes de um pesquisador. O método "**boards()**" permite extrair todas as participações em bancas de um pesquisador, tanto em nível de mestrado quanto em nível de doutorado.

O método "**save()**" nesse caso tem uma funcionalidade diferente. Este método armazena as informações extraídas, de acordo com os métodos de extração utilizados, em um banco de dados definido em um arquivo de propriedades (nesse caso, "mysql.properties"). O arquivo de propriedades é definido através do método "**props()**". Outra possibilidade seria armazenar as informações

extraídas em arquivos XML. Nesse caso, o método "**xml()**" deveria ter sido utilizado ao invés do método "**save()**".

A Tabela 4.3 apresenta todos os métodos disponíveis na Linguagem LattesMiner que permitem extrair informações de um currículo Lattes. Esses métodos são apresentados em Português e Inglês.

Tabela 4.3 - Métodos da linguagem LattesMiner para extração de informações.

Português	Inglês	Português	Inglês
perfil()	biodata()	endereco()	address()
areas()	areas()	formacoes()	formations()
idiomas()	languages()	contatos()	contacts()
bancas()	boards()	orientacoes()	advisories()
publicacoes(PERIODICO)	publications(JOURNAL)	publicacoes(CONGRESSO)	publications(CONFERENCE)
publicacoes(LIVRO)	publications(BOOK)	publicacoes(CAPITULO)	publications(CHAPTER)
publicacoes(RESUMO)	publications(RESUME)	publicacoes(EXPANDIDO)	publications(EXPANDED)
publicacoes(OUTRA)	publications(OTHER)	tecnicas()	techniques()
patentes()	patents()	processos()	processes()
programas()	programs()	produtos()	products()
pesquisas()	researches()	projetos()	projects()
trabalhos()	works()	cursos()	courses()
premios()	awards()	revisores()	referees()
citacoes()	citations()	editorial()	editorial()
eventos(PARTICIPACAO)	events(PARTICIPATION)	eventos(ORGANIZACAO)	events(ORGANIZATION)

Depois que as informações extraídas estão armazenadas em um banco de dados, outras consultas podem ser feitas e informações diferentes podem ser obtidas. Apesar de ser possível obter essas informações diretamente dos currículos Lattes armazenados como páginas HTML, isso não é viável principalmente quando o grupo que está sendo analisado contém muitos pesquisadores.

Nos trechos de código a seguir as informações são obtidas diretamente do banco de dados criado no exemplo anterior. Para isso, o método "**database()**" deve ser utilizado. Há também um conjunto de métodos que foram definidos para realizar consultas SQL no banco de dados. O trecho de código a seguir

mostra como obter as primeiras informações sobre o grupo que está sendo analisado.

```
database();
total().println();
update().year(2012).println();
```

Para saber quantos pesquisadores estão sendo analisados, bastar utilizar o método “**total()**”. O resultado pode ser impresso na tela utilizando o método “**print()**” ou “**println()**”. Outra opção é retornar o resultado utilizando o método “**result()**”. O método “**update()**” permite verificar quantos pesquisadores atualizaram o currículo Lattes no ano especificado no método “**year()**”. Se não for especificado nenhum ano, o método “**update()**” considera o ano atual. A saída da execução desse trecho de código é apresentada a seguir.

SAÍDA

```
5
4
```

No estudo em questão, há 5 pesquisadores sendo que 4 deles atualizaram o currículo Lattes em 2012. Outra informação que pode ser obtida é quantos pesquisadores possuem bolsa PQ do CNPq. Isso pode ser feito utilizando o método “**scholarship()**”, ilustrado no trecho de código a seguir.

```
scholarship().println();
scholarship(PQ_1A).println();
```

Nesse trecho de código a primeira informação obtida é quantos pesquisadores possuem bolsa PQ. Também, é obtido, quantos pesquisadores são bolsistas da categoria 1A. Neste caso, todos os 5 pesquisadores são bolsistas e da categoria 1A, conforme ilustra a saída apresentada a seguir.

SAÍDA

```
5
5
```

O método “**institution()**” permite obter as instituições em que os pesquisadores declaram trabalhar. Se o método “**top(int n)**” for utilizado, as *n* instituições mais informadas podem ser obtidas. O método **“get(int i)**” permite obter informações sobre uma determinada instituição da lista ranqueada. O trecho de código a seguir ilustra como isso pode ser feito.

```
institution().top(3).println();
institution().get(1).println();
institution().get(1).total().println();
```

Nesse trecho de código, o método “**println()**” imprimiu na tela as 3 instituições mais informadas entre os pesquisadores. A principal instituição foi a “Universidade Federal do Rio de Janeiro”, tendo sido informada por 2 pesquisadores. Esse resultado pode ser observado na saída ilustrada a seguir.

SAÍDA

```
Universidade Federal do Rio de Janeiro
Universidade Estadual de Campinas
Universidade Federal de Santa Catarina
Universidade Federal do Rio de Janeiro
2
```

Também é possível obter informações sobre os estados em que os pesquisadores declaram trabalhar utilizando o método “**state()**”. Esse método permite também obter informações sobre um determinado estado ou uma região. O trecho de código a seguir ilustra como esse método pode ser utilizado, mostrando inclusive como utilizá-lo como o método “**scholarship()**”.

```
state().top(3).println();
state().get(1).println();
state().get(1).total().println();
state(SP).println();
state(SOUTH).println();
state(RJ).scholarship(PQ_1A).println();
```

Os dois principais estados informados foram São Paulo e Rio de Janeiro, com 2 pesquisadores cada um. A região Sul possui apenas 1 pesquisador e no Rio

de Janeiro há 2 pesquisadores com bolsa PQ da categoria 1A. Esse resultado é apresentado na saída a seguir.

SAÍDA

```
RJ
SP
SC
RJ
2
2
1
2
```

No currículo Lattes, um pesquisador pode, atualmente, informar até 6 áreas de atuação, de acordo com as áreas do conhecimento do CNPq. Essas áreas podem ser obtidas utilizando o método "**area()**", ilustrado no trecho de código apresentado a seguir.

```
area().top(1).println();
area().get(1).total().println();
```

A principal área de atuação desses pesquisadores é "Engenharia Mecânica", tendo sido informado por 3 pesquisadores, conforme ilustra a saída a seguir.

SAÍDA

```
Engenharia Mecânica
3
```

O número de publicações em periódicos e em congressos também podem ser obtidos utilizando o método "**publication()**". É possível obter o número de publicações em um determinado ano (**year()**), período (**from()** e **to()**), antes (**before()**) e depois (**after()**) de um determinado ano. O trecho de código a seguir ilustrado como isso pode ser feito, além de apresentar o número total de publicações em periódicos ano a ano em um determinado período.

```
publication(JOURNAL).year(2011).println();
publication(JOURNAL).from(2001).to(2010).println();

publication(JOURNAL).after(2005).println();
publication(JOURNAL).before(2005).println();

publication(CONFERENCE).top(1).println();
publication(CONFERENCE).get(1).total().println();

for (int ano = 2006; ano <= 2012; ano++)
  publication(JOURNAL).year(ano).println();
```

Em 2011 foram publicados 18 artigos em periódicos pelos 5 pesquisadores. No período de 2001 a 2010, foram publicados 199 artigos em periódicos. Depois de 2005 foram publicados 121 artigos e 366 foram publicados antes. O pesquisador que mais publicou em congressos foi “Alvaro Toubes Prata”, com 140 artigos. E no período de 2006 a 2012, 2009 foi o ano com mais publicações em periódicos, com um total de 25 artigos. Essas informações são apresentadas na saída ilustrada a seguir.

SAÍDA

```
18
199
121
336

Alvaro Toubes Prata
140

16
16
18
25
23
18
5
```

Diversas outras informações podem ser obtidas diretamente do banco de dados e mais métodos podem ser implementados na linguagem LattesMiner.

4.1.2. ScopusMiner

A linguagem **ScopusMiner** é uma LDE interna que permite a extração automática de informações de documentos indexados na base Scopus. É composta por um conjunto de classes escritas em Java que permite que outros desenvolvedores implementem suas próprias aplicações com alto nível de abstração e poder de expressão.

A linguagem ScopusMiner permite extrair informações de até 2.000 documentos por vez. Essa é uma limitação imposta pela base Scopus para qualquer tipo de consulta. As informações extraídas podem ser utilizadas para analisar a produção de um pesquisador, de um periódico, de uma instituição, de uma área e até mesmo de um país.

4.1.2.1. Domínio do problema

Para definir os termos do problema foi utilizado como base os termos sugeridos na busca avançada da Scopus, conforme ilustrado na Figura 4.6. Esses termos permitem aos usuários definirem buscas mais específicas e elaboradas.

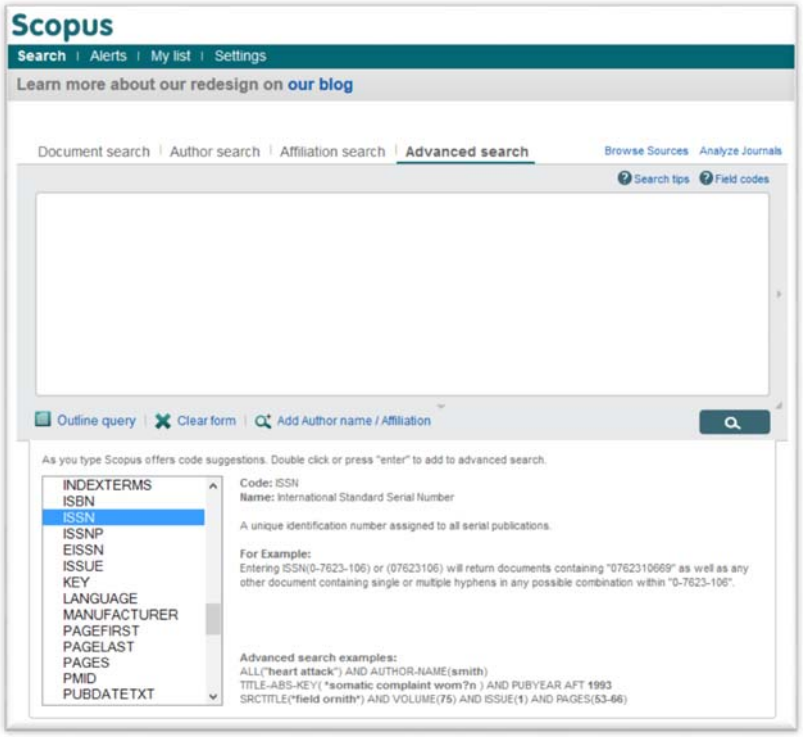


Figura 4.6 - Interface para consulta avançada na Scopus.

Portanto, quando a linguagem ScopusMiner foi projetada esses termos foram considerados. Em alguns casos houve a necessidade de adaptações, porém sempre na tentativa de tornar os termos mais intuitivos e significativos para o usuário. Também foi necessário definir novos termos, pois a linguagem ScopusMiner possui funcionalidades que não estão disponíveis na busca avançada da base Scopus. Assim como os termos sugeridos pela Scopus, os termos do problema também foram definidos somente em inglês.

4.1.2.2. Implementação

A linguagem ScopusMiner é composta por um conjunto de classes escritas em Java e sua classe principal fornece a maioria das funcionalidades da LDE.

A classe **ScopusMiner** é composta por instâncias das classes **Document** e **Author**, além de outras classes aqui não apresentadas. A classe **Document**, por exemplo, contém todos os dados de um documento. A classe **DocumentIE** é responsável pela extração de informações de um documento e a classe **DocumentDao** é responsável pela persistência dessas informações em um banco de dados.

A linguagem ScopusMiner também foi implementada utilizando uma interface fluente, que é implementada utilizando o método de encadeamento, permitindo que qualquer método da linguagem ScopusMiner seja utilizado em qualquer ordem e várias vezes. Além do método de encadeamento, a linguagem ScopusMiner também faz uso de métodos estáticos que permitem criar códigos mais compactos e ainda sim legíveis.

Uma questão que mereceu atenção na implementação da linguagem ScopusMiner é o fato de que a base Scopus limita a visualização dos resultados de uma consulta em 2.000 registros. Para contornar essa limitação, foram desenvolvidos métodos que permitem refinar as consultas, ou seja, realizar consultas que retornem um número menor de registros.

Os dados extraídos pela linguagem ScopusMiner são armazenados em um banco de dados. A Figura 4.7 ilustra as tabelas desse banco de dados. Assim

como na linguagem LattesMiner, a criação do banco de dados e das tabelas também é feita automaticamente na linguagem ScopusMiner. Também há um arquivo de propriedades que pode ser configurado permitindo que os dados sejam armazenados em qualquer banco de dados.

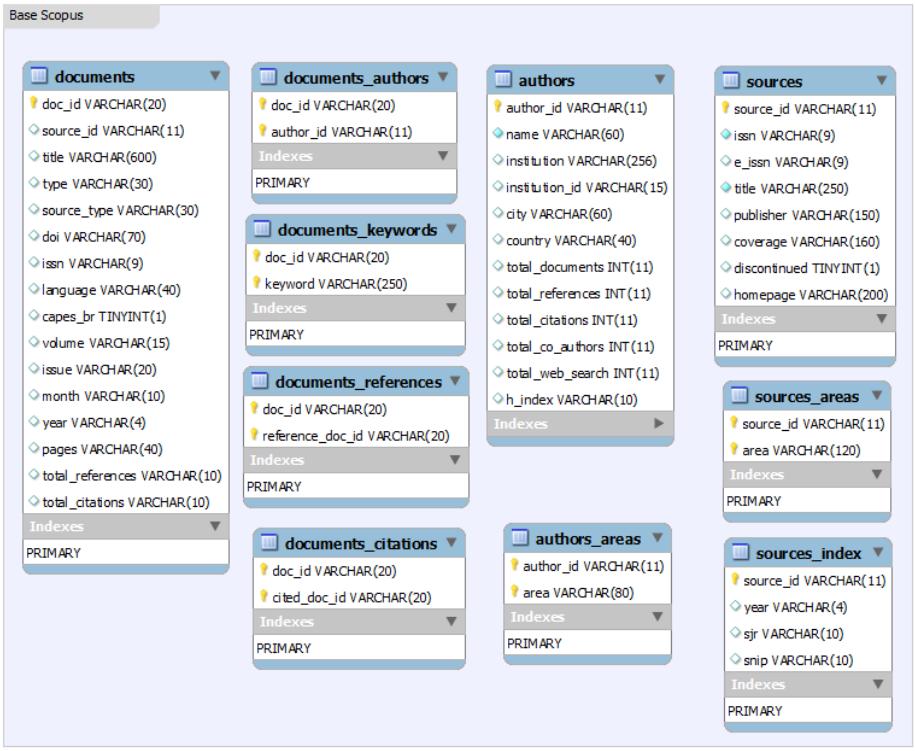


Figura 4.7 - Diagrama das tabelas que armazenam os dados extraídos utilizando a linguagem ScopusMiner.

4.1.2.3. Exemplo de uso

Nesta Subseção é apresentado um exemplo de uso da linguagem ScopusMiner, mostrando como utilizar as principais funcionalidades da linguagem, conforme ilustrado na Listagem 4.6.

O método "**issn()**" permite obter os dados de um determinado periódico de acordo com o seu ISSN. O método "**doctype()**" permite definir o tipo de documento que será consultado. Caso não seja informado, o tipo default é "AR" (Artigo). Há outros tipos que podem ser definidos: CP (artigo de conferência), RE (artigo de revisão), LE (carta), entre outros. Nesse método são permitidos todos os tipos de documentos sugeridos pela base Scopus. E o método

“**extract()**” é o responsável por realizar a extração dos dados e armazená-los em um banco de dados.

ExemploScopus.java

```
import static scopus.miner.ScopusMiner.*;

public class ExemploScopus
{
 public static void main(String[] args)
 {
  // todos os artigos
  issn(“1751-1577”).doctype(AR).extract();

  // somente os artigos de 2011
  issn(“1751-1577”).doctype(AR).year(2011).extract();

  // somente os artigos do volume 4 e da edição 2
  issn(“1751-1577”).volume(4).issue(2).extract();

  // referências
  issn(“1751-1577”).doctype(AR).references().extract();

  // citações
  issn(“1751-1577”).citations().extract();

  // artigos com as referências e citações
  issn(“1751-1577”).references().citations().extract();
 }
}
```

Listagem 4.6 - Exemplo de uma aplicação Java para extração de informações da base Scopus utilizando a linguagem ScopusMiner.

Além desses métodos, há outros métodos que permitem que uma consulta mais refinada seja realizada. Por exemplo, o método “**year()**” permite definir um ano para a consulta. Também é possível definir um determinado volume utilizando o método “**volume()**” e uma determinada edição utilizando o método “**issue()**”. As referências e as citações também podem ser obtidas. Se o método “**references()**” for utilizado o método “**extract()**” também extrai os dados relativos as referências dos documentos. O mesmo é válido para as citações se o método “**citations()**” for utilizado. É válido ressaltar que todos esses métodos podem ser utilizados juntos em uma única linha. O único problema é que obviamente isso demandará mais tempo.

4.2. Sistema SUCUPIRA

O sistema SUCUPIRA é um sistema de extração de informações da PL para identificação de redes sociais acadêmicas. Este sistema é a principal ferramenta do projeto SUCUPURA.

A Figura 4.8 ilustra a página inicial do sistema SUCUPIRA que exibe a localização geográfica atual do usuário.

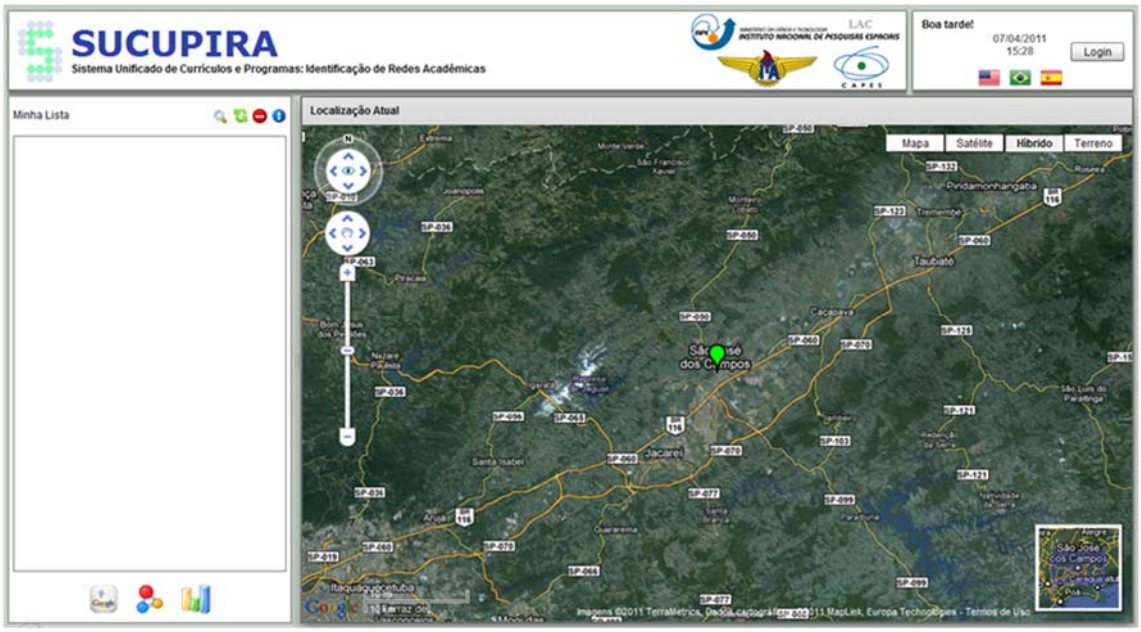


Figura 4.8 - Página inicial do sistema SUCUPIRA.

4.2.1. Arquitetura

O sistema SUCUPIRA foi implementado utilizando a tecnologia Adobe Flex 4.5, que suporta o desenvolvimento de aplicações ricas para Internet e compatíveis com várias plataformas. Para acessar o sistema é necessário que o usuário tenha no seu navegador Web o Adobe Flash Player instalado, aplicativo pelo qual são visualizadas as aplicações Flex.

No lado servidor foi utilizada a linguagem Java. Para que o Adobe Flex consiga se integrar perfeitamente com Java, é necessário ter um *gateway* que possa converter os tipos de dados nativos do Flex (especificamente, do ActionScript) para os tipos de dados nativos do Java e vice-versa. Para isso foi utilizado o

AMF (*Action Message Format*), que é um protocolo de especificação aberta, compacto e que trafega os dados em formato binário. Há várias implementações que suportam o AMF. No sistema SUCUPIRA foi utilizado o *gateway* BlazeDS (http://sourceforge.net/adobe/blazeds/) que foi desenvolvido em Java e é de código aberto.

A Figura 4.9 apresenta uma descrição da arquitetura do sistema SUCUPIRA, destacando os principais componentes.

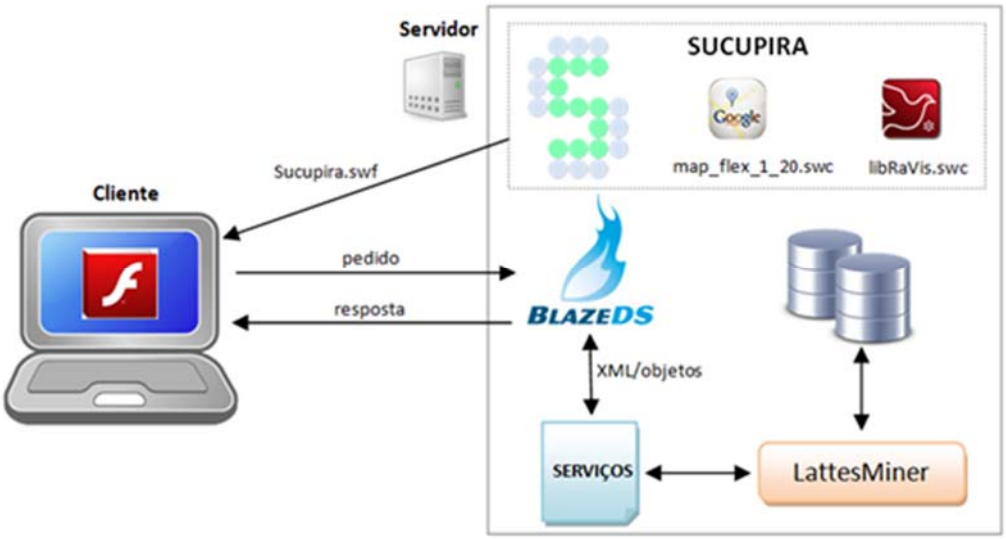


Figura 4.9 - Arquitetura do sistema SUCUPIRA.

Quando um usuário acessa a página do sistema SUCUPIRA, basicamente é retornado o arquivo Flash do sistema (*Sucupira.swf*). O sistema faz uso da biblioteca do *Google Maps* para Flash (*map_flex_1_20.swc*) para exibir a localização atual do usuário e para exibir a distribuição geográfica dos pesquisadores que fazem parte da lista desse usuário. Também faz uso do componente RaVis (*Relational Analysis Visualization*) (*libRaVis.swc*) da biblioteca BirdEye (http://code.google.com/p/birdeye/) para exibir os grafos de relacionamentos entre os pesquisadores.

Em seguida, o sistema fica aguardando as solicitações do usuário. Toda solicitação é atendida pelo *gateway* BlazeDS que, por sua vez, encaminha a solicitação para uma classe escrita em Java que implementa os serviços

oferecidos. Cada serviço oferecido é implementado por um método dessa classe (*Servicos.class*), que faz acesso à linguagem LattesMiner, responsável por extrair as informações dos currículos Lattes dos pesquisadores. As informações extraídas são armazenadas em um banco de dados qualquer que pode ser definido no momento da implantação do sistema. Isso é possível porque a linguagem LattesMiner possui um arquivo texto de propriedades que permite configurar qual gerenciador de banco de dados será utilizado, podendo ser alterado facilmente a qualquer instante.

A comunicação entre o *gateway* BlazeDS e a classe de serviços é feita através de objetos Java ou através de documentos XML. Por fim, o resultado é enviado ao cliente e as informações apresentadas ao usuário na forma de mapas, grafos de relacionamentos, gráficos ou tabelas.

4.2.2. Principais funcionalidades

Uma das principais funcionalidades do sistema SUCUPIRA é gerenciar uma lista de pesquisadores definida por cada usuário do sistema. A ideia é que o usuário adicione na sua lista, chamada no sistema de “Minha Lista”, os pesquisadores que ele deseja comparar e analisar.

Inicialmente, o usuário deve fazer o *login* no sistema. Caso seja o seu primeiro acesso, é necessário se cadastrar. Para isso, basta informar o *login*, o nome, o e-mail (opcional), a senha e se o usuário tiver uma webcam na sua máquina é possível capturar sua foto, conforme ilustra a Figura 4.10.

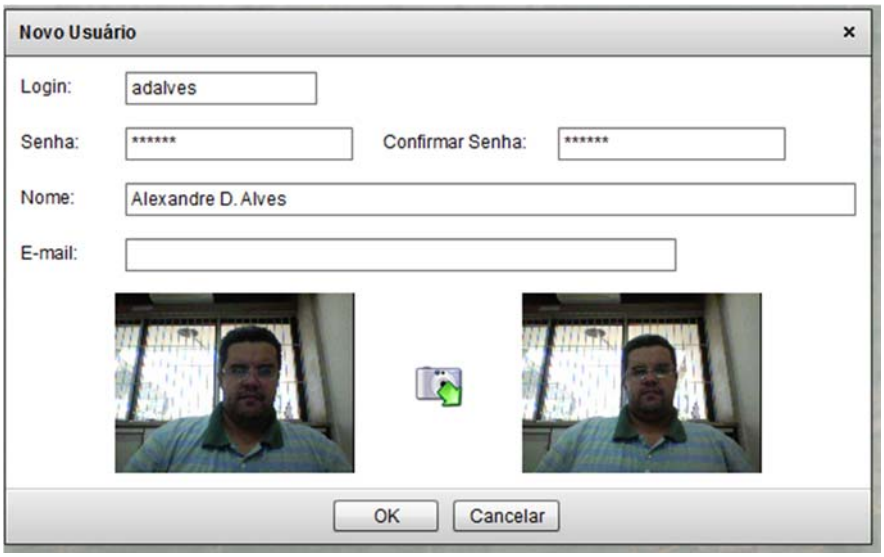


Figura 4.10 - Janela para cadastro de novo usuário no sistema SUCUPIRA.

Após o *login*, é permitido realizar a busca por pesquisadores, bastando clicar no ícone  do componente “Minha Lista” do sistema. Feito isso, é exibida a janela de “Busca por Pesquisadores”, ilustrada parcialmente pela Figura 4.11.

Figura 4.11 - Janela para busca por pesquisadores na PL.

Na janela de “Busca por Pesquisadores”, o usuário deve informar o nome do pesquisador e clicar no ícone  para realizar a busca. Se o pesquisador for encontrado, são exibidas diversas informações que permitem ao usuário confirmar se é realmente o pesquisador procurado. Caso seja, o usuário pode clicar no ícone  para inseri-lo na sua lista. Feito isso, todos os dados desse pesquisador são extraídos e armazenados no banco de dados; o componente

“Minha Lista” do usuário é automaticamente atualizado e um novo pesquisador pode ser procurado.

Para demonstrar as principais funcionalidades do sistema SUCUPIRA foram escolhidos, de forma aleatória, seis pesquisadores, todos Bolsistas de Produtividade em Pesquisa do CNPq de nível 1A. Esses pesquisadores foram adicionados ao componente “Minha Lista”, conforme ilustra a Figura 4.12. Os três primeiros pesquisadores na lista são da área de Medicina e os outros três são da área de Ciência da Computação.

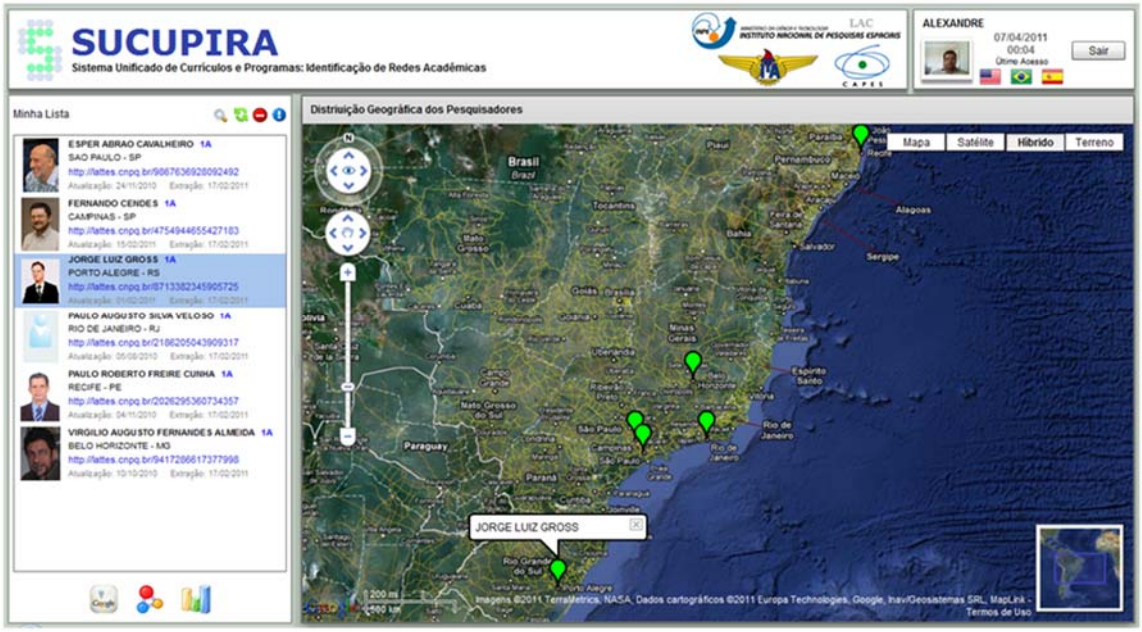


Figura 4.12 - Distribuição geográfica dos pesquisadores.

A Figura 4.12 também ilustra a distribuição geográfica dos pesquisadores que fazem parte do componente “Minha Lista” do usuário. Essa funcionalidade pode ser acionada clicando no ícone . Com isso, é possível visualizar no mapa onde estão trabalhando esses pesquisadores, uma vez que a localização é baseada no endereço profissional informado no currículo Lattes de cada pesquisador.

Também é possível visualizar o gráfico de publicações desses pesquisadores, conforme ilustra a Figura 4.13. Essa funcionalidade é acionada clicando no

ícone , sendo possível aumentar a área de visualização do gráfico, ocultando o componente “Minha Lista”.

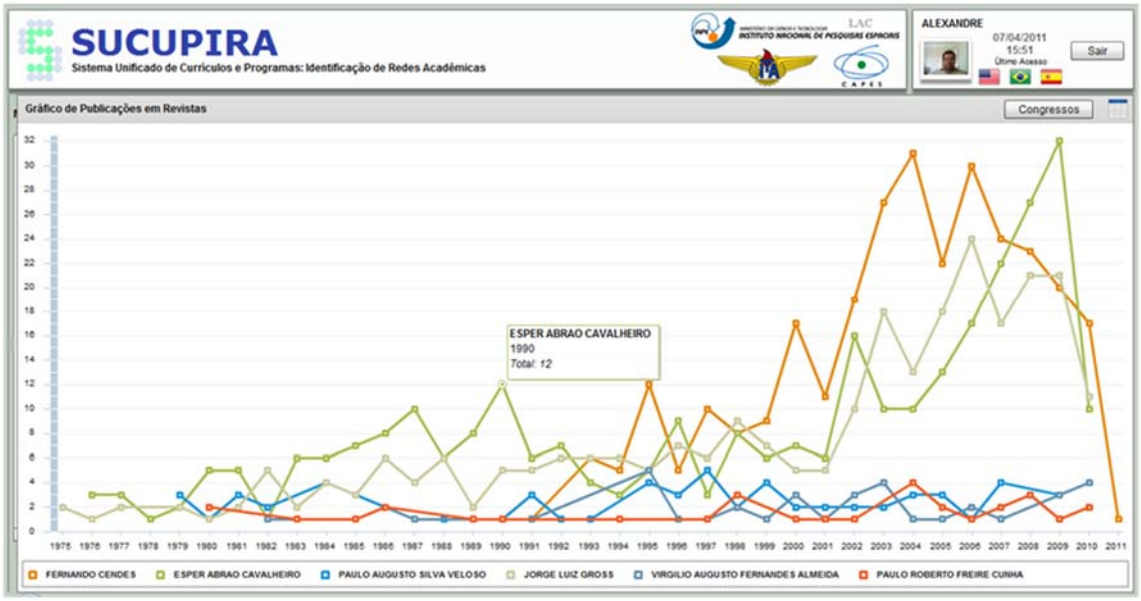
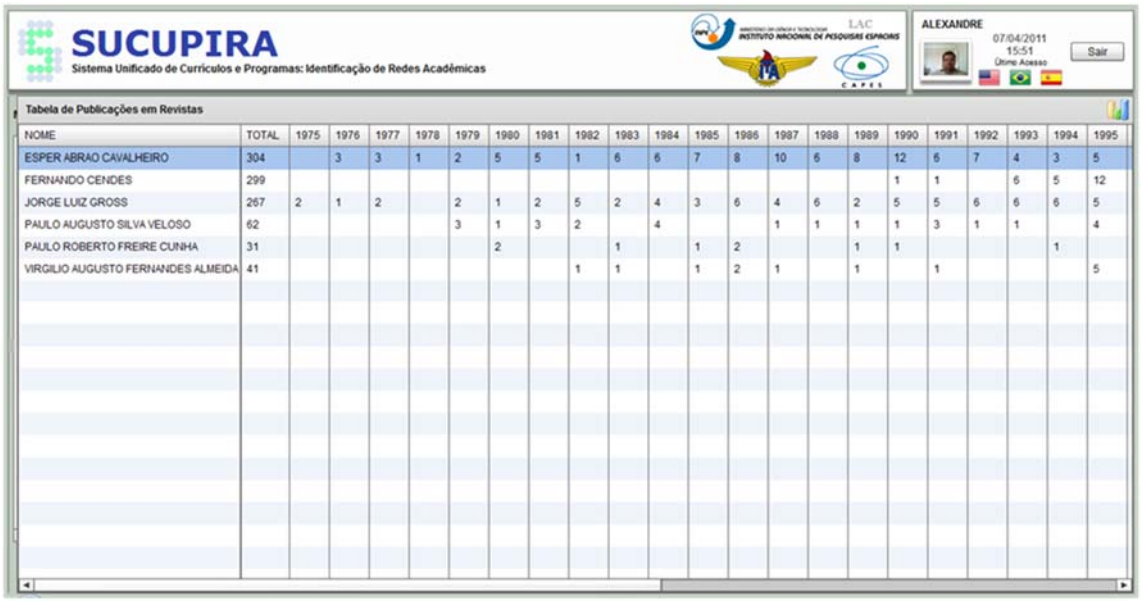


Figura 4.13 - Gráfico de publicações em periódicos.

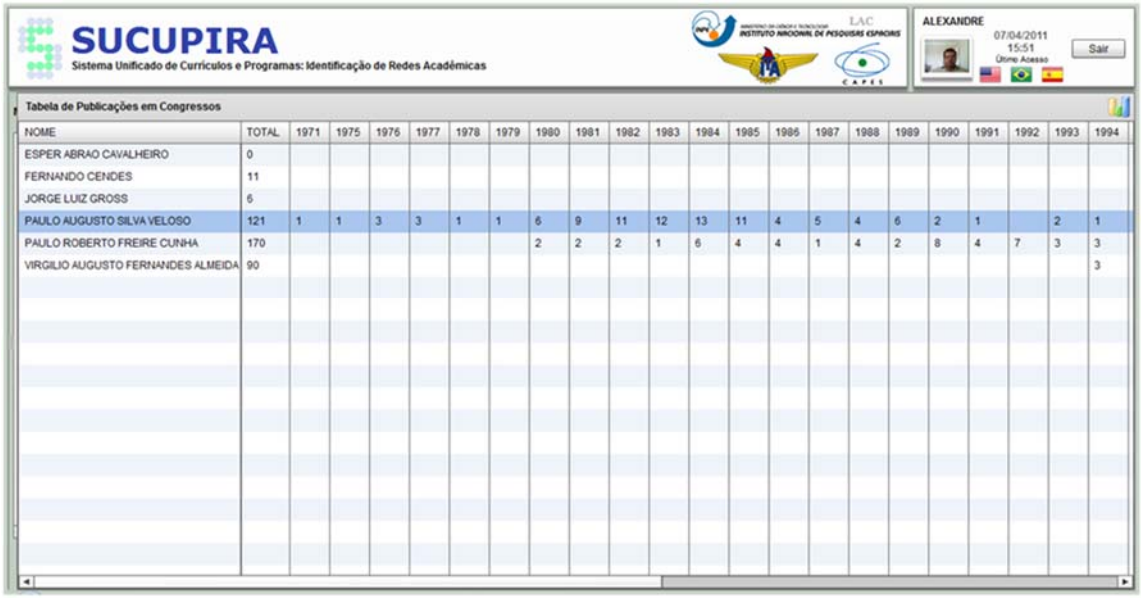
O gráfico apresentado é referente às publicações dos pesquisadores em periódicos. Para visualizar o gráfico de publicações em congressos basta clicar no botão “Congressos”. Para visualizar qualquer um dos gráficos na forma de tabela, basta clicar no ícone e a tabela correspondente é exibida. A Figura 4.14 ilustra a tabela de publicações em periódicos e a Figura 4.15 ilustra a tabela de publicações em congressos.

Analisando essas tabelas, é possível perceber que os três primeiros pesquisadores, todos da área de Medicina, praticamente não publicam em congressos. Por outro lado, o número de publicações em periódicos é significativo. Já os pesquisadores da Ciência da Computação, apesar de também publicarem em periódicos, publicam bem mais em congressos.



NOME	TOTAL	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
ESPER ABRAO CAVALHEIRO	304		3	3	1	2	5	5	1	6	6	7	8	10	6	8	12	6	7	4	3	5
FERNANDO CENDES	299																1	1		6	5	12
JORGE LUIZ GROSS	267	2	1	2		2	1	2	5	2	4	3	6	4	6	2	5	5	6	6	6	5
PAULO AUGUSTO SILVA VELOSO	62					3	1	3	2		4			1	1	1	1	3	1	1		4
PAULO ROBERTO FREIRE CUNHA	31						2			1		1	2			1	1				1	
VIRGILIO AUGUSTO FERNANDES ALMEIDA	41								1	1		1	2	1		1		1				5

Figura 4.14 - Tabela de publicações em periódicos.



NOME	TOTAL	1971	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
ESPER ABRAO CAVALHEIRO	0																					
FERNANDO CENDES	11																					
JORGE LUIZ GROSS	6																					
PAULO AUGUSTO SILVA VELOSO	121	1	1	3	3	1	1	6	9	11	12	13	11	4	5	4	6	2	1		2	1
PAULO ROBERTO FREIRE CUNHA	170							2	2	2	1	6	4	4	1	4	2	8	4	7	3	3
VIRGILIO AUGUSTO FERNANDES ALMEIDA	90																					3

Figura 4.15 - Tabela de publicações em congressos.

A principal funcionalidade do SUCUPIRA é a visualização das redes sociais acadêmicas identificadas entre os pesquisadores. Esta funcionalidade pode ser acionada clicando no ícone  . As redes sociais são representadas no sistema na forma de grafos. Esse grafo é definido verificando os contatos (links para outros currículos Lattes) contidos no currículo Lattes de cada pesquisador.

Todo contato contém o número (ID) do pesquisador, o que permite identificar os relacionamentos entre os pesquisadores.

A Figura 4.16 ilustra o grafo de contatos entre os seis pesquisadores, destacados em vermelho. Esse grafo foi exibido definindo o grau de separação entre os vértices igual a 2. A Figura 4.17 exibe o mesmo grafo considerando o grau de separação igual a 1. Também é possível trocar o tipo de layout, além de outros controles que são fornecidos pelo componente RaVis que são utilizados no sistema.

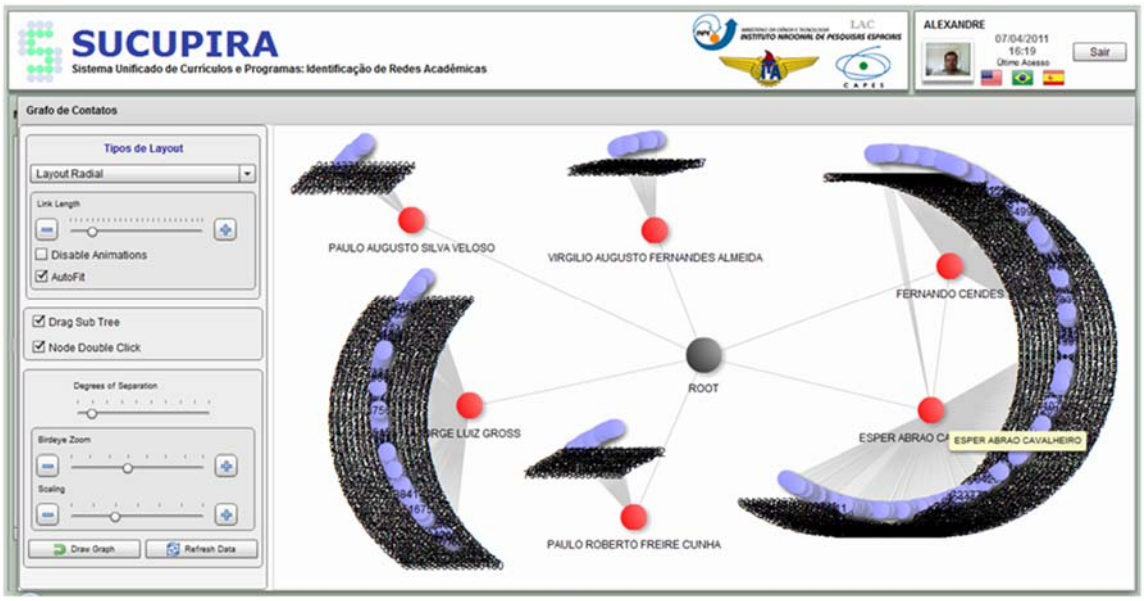


Figura 4.16 - Grafo dos contatos dos pesquisadores com grau 2 de separação.

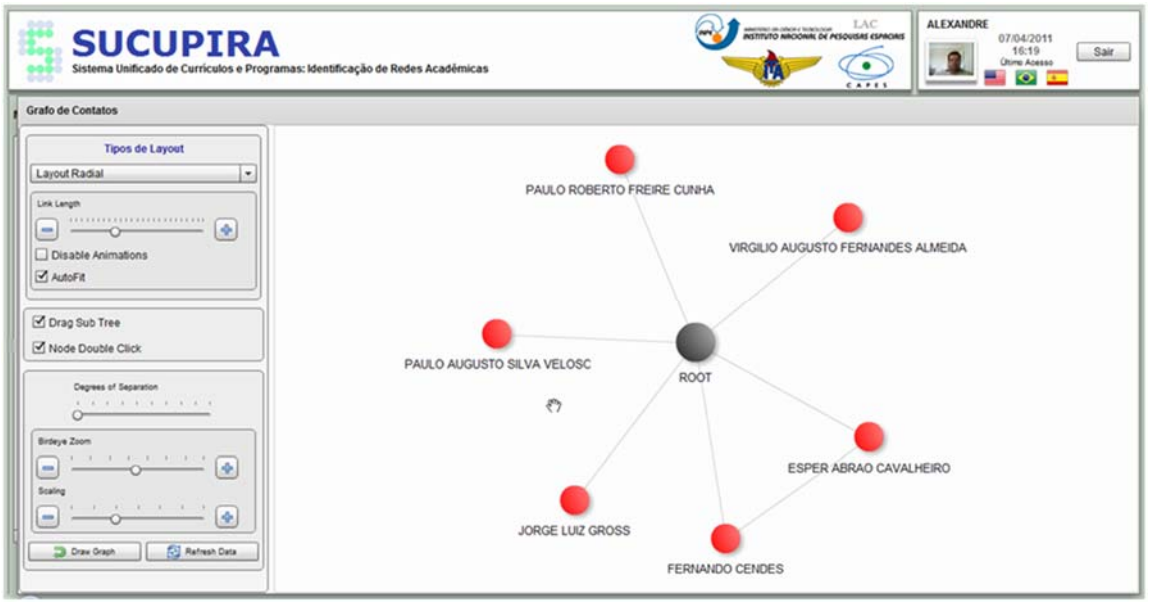


Figura 4.17 - Grafo dos contatos dos pesquisadores com grau 1 de separação.

No grafo da Figura 4.17 observa-se um relacionamento entre dois pesquisadores, ambos da área de Medicina, sendo que um trabalha na Universidade Federal de São Paulo (UNIFESP) e o outro na Universidade Estadual de Campinas (UNICAMP).

Atualmente, apenas um tipo de relacionamento é identificado no sistema SUCUPIRA. Entretanto, com as informações que já são extraídas dos currículos Lattes dos pesquisadores é possível identificar outros tipos de relacionamentos como, por exemplo, relações de orientado/orientador, participação em bancas etc. Além disso, pode-se adicionar informação à própria representação da rede social, por exemplo, incluindo peso nas arestas representando o número de vezes que um determinado relacionamento acontece. Com isso, é possível visualmente observar o quão intenso tais relacionamentos ocorrem.

4.3. Extratores

Algumas ferramentas foram desenvolvidas como extratores, ou seja, ferramentas que permitem extrair automaticamente informações de bases de dados. Além disso, essas ferramentas também permitem armazenar os dados

extraídos em um banco de dados. Pelo fato desses extratores realizarem funções básicas e simples, não houve a necessidade de que essas ferramentas fossem implementadas como uma LDE.

O primeiro extrator desenvolvido permite extrair informações do JCR®. Para isso foi implementado um conjunto de classes escritas em Java. A classe principal possui um único método que permite extrair e armazenar os dados em um banco de dados. Esse método exige como parâmetros a edição do JCR® ("*Science*" ou "*Social Sciences*") que deve ser considerada na extração e o ano. A Listagem 4.7 ilustra como isso é feito.

ExemploJCR.java

```
import jcr.miner.*;

public class ExemploJCR
{
 public static void main(String[] args)
 {
  JCRMiner jcr = new JCRMiner();
  jcr.searchAllJournals("Science", 2012);
 }
}
```

Listagem 4.7 - Exemplo de uma aplicação Java para extrair informações do JCR®.

Os dados extraídos são armazenados em um banco de dados que contém duas tabelas, conforme ilustrado no diagrama da Figura 4.18. Como pode ser observado, é extraída uma grande quantidade de dados. Para isso é necessário realizar diversas buscas no JCR® pois os dados estão disponíveis em mais de uma página Web. Esse processo é um pouco demorado, principalmente quando está sendo feita a extração da edição "*Science*" do JCR®, que possui bem mais periódicos do que a edição "*Social Sciences*". Além disso, deve-se observar que é necessário armazenar os dados, que também é um processo demorado, devido principalmente a grande quantidade de dados. Contudo, todo o processo pode ser realizado em poucas horas.

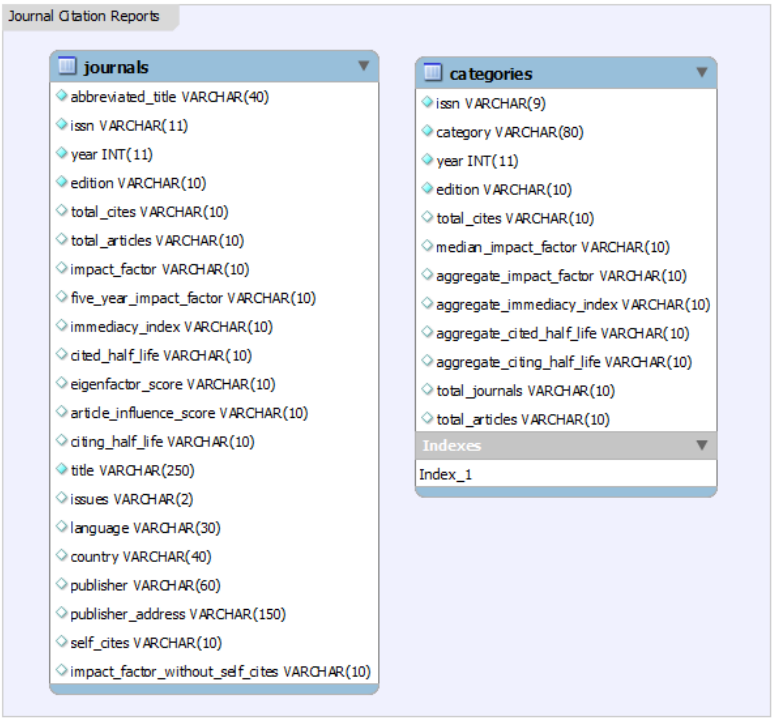


Figura 4.18 - Diagrama das tabelas que armazenam os dados extraídos do JCR®.

Atualmente, estão disponíveis no JCR® os dados das duas edições desde 2007. O número de periódicos em cada uma das edições e de periódicos brasileiros indexados no JCR® são apresentados na Tabela 4.4.

Tabela 4.4 - Número de periódicos indexados no JCR® nas edições "*Science*" e "*Social Sciences*".

Ano	*Science*	*Social Sciences*	Brasil
2007	6.426	1.866	30
2008	6.620	1.985	31
2009	7.387	2.257	71
2010	8.073	2.731	103
2011	8.336	2.966	114
2012	8.471	3.047	118

Neste trabalho já foram extraídos os dados das duas edições de todos os anos disponíveis. Com isso, já é possível realizar uma análise sobre esses dados, permitindo entender, por exemplo, como ocorreu a evolução do FI de um determinado periódico.

Também foi desenvolvida uma ferramenta que permite extrair e armazenar automaticamente informações da base WoS. Para isso foi implementado um conjunto de classes escritas em Java. A classe principal possui somente dois métodos. Um dos métodos permite baixar os dados bibliográficos dos documentos indexados na base WoS e armazená-los em arquivos textos de acordo com a consulta realizada. O outro método permite extrair os dados que estão nos arquivos baixados e armazená-los em um banco de dados.

Na base WoS cada busca permite que 100.000 registros sejam consultados. Porém, os dados bibliográficos dos documentos consultados só podem ser obtidos para 500 registros por vez. Dessa forma, uma busca que retorne 9.850 registros exige que sejam criados 20 arquivos para conter os dados bibliográficos desses documentos. Os 19 primeiros arquivos conterão 500 registros cada e o vigésimo arquivo conterá 350 registros. Esse exemplo é ilustrado na Listagem 4.8. O método "**download()**" é responsável por baixar e armazenar os dados bibliográficos dos documentos em arquivos textos. Esse método exige como parâmetros o número total de registros retornados na busca realizada na base WoS e um diretório indicando onde os arquivos devem ser armazenados. Cada arquivo é criado tendo como nome um número sequencial. Nesse caso, os arquivos terão nomes variando de 1 a 20. O método "**extract()**" é responsável por extrair os dados dos arquivos e armazená-los em um banco de dados.

ExemploWoS.java

```
import isi.miner.*;

public class ExemploWoS
{
 public static void main(String[] args)
 {
  ISIMiner isi = new ISIMiner();
  isi.download(9850, "diretório");

  isi.extract(9850, "diretório");
 }
}
```

Listagem 4.8 - Exemplo de uma aplicação Java para extrair informações da base WoS.

Os dados extraídos dos arquivos são armazenados em um banco de dados contendo várias tabelas, conforme ilustrado no diagrama da Figura 4.19. Apesar de serem muitos dados, o processo de extração e armazenamento em um banco dados é muito rápido. Isso se deve ao fato dos dados estarem em arquivos textos contendo poucos registros (no máximo 500) e contendo poucas informações em cada registro.

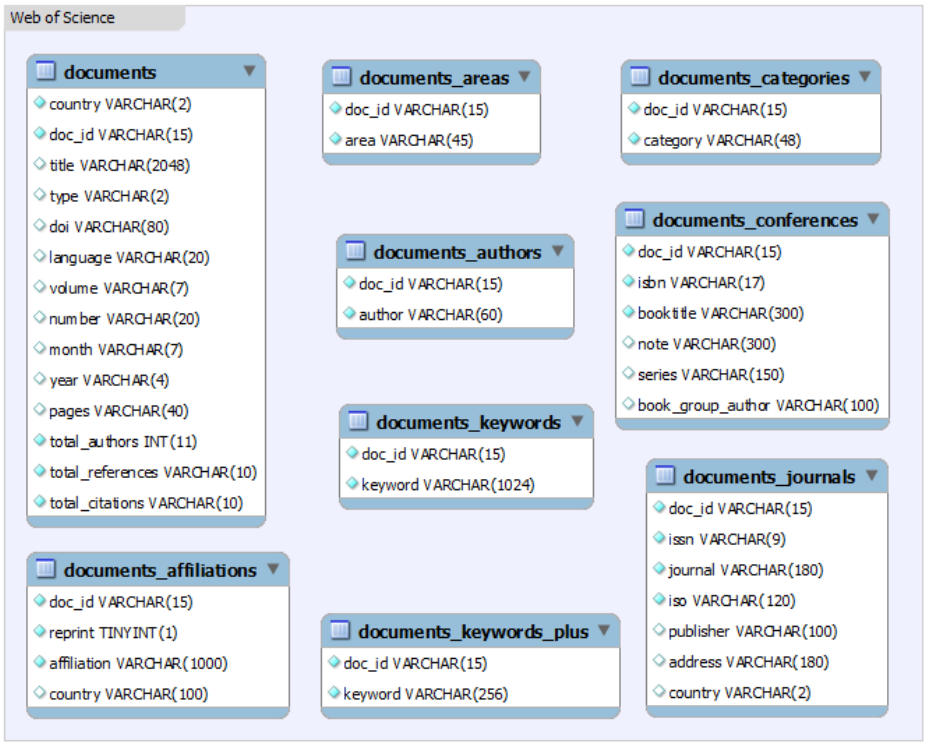


Figura 4.19 - Diagrama das tabelas que armazenam os dados extraídos da base WoS.

Neste trabalho, essa ferramenta foi utilizada para obter os dados bibliográficos de 28.864.820 artigos publicados em periódicos indexados na WoS. Esses artigos são de autores de 68 países e de 6 continentes diferentes, conforme ilustra a Tabela 4.5. Esses dados foram obtidos no período de 23 de dezembro de 2012 a 8 de janeiro de 2013. O número médio de autores desses artigos é 6,09 e o número médio de citações é 18,53 por artigo. A América do Norte é o continente com a maior porcentagem de artigos citados, apesar desse valor não variar muito para os outros continentes.

Tabela 4.5 - Número de artigos e citações de continentes de acordo com dados da base WoS.

Continente	Países	Artigos	Média de Autores	Citações	Média	%
África	7	357.460	7,62	3.226.726	9,03	76,18
América do Norte	3	10.332.624	4,03	252.311.571	24,42	85,72
América do Sul	7	625.890	10,66	6.236.852	9,96	75,54
Ásia	18	5.691.801	6,00	64.306.163	11,30	78,27
Europa	31	11.014.422	7,84	193.939.734	17,61	82,54
Oceania	2	842.623	5,05	14.771.246	17,53	85,59
Total	**68**	**28.864.820**	**6,09**	**534.792.292**	**18,53**	**82,73**

No continente africano, de acordo com os dados dos países analisados na Tabela 4.6, o número médio de autores é 7,62, sendo que esse número é bem maior em “Marrocos”. O número médio de citações é 9,03 por artigo. Nesse quesito o país com mais citações por artigo é o Quênia, que também é o país com a maior porcentagem de artigos citados.

Tabela 4.6 - Número de artigos e citações de países da África de acordo com dados da base WoS.

País	Artigos	Média de Autores	Citações	Média	%
África do Sul	143.704	7,33	1.682.522	11,71	80,03
Argélia	17.696	4,15	108.548	6,13	69,21
Egito	86.180	7,20	579.133	6,72	74,91
Marrocos	21.969	28,89	162.168	7,38	73,51
Nigéria	42.338	2,80	249.259	5,89	70,83
Quênia	19.737	4,76	290.988	14,74	83,44
Tunísia	25.836	4,94	154.108	5,96	69,27

Fonte: Web of Science (08/01/2013)

Na América do Norte o número médio de autores é 4,03 por artigo, o menor valor entre os continentes. O número médio de citações é 24,42 por artigo. O “Estados Unidos” é o país com o menor número médio de autores e o maior número médio de citações, conforme pode ser observado na Tabela 4.7. É interessante observar que tanto o “Canadá” quanto os “Estados Unidos” possuem a mesma porcentagem de artigos citados.

Tabela 4.7 - Número de artigos e citações de países da América do Norte de acordo com dados da base WoS.

País	Artigos	Média de Autores	Citações	Média	%
Canadá	1.217.258	4,92	24.505.886	20,13	85,85
Estados Unidos	8.976.128	3,82	226.366.634	25,22	85,85
México	139.238	9,64	1.439.051	10,34	76,02

Fonte: Web of Science (30/12/2012 a 06/01/2013)

Na América do Sul, a “Venezuela” possui o menor número médio de autores e o “Peru” o maior número médio de citações, conforme pode ser observado na Tabela 4.8. O “Brasil”, entre os países considerados da América do Sul nessa análise, possui o menor número médio de citações, abaixo inclusive do número médio de citações desse continente que é de 9,96 por artigo. É interessante também destacar que a América do Sul possui o maior número médio de autores entre os continentes analisados, sendo que a “Colômbia” possui um número médio de autores muito alto. Além disso, a América do Sul é o continente com a menor porcentagem de artigos citados.

Tabela 4.8 - Número de artigos e citações de países da América do Sul de acordo com dados da base WoS.

País	Artigos	Média de Autores	Citações	Média	%
Argentina	122.677	10,08	1.364.983	11,13	79,95
Brasil	366.068	8,48	3.298.073	9,01	74,06
Chile	69.175	12,62	826.184	11,94	76,47
Colômbia	24.024	50,43	230.872	9,61	69,12
Peru	8.259	9,76	114.789	13,90	78,65
Uruguai	8.605	4,94	110.649	12,86	81,13
Venezuela	27.082	4,45	291.302	10,76	76,09

Fonte: Web of Science (23/12/2012 a 03/01/2013)

Na Ásia, o “Paquistão” é o país com o maior número médio de autores e “Israel” é o país com o maior número médio de citações por artigo, sendo também o país com a maior porcentagem de artigos citados, conforme pode ser observado na Tabela 4.9. A “China” apesar de ser um país que publica muito, o número médio de citações é menor do que o número médio do

continente asiático. O “Japão” é o país que mais publicou e já teve mais de 85% dos seus artigos citados.

Tabela 4.9 - Número de artigos e citações de países da Ásia de acordo com dados da base WoS.

País	Artigos	Média de Autores	Citações	Média	%
Arábia Saudita	49.104	3,73	312.129	6,36	68,96
Bangladesh	13.521	4,48	124.784	9,23	72,73
China	1.314.959	5,94	10.266.234	7,81	71,79
Coreia	436.313	7,74	4.104.905	9,41	75,87
Filipinas	13.428	4,77	183.604	13,67	77,00
Índia	687.803	4,73	5.043.407	7,33	73,88
Irã	123.937	6,64	658.399	5,31	64,91
Israel	296.457	6,65	5.640.749	19,03	86,85
Japão	1.929.891	5,42	30.457.695	15,78	85,58
Jordânia	14.270	3,25	94.409	6,62	74,29
Kuwait	14.107	3,25	109.315	7,75	76,88
Malásia	48.276	4,52	295.363	6,12	66,47
Paquistão	38.855	14,55	214.759	5,53	64,64
Singapura	107.666	4,15	1.355.562	12,59	81,70
Tailândia	54.994	4,96	575.128	10,47	77,41
Taiwan	305.796	8,58	3.130.193	10,24	79,80
Turquia	229.593	8,05	1.631.839	7,11	72,44
Vietnã	12.831	7,69	107.689	8,39	68,12

Fonte: Web of Science (23/12/2012 a 04/01/2013)

Na Europa foram analisados os dados de 31 países e a “Bielorrússia” é o país com o maior número médio de autores, conforme pode ser observado na Tabela 4.10. A “Suíça” é o país com o maior número médio de citações por artigo e a “Sérvia” o país com o menor número. A “Suécia” também é um país com um alto número médio de citações por artigo e é o país com a maior porcentagem de artigos citados. A “Bielorrússia” é o país da Europa com a menor porcentagem de artigos citados. É interessante destacar que na Europa há 8 países que possuem mais de 85% dos artigos citados.

Tabela 4.10 - Número de artigos e citações de países da Europa de acordo com dados da base WoS.

País	Artigos	Média de Autores	Citações	Média	%
Alemanha	1.556.167	6,32	27.847.774	17,90	83,34
Áustria	215.504	10,74	3.571.470	16,57	82,35
Bélgica	315.664	7,28	5.939.936	18,82	84,60
Bielorrússia	21.291	47,87	122.222	5,74	60,73
Croácia	36.276	18,84	275.756	7,60	69,53
Dinamarca	241.281	7,88	5.676.357	23,53	89,21
Eslováquia	43.376	24,76	357.016	8,23	73,59
Eslovênia	38.197	23,84	367.219	9,61	75,18
Espanha	653.588	8,12	8.946.168	13,69	81,18
Estônia	15.083	30,80	187.855	12,45	80,10
Finlândia	209.392	8,94	4.188.974	20,01	87,94
França	1.494.036	6,16	25.883.919	17,32	81,95
Grécia	150.046	14,02	1.791.942	11,94	82,80
Holanda	586.301	7,39	13.457.583	22,95	88,37
Hungria	136.764	12,83	1.618.427	11,83	78,26
Inglaterra	1.876.482	5,08	41.623.822	22,18	86,25
Irlanda	134.541	6,58	2.063.303	15,34	82,80
Islândia	10.197	7,59	237.465	23,29	85,64
Itália	933.359	8,08	15.350.341	16,45	84,05
Lituânia	19.130	25,05	136.464	7,13	68,88
Luxemburgo	4.615	5,86	50.207	10,88	73,85
Noruega	170.218	10,11	3.147.720	18,49	86,66
Polônia	338.575	8,45	3.063.392	9,05	75,84
Portugal	104.233	15,50	1.263.260	12,12	80,77
República Checa	110.011	16,72	1.122.901	10,21	77,92
Romênia	77.994	14,94	442.998	5,68	64,13
Rússia	524.970	8,58	3.347.827	6,38	62,46
Sérvia	24.592	44,05	100.136	4,07	59,35
Suécia	460.933	6,64	10.821.972	23,48	89,40
Suíça	436.798	9,50	10.726.476	24,56	85,64
Ucrânia	85.005	9,65	446.297	5,25	60,99

Fonte: Web of Science (26/12/2012 a 03/01/2013)

Na Oceania foram analisados os dados de 2 países, conforme pode ser observado na Tabela 4.11. A “Austrália” é o país com o menor número médio de autores e o maior número médio de citações por artigo.

Tabela 4.11 - Número de artigos e citações de países da Oceania de acordo com dados da base WoS.

País	Artigos	Média de Autores	Citações	Média	%
Austrália	702.654	4,79	12.541.516	17,85	85,61
Nova Zelândia	139.969	6,34	2.229.730	15,93	85,51

Fonte: Web of Science (26/12/2012)

A principal vantagem de obter dados da base WoS é o fato de cada busca retornar até 100.000 registros. Isso facilita muito o trabalho de obter uma grande quantidade de dados. Com a limitação que a base Scopus impõe de retornar apenas 2.000 registros, esse trabalho seria muito custoso e ainda mais demorado.

Para obter os dados dos Cursos de Pós-Graduação recomendados e reconhecidos pela CAPES foi desenvolvido um conjunto de classes escritas em Java. A classe principal permite executar os métodos responsáveis pela extração e armazenamento dos dados. Um exemplo é mostrado na Listagem 4.9.

ExemploCursosPosGraduacaoCAPES.java

```
import capes.miner.*;

public class ExemploCursosPosGraduacaoCAPES
{
 public static void main(String[] args)
 {
  CapesMiner cm = new CapesMiner();
  cm.largesAreas();
  cm.areas();
  cm.programs();
  cm.institutions();
  cm.courses();
 }
}
```

Listagem 4.9 - Exemplo de uma aplicação Java para extrair os cursos de Pós-Graduação recomendados e reconhecidos pela CAPES.

Cada um dos métodos pode ser executado de maneira independente. Entretanto, internamente, pode ser necessário dados que são obtidos através

dos outros métodos. Por exemplo, para obter os dados dos cursos é necessário primeiro obter os dados das grandes áreas, áreas, instituições e programas. Isso pode ser observado no diagrama da Figura 4.20 que ilustra as tabelas que armazenam os dados dos cursos de Pós-Graduação. Note que na tabela que armazena os cursos é necessário informar o código da grande área, da área, da instituição e do programa no qual o curso está relacionado.

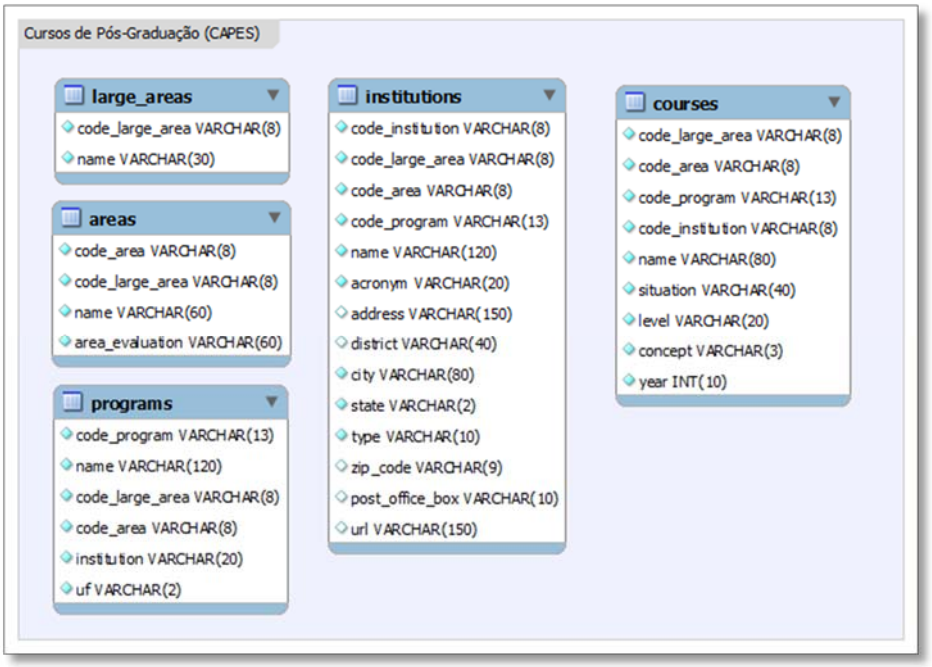


Figura 4.20 - Diagrama das tabelas que armazenam os dados extraídos da base de Cursos de Pós-Graduação recomendados e reconhecidos pela CAPES.

Atualmente, há 3.801 programas de Pós-Graduação e 5.661 cursos de mestrado (3.158; 55,79%), doutorado (1.922; 33,95%) ou mestrado profissional (581; 10,26%) recomendados e reconhecidos pela CAPES. Por isso, obter e armazenar todos esses dados é um processo um pouco demorado, mas normalmente possível de ser realizado em poucas horas.

No caso das Bolsas em curso do CNPq, foi desenvolvido apenas um método em Java que permite extrair os nomes dos pesquisadores com bolsas PQ ativas e armazená-los em um arquivo texto. A Listagem 4.10 mostra como o método pode ser utilizado. É necessário informar apenas a URL da página das Bolsas em curso que está sendo consultada. A consulta pode ser realizada

pela área de conhecimento, por estado ou por uma instituição. Também é necessário informar o nome do arquivo texto em que os nomes dos bolsistas serão armazenados. Basicamente, o método utiliza uma expressão regular que permite extrair os nomes dos bolsistas.

ExemploBolsasEmCurso.java

```
public class ExemploBolsasEmCurso
{
 public static void main(String[] args)
 {
  BolsasPQ bpq = new BolsasPQ();
  bpq.extrair("URL", "nomeArquivo.txt");
 }
}
```

Listagem 4.10 - Exemplo de uma aplicação Java para extrair os nomes dos pesquisadores com bolsas PQ ativas no CNPq.

Em seguida, é possível utilizar a linguagem LattesMiner para identificar o ID desses pesquisadores, conforme já mostrado anteriormente.

4.4. Conversores

Inicialmente, o Qualis Periódicos da CAPES permitia exportar os dados de cada uma das áreas de avaliação para uma planilha. Atualmente, o *SCImago Journal & Country Rank* permite que os dados disponíveis possam ser exportados para planilhas. Por isso, foi necessário definir um mecanismo que permitisse converter planilhas para tabelas em banco de dados.

Os seguintes passos devem ser realizados para converter uma planilha qualquer em uma tabela em um banco de dados no MySQL:

1. Abrir a planilha no Excel e remover todos os títulos das colunas. Na planilha devem permanecer somente os dados.

2. Verificar se há a ocorrência do caractere ";". Caso haja, o mesmo deve ser colocado entre aspas simples ou duplas. Normalmente, esse caractere é utilizado como caractere de separação para delimitar um

campo. O caractere quebra de linha também é utilizado como um delimitador.

3. Salvar a planilha no formato de arquivo de texto “CSV” (*comma-separated values*). No Excel, há essa opção no “Salvar como” (Tipo).

4. Criar uma tabela em um banco de dados no MySQL de acordo com as colunas definidas na planilha. O tipo de cada campo na tabela deve ser compatível com os tipos dos dados da planilha.

5. Executar o seguinte comando no MySQL:

   LOAD DATA LOCAL INFILE ‘D:/planilhas/planilha.csv’ INTO TABLE tabela FIELDS TERMINATED BY ‘;’ LINES TERMINATED BY ‘\n’

   Caso não ocorra nenhum erro, será criada uma tabela no banco de dados que estiver ativo no MySQL no momento da execução desse comando. Nesse caso, a tabela conterá um número de registros igual ao número de linhas da planilha em questão.

Neste trabalho, esses passos foram executados para essas duas bases. Com isso, foram criadas 3 tabelas, conforme ilustra o digrama na Figura 4.21.

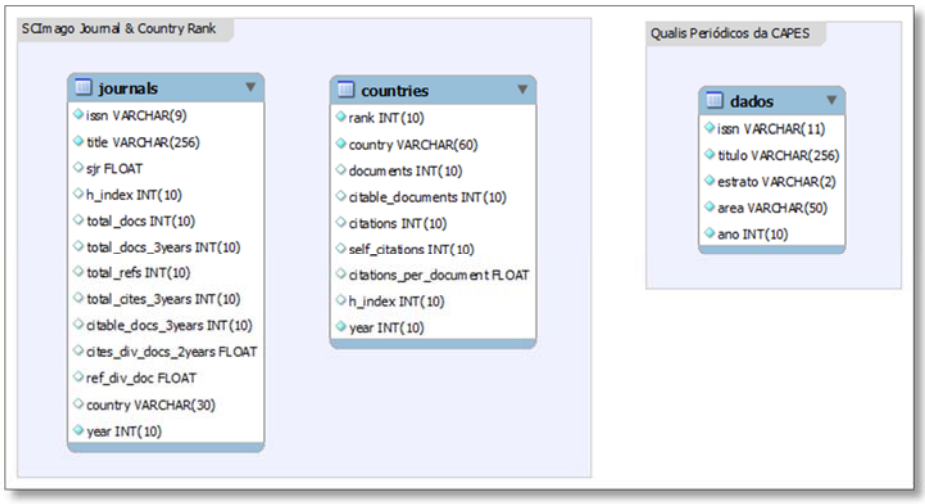


Figura 4.21 - Diagrama das tabelas que armazenam os dados extraídos das bases *SCImago Journal & Country Rank* e Qualis Periódicos da CAPES.

4.5. Considerações finais

Neste Capítulo foram apresentadas todas as ferramentas computacionais desenvolvidas neste trabalho. Cada uma das ferramentas permite extrair informações de uma base de dados. Além disso, permitir análises mais abrangentes, isso também possibilita contornar as limitações das bases de dados. Por exemplo, algumas bases de dados possuem dados somente de pesquisadores brasileiros ou de interesse apenas em âmbito nacional. Há bases de dados que limitam o acesso aos dados e dificulta a obtenção de uma grande quantidade de dados. Por isso, as bases de dados consideradas neste trabalho se integram de maneira a permitir que estudos mais complexos possam ser realizados. No próximo Capítulo é proposto um novo índice que permite medir a colaboração entre os autores de um artigo.

5 ÍNDICE DE COLABORAÇÃO

Neste Capítulo é proposto um novo índice, denominado Índice de Colaboração (IC), que busca medir a colaboração entre autores em um determinado artigo. É demonstrada a sua utilização com exemplos e estudos de casos, além de destacar suas principais vantagens e limitações.

5.1. Motivação

Para compreender a produção e o uso do conhecimento científico, é necessário entender como os pesquisadores se comportam, se relacionam, se organizam e como transmitem informações entre si (VANZ; STUMPF, 2010). A colaboração entre duas pessoas, uma forma de relacionamento, é um processo social e de interação humana que pode ocorrer de várias formas e por diferentes razões. A colaboração científica tem sido definida como dois ou mais pesquisadores trabalhando juntos em um projeto de pesquisa, compartilhando recursos intelectuais, econômicos e/ou físicos (VANZ; STUMPF, 2010).

Vanz e Stumpf (2010) definiram uma lista de 17 motivos que levam um pesquisador a colaborar:

1. Desejo de aumentar a popularidade científica, a visibilidade e o reconhecimento pessoal;
2. Aumento da produtividade;
3. Racionalização do uso da mão-de-obra científica e do tempo dispensado à pesquisa;
4. Redução da possibilidade de erro;
5. Obtenção e/ou ampliação de financiamentos, recursos, equipamentos especiais, materiais;
6. Aumento da especialização na Ciência;
7. Possibilidade de “ataque” a grandes problemas de pesquisa;

8. Crescente profissionalização da ciência;
9. Desejo de aumentar a própria experiência através da experiência de outros pesquisadores;
10. Desejo de realizar pesquisa multidisciplinar;
11. União de forças para evitar a competição;
12. Treinamento de pesquisadores e orientandos;
13. Necessidade de opiniões externas para confirmar ou avaliar um problema;
14. Possibilidade de maior divulgação da pesquisa;
15. Como forma de manter a concentração e a disciplina na pesquisa até a entrega dos resultados ao restante da equipe;
16. Compartilhamento do entusiasmo por uma pesquisa com alguém;
17. Necessidade de trabalhar fisicamente próximo a outros pesquisadores, por amizade e desejo de estar com quem se gosta.

O advento da Internet e das redes sem fio, permitindo uma comunicação com baixo custo, facilitou o contato de pesquisadores separados por grandes distâncias e ampliou as possibilidades de colaboração entre eles (VANZ; STUMPF, 2010). As motivações para a colaboração não são as mesmas em todas as áreas do conhecimento. Na Matemática, por ser uma área teórica, as parcerias tendem a resultar da necessidade de trocar ideias e debater problemas. Na Física, a colaboração ocorre mais pela necessidade de compartilhar equipamentos cada vez mais caros e complexos, como telescópios e aceleradores de partículas (VANZ, 2009). Katz e Martin (1997) afirmam que, em geral, os trabalhos teóricos produzem artigos com poucos autores comparados com trabalhos experimentais.

Segundo Glänzel e Lange (2002), trabalhos que possuem colaboração internacional usualmente apresentam maior visibilidade e impacto, o que é observado pelo maior número de citações.

É fato que a colaboração entre pesquisadores está aumentando (STALLINGS et al., 2013). Porém, isso é observado devido ao aumento no número de autores nos artigos. Vanz e Stumpf (2010), também observaram que a colaboração científica aparece, muitas vezes, na literatura, relacionada à coautoria. Porém, apesar dos dois termos serem considerados sinônimos pelos pesquisadores, a coautoria é apenas uma faceta da colaboração. Segundo as autoras, a coautoria tem sido utilizada por pesquisadores das áreas de bibliometria e cientometria para investigar a colaboração entre pesquisadores, instituições e países. Segundo Katz e Martin (1997) a colaboração científica pode ser estudada considerando outros indicadores, porém a coautoria é o indicador mais utilizado.

5.2. Trabalhos relacionados

Há outros índices que medem a colaboração entre pesquisadores. Do melhor do nosso conhecimento, o primeiro deles foi proposto por Lawani (1986). O índice proposto por ele descreve o número médio de autores por artigo para um determinado conjunto de artigos. Segundo Lawani (1986), um número alto de coautores é usualmente associado a um alto número de citações. A limitação desse índice é o cálculo para artigos de um único autor, uma vez que não representam colaboração (MOHAMMADHASSANZADEH et al., 2011). Outro índice, denominado Grau de Colaboração (GC), é definido como a proporção entre o número de artigos com um único autor e o número total de artigos (SUBRAMANYAM, 1983). Porém, o GC não diferencia artigos com muitos autores quando o número de autores varia (AJIFERUKE et al., 1988; MOHAMMADHASSANZADEH et al., 2011).

O Coeficiente de Colaboração (CC) é um índice que permite definir o nível de colaboração entre pesquisadores a partir dos artigos publicados em um

determinado intervalo de tempo (AJIFERUKE et al., 1988). O cálculo do CC é feito utilizando a seguinte equação:

$$CC = 1 - \frac{\sum_{j=1}^{k} \left(\frac{1}{j}\right) f_j}{N}$$

em que f_j é o número de artigos publicados com *j* autores; *N* é o número total de artigos publicados e *k* é o maior número de autores por artigo. Esse índice incorpora características dos índices anteriores, uma vez que ele reflete tanto o número médio de autores por artigo bem como a proporção de artigos com vários autores. O CC diferencia vários níveis de múltiplas autorias, ou seja, um valor de CC muito alto indica uma maioria de artigos com muitos autores. Além disso, quando artigos com um único autor são maioria, esse índice tenderá a zero (MOHAMMADHASSANZADEH et al., 2011).

A colaboração como índice também é utilizada para medir a contribuição de cada autor em um artigo, uma vez que os autores não contribuem igualmente e que quanto maior o número de autores, maior também a dificuldade para medir a contribuição de cada um (STALLINGS et al., 2013). Além da colaboração, a internacionalização também é calculada através de um índice que considera quantos países citaram um artigo (KOSMULSKI, 2010). Com esse índice é possível medir o quanto um pesquisador, um periódico ou uma instituição tem abrangência geográfica.

5.3. Definição

O IC proposto mede a colaboração entre os autores de um determinado artigo levando em consideração o número de autores, suas instituições, cidades e países. Apesar do índice se chamar índice de colaboração, na verdade ele é baseado apenas em coautorias. Dessa forma, a colaboração pode ser medida pelo número de autores por artigo, sendo esse o principal atributo do IC e todo o cálculo realizado em função desse atributo.

O IC tem um valor entre 0 e 1, sendo que “0” indica que o artigo para o qual o índice está sendo calculado possui apenas um único autor e, “1” que indica que

todos os autores são de países distintos, representando uma colaboração internacional maximal. O IC fornece um indicativo do nível de abrangência geográfica de colaboração que ocorreu entre os autores de um artigo. Assim, é possível verificar como se dá a abrangência geográfica de colaboração de um pesquisador, de uma instituição, de uma área ou até mesmo de um país.

5.4. Cálculo

Para um número de autores igual a 1, o IC é definido igual a 0. Para um número de autores maior do que 1, o IC é calculado da seguinte maneira:

$$\mathbf{IC} = \frac{(\mathbf{NA}-\mathbf{1}).\mathbf{p_1} + (\mathbf{NI}-\mathbf{1}).\mathbf{p_2} + (\mathbf{NC}-\mathbf{1}).\mathbf{p_3} + (\mathbf{NP}-\mathbf{1}).\mathbf{p_4}}{\mathbf{NA}-\mathbf{1}}$$

sendo que:

- NA: é o número de autores em um artigo.
- NI: é o número de instituições distintas dos autores. É um valor entre 1 e NA.
- NC: é o número de cidades distintas dos autores. É um valor entre 1 e NA. A seguinte relação deve ser observada: NC ≤ NI.
- NP: é o número de países distintos dos autores. É um valor entre 1 e NA. As seguintes relações devem ser observadas: NP ≤ NI e NP ≤ NC.
- p_1, p_2, p_3 e p_4 são pesos atribuídos de forma que $p_1 + p_2 + p_3 + p_4 = 1$. Como uma forma de representar a abrangência geográfica da colaboração entre os autores, sendo que o mais valorado é o número de países, devem-se utilizar os seguintes valores: 0,1; 0,2; 0,3 e 0,4 respectivamente.

Os dados para o cálculo do IC podem ser obtidos diretamente de um artigo ou a partir de bases de dados como a WoS e a Scopus.

5.5. Exemplos

Na Tabela 5.1 são apresentados alguns exemplos do cálculo do IC.

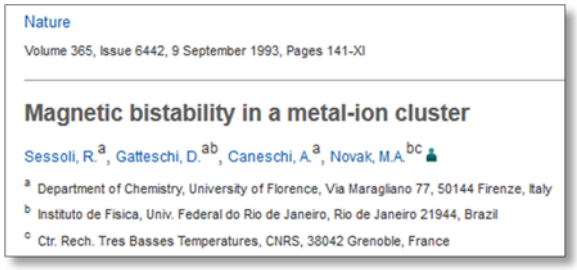
Tabela 5.1 - Exemplos de cálculo do IC.

n	NA	NI	NC	NP	Cálculo	IC
1	3	1	1	1	((3 – 1).0,1 + (1 – 1).0,2 + (1 – 1).0,3 + (1 – 1).0,4) / (3 – 1)	0,10
2	4	3	1	1	((4 – 1).0,1 + (3 – 1).0,2 + (1 – 1).0,3 + (1 – 1).0,4) / (4 – 1)	0,23
3	2	2	1	1	((2 – 1).0,1 + (2 – 1).0,2 + (1 – 1).0,3 + (1 – 1).0,4) / (2 – 1)	0,30
4	5	5	2	1	((5 – 1).0,1 + (5 – 1).0,2 + (2 – 1).0,3 + (1 – 1).0,4) / (5 – 1)	0,38
5	4	4	3	1	((4 – 1).0,1 + (4 – 1).0,2 + (3 – 1).0,3 + (1 – 1).0,4) / (4 – 1)	0,50
6	3	2	2	2	((3 – 1).0,1 + (2 – 1).0,2 + (2 – 1).0,3 + (2 – 1).0,4) / (3 – 1)	0,55
7	8	4	2	2	((8 – 1).0,1 + (4 – 1).0,2 + (2 – 1).0,3 + (2 – 1).0,4) / (8 – 1)	0,29
8	4	4	2	1	((4 – 1).0,1 + (4 – 1).0,2 + (2 – 1).0,3 + (1 – 1).0,4) / (4 – 1)	0,40
9	6	6	6	5	((6 – 1).0,1 + (6 – 1).0,2 + (6 – 1).0,3 + (5 – 1).0,4) / (6 – 1)	0,92
10	4	4	4	4	((4 – 1).0,1 + (4 – 1).0,2 + (4 – 1).0,3 + (4 – 1).0,4) / (4 – 1)	1

Os valores calculadores de IC na Tabela 5.1 permitem entender como a colaboração entre os autores de um artigo ocorreu. Por exemplo, é possível verificar que quando o IC é igual a 0,10, independentemente do número de autores, houve apenas colaboração “local”, ou seja, a coautoria do artigo foi apenas com pesquisadores de uma única instituição. Também é possível verificar que toda vez que o número de países for 1, o IC sempre será menor ou igual a 0,60. Isso não significa que o IC deverá ser maior que 0,60 quando a colaboração ocorreu com autores de mais de um país. Isso pode ser verificado no exemplo 6 da Tabela 5.1, em que o IC é 0,55 e ocorreu a colaboração com autores de mais de um país. Porém, sempre que o IC for maior que 0,60 significa que houve colaboração “internacional”, ou seja, uma colaboração com autores de pelo menos 2 países distintos. Pode-se afirmar também que sempre que o IC resultar em um valor maior que 0,30 e menor ou igual a 0,60, houve colaboração “regional”, ou seja, uma colaboração entre autores de pelo menos

2 cidades distintas. Toda vez que o IC resultar em um valor maior que 0,10 e menor ou igual a 0,30, significa que houve colaboração “institucional”, ou seja, uma colaboração entre autores de pelo menos 2 instituições distintas. E é possível observar também que toda vez que o número de autores for igual ao número de países, o IC sempre será 1, exceto é claro quando o número de autores for 1, pois nesse caso o IC é 0 e significa que não houve tipo algum de colaboração.

Consideramos a seguir exemplos baseados em artigos indexados na base Scopus escolhidos de forma aleatória. O primeiro exemplo foi obtido na base Scopus e é um artigo publicado na “Nature” por 4 autores em 1993, que contou com a colaboração de autores brasileiros, conforme ilustra a Figura 5.1.


Nature

Volume 365, Issue 6442, 9 September 1993, Pages 141-XI

Magnetic bistability in a metal-ion cluster

Sessoli, R.[a], Gatteschi, D.[ab], Caneschi, A.[a], Novak, M.A.[bc]

[a] Department of Chemistry, University of Florence, Via Maragliano 77, 50144 Firenze, Italy

[b] Instituto de Fisica, Univ. Federal do Rio de Janeiro, Rio de Janeiro 21944, Brazil

[c] Ctr. Rech. Tres Basses Temperatures, CNRS, 38042 Grenoble, France

Figura 5.1 - Dados de um artigo indexado na Scopus para o cálculo do IC (Exemplo 1).

O IC para esse artigo é calculado da seguinte maneira:

$$IC = ((4 – 1).0,1 + (3 – 1).0,2 + (3 – 1).0,3 + (3 – 1).0,4) / (4 – 1) = 0,70$$

Como o IC nesse exemplo é maior do que 0,60, pode-se afirmar que houve colaboração “internacional” na coautoria desse artigo, o que pode ser comprovado facilmente verificando as afiliações dos autores na Figura 5.1. Pode-se então verificar que os 4 autores desse artigo são de 3 países distintos.

O segundo exemplo é um artigo publicado na “Science” por 14 autores em 2007, conforme ilustra a Figura 5.2.

Science

Volume 317, Issue 5842, 31 August 2007, Pages 1196-1199

Superconducting interfaces between insulating oxides

Reyren, N.[a], Thiel, S.[b], Caviglia, A.D.[a], Fitting Kourkoutis, L.[c], Hammerl, G.[b], Richter, C.[b], Schneider, C.W.[b], Kopp, T.[b], Rüetschi, A.-S.[a], Jaccard, D.[a], Gabay, M.[d], Muller, D.A.[c], Triscone, J.-M.[a], Mannhart, J.[b]

[a] Département de Physique de la Matière Condensée, University of Geneva, 24 quai Ernest-Ansermet, 1211 Genève 4, Switzerland

[b] Center for Electronic Correlations and Magnetism, Institute of Physics, University of Augsburg, D-86135 Augsburg, Germany

[c] School of Applied and Engineering Physics, Cornell University, Ithaca, NY 14853, United States

[d] Laboratoire de Physique des Solides, Bat. 510, Centre d'Orsay, 91405 Orsay, Cedex, France

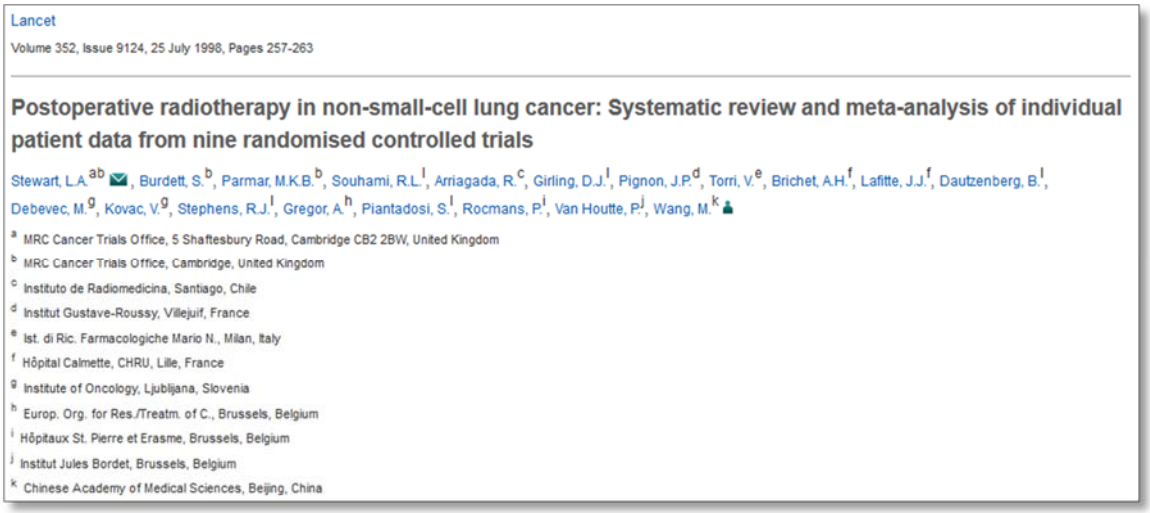
Figura 5.2 - Dados de um artigo indexado na Scopus para o cálculo do IC (Exemplo 2).

O IC para esse artigo é calculado da seguinte maneira:

$$IC = ((14 - 1).0,1 + (4 - 1).0,2 + (4 - 1).0,3 + (4 - 1).0,4) / (14 - 1) = 0,31$$

Nesse exemplo, como o IC é superior a 0,30, pode-se afirmar que houve colaboração “regional”. Porém, verificando a afiliação dos autores na Figura 5.2 é possível perceber que também houve colaboração “internacional”, pois os autores desse artigo são de 4 países distintos. Infelizmente, isso não pode ser afirmado baseando-se apenas no IC desse artigo.

O terceiro exemplo é um artigo publicado na “Lancet” por 19 autores em 1998, conforme ilustra a Figura 5.3.



Lancet

Volume 352, Issue 9124, 25 July 1998, Pages 257-263

Postoperative radiotherapy in non-small-cell lung cancer: Systematic review and meta-analysis of individual patient data from nine randomised controlled trials

Stewart, L.A.[ab], Burdett, S.[b], Parmar, M.K.B.[b], Souhami, R.L.[l], Arriagada, R.[c], Girling, D.J.[l], Pignon, J.P.[d], Torri, V.[e], Brichet, A.H.[f], Lafitte, J.J.[f], Dautzenberg, B.[l], Debevec, M.[g], Kovac, V.[g], Stephens, R.J.[l], Gregor, A.[h], Piantadosi, S.[l], Rocmans, P.[i], Van Houtte, P.[j], Wang, M.[k]

[a] MRC Cancer Trials Office, 5 Shaftesbury Road, Cambridge CB2 2BW, United Kingdom

[b] MRC Cancer Trials Office, Cambridge, United Kingdom

[c] Instituto de Radiomedicina, Santiago, Chile

[d] Institut Gustave-Roussy, Villejuif, France

[e] Ist. di Ric. Farmacologiche Mario N., Milan, Italy

[f] Hôpital Calmette, CHRU, Lille, France

[g] Institute of Oncology, Ljubljana, Slovenia

[h] Europ. Org. for Res./Treatm. of C., Brussels, Belgium

[i] Hôpitaux St. Pierre et Erasme, Brussels, Belgium

[j] Institut Jules Bordet, Brussels, Belgium

[k] Chinese Academy of Medical Sciences, Beijing, China

Figura 5.3 - Dados de um artigo indexado na Scopus para o cálculo do IC (Exemplo 3).

O IC para esse artigo é calculado da seguinte maneira:

$$IC = ((19 – 1).0,1 + (10 – 1).0,2 + (8 – 1).0,3 + (7 – 1).0,4) / (19 – 1) = 0,45$$

Nesse exemplo também houve colaboração “regional” que pode ser observada através do IC. Porém, também nesse artigo houve colaboração “internacional” que não é possível de ser detectada pelo IC.

O quarto exemplo é um artigo publicado no “European Journal of Operational Research” por 3 autores em 1978, conforme ilustra a Figura 5.4.

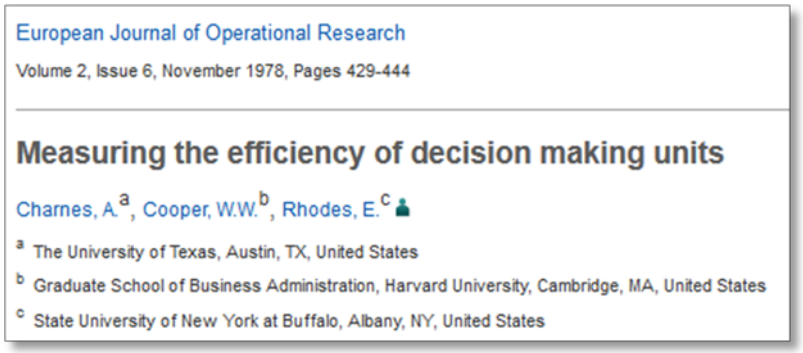


Figura 5.4 - Dados de um artigo indexado na Scopus para o cálculo do IC (Exemplo 4).

O IC para esse artigo é calculado da seguinte maneira:

$$IC = ((3 – 1).0,1 + (3 – 1).0,2 + (3 – 1).0,3 + (1 – 1).0,4) / (3 – 1) = 0,60$$

Nesse exemplo é possível afirmar que houve colaboração “regional” e tem-se um valor no limiar de colaboração “internacional”. Neste artigo especificamente não houve colaboração “internacional”.

O último exemplo, ilustrado na Figura 5.5, é um artigo publicado no periódico “Operations Research” por 3 autores em 2011.


Operations Research
Volume 59, Issue 3, May 2011, Pages 684-695

Structured replacement policies for components with complex degradation processes and dedicated sensors

Elwany, A.H.[a], Gebraeel, N.Z.[b], Maillart, L.M.[c]

[a] Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, 5600 MB Eindhoven, Netherlands
[b] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, United States
[c] Department of Industrial Engineering, University of Pittsburgh, Pittsburgh, PA 15260, United States

Figura 5.5 - Dados de um artigo indexado na Scopus para o cálculo do IC (Exemplo 5).

O IC para esse artigo é calculado da seguinte maneira:

$$IC = ((3 – 1).0,1 + (3 – 1).0,2 + (3 – 1).0,3 + (2 – 1).0,4) / (3 – 1) = 0,80$$

Como o IC desse artigo é superior a 0,60, pode-se afirmar com certeza que houve colaboração “internacional” na coautoria desse artigo.

5.6. Estudos de casos

Apresenta-se a seguir alguns estudos que ilustram a utilização do IC. Para isso, serão utilizados dados das publicações de 3 pesquisadores, de todos os artigos publicados em um determinado periódico, dos artigos mais citados de 2 países, e os artigos mais citados de uma determinada área. Todos os dados utilizados nos estudos foram obtidos a partir da base de dados Scopus.

O primeiro pesquisador analisado foi “Carlos José Pereira de Lucena”. Ele é pesquisador na Pontifícia Universidade Católica do Rio de Janeiro e possui 51 artigos publicados em periódicos indexados na Scopus entre 1976 e 2013. Esses artigos receberam 303 citações, com uma média de 5,94 citações por artigo. Porém, 14 (27,45%) desses artigos não foram citados. A média de autores dos artigos é 3,74 e o IC médio é 0,41. A Figura 5.6 apresenta a distribuição do IC desses artigos. É possível perceber que 9 (17,65%) artigos possuem IC igual a 1, indicando que houve colaboração “internacional” na coautoria desses artigos. O número de autores nesses artigos variou de 2 a 3 autores e a maioria (7) ocorreu até o ano 2001. Porém, a maioria dos artigos (15; 29,41%) possui IC igual a 0,1, o que indica apenas colaboração “local”.

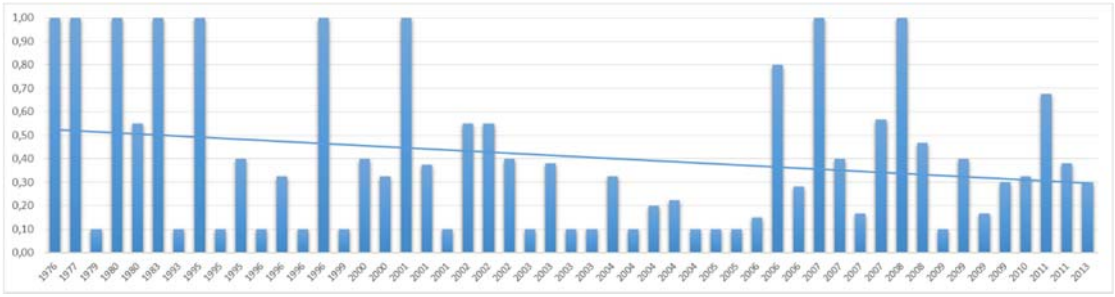


Figura 5.6 - Distribuição do IC dos artigos publicados em periódicos indexados na Scopus pelo pesquisador “Carlos José Pereira de Lucena”.

O IC médio dos artigos publicados até 2002 é 0,50. No período restante, o IC médio diminuiu para 0,33, conforme pode ser observado pela linha de tendência destacada na Figura 5.6. O número médio de citações dos 5 artigos com mais autores (6) é 14,6 e o IC médio é 0,31. Nos 6 artigos mais citados (entre 16 e 49 citações) o número médio de autores é 4,33 e o IC médio é 0,34. Nos artigos com IC igual a 1, o número médio de citações é 4,66. Percebe-se então que os artigos com mais autores são realmente os mais citados, sendo citados mais de 3 vezes mais que os artigos com IC igual a 1. Também é interessante observar que o número médio de autores é bem maior nos artigos mais citados.

O segundo pesquisador analisado foi “Miguel Afonso Sellitto”. Ele é pesquisador na Universidade Vale do Rio dos Sinos e possui 30 artigos publicados em periódicos indexados na Scopus entre 2006 e 2013. Esses artigos foram citados 58 vezes, com uma média de 1,93 citações e 3,43 autores por artigo. Desses artigos, 9 (30%) não foram citados. A distribuição do IC dos artigos desse pesquisador é apresentada na Figura 5.7. O IC médio é 0,18, o que indica colaboração institucional na coautoria desses artigos. Porém, a grande maioria dos artigos (18; 60%) possui IC igual a 0,1, indicando colaboração “local”. O IC médio até 2010 era 0,17 e passou para 0,20 nos anos seguintes. Isso pode ser percebido na linha de tendência destacada na Figura 5.7, que ilustra uma pequena variação no IC ao longo dos anos.

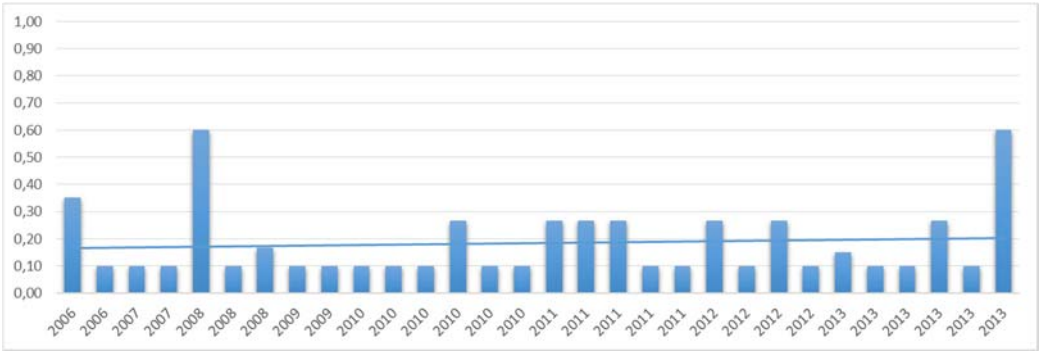


Figura 5.7 - Distribuição do IC dos artigos publicados em periódicos indexados na Scopus pelo pesquisador “Miguel Afonso Sellitto”.

O último pesquisador analisado foi “Alan Solon Ivor Zinober”. Ele é pesquisador na Universidade de Sheffield e possui 47 artigos publicados em periódicos indexados na Scopus entre 1984 e 2013. Esses artigos receberam 899 citações, com uma média de 19,13 citações e um número médio de autores de 2,89 por artigo. Desses artigos, 10 (21,28%) não foram citados e 2 foram publicados sem coautores. A distribuição do IC dos artigos desse pesquisador é apresentada na Figura 5.8. O IC médio é 0,42, o que indica colaboração “regional” na coautoria desses artigos.

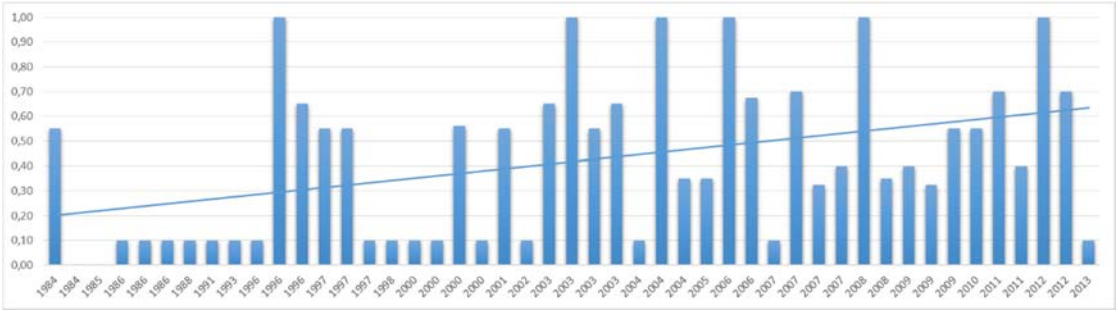


Figura 5.8 - Distribuição do IC dos artigos publicados em periódicos indexados na Scopus pelo pesquisador “Alan Solon Ivor Zinober”.

É possível perceber que 6 (12,77%) artigos possuem IC igual a 1, indicando que houve colaboração “internacional” na coautoria desses artigos. O número de autores nesses artigos variou de 2 a 4 autores e a maioria (5) ocorreu depois de 2003. Porém, a maioria dos artigos (16; 34,04%) também possui IC igual a 0,1, indicando apenas colaboração “local”.

O número médio de citações dos 5 artigos com mais autores (5 ou 6) é 3,4 e o IC médio é 0,52. Nos 5 artigos mais citados (entre 48 e 128 citações) o número médio de autores é 2,20 e o IC médio é 0,15. Nos artigos com IC igual a 1, o número médio de citações é 3,4. Nesse caso, pode-se perceber que os artigos com mais autores não foram os mais citados, tendo isso ocorrido com os artigos com IC igual a 1. É interessante observar que o IC médio entre os artigos com mais autores é bem maior que os artigos mais citados. Também é possível observar que houve uma grande variação no IC desses artigos. Até 2002 o IC era 0,26 e no período restante passou para 0,56. Isso pode ser observado na linha de tendência destacada na Figura 5.8.

A seguir analisamos o periódico “Journal of Informetrics”. De 2007 a 2012 foram publicados 290 artigos nesse periódico de acordo com a Scopus. Porém, nesse estudo foram considerados 283 artigos, pois 7 artigos não possuíam os dados necessários para o cálculo do IC. A Figura 5.9 apresenta a distribuição do IC desses artigos.

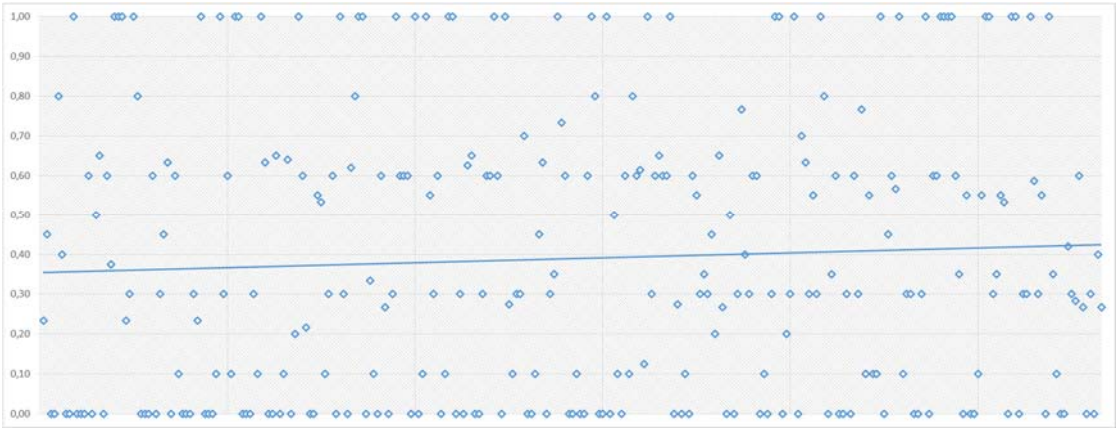


Figura 5.9 - Distribuição do IC dos artigos publicados no periódico “Journal of Informetrics” de acordo com dados da Scopus no período de 2007 a 2012.

Ao se analisar a tendência, observa-se que o termo independente do ajuste linear é dominante e muito próximo do IC médio, que nesse caso é 0,39. O número médio de autores desses artigos é 2,35 e o número médio de citações é 8,39 citações por artigo.

Na Tabela 5.2 é apresentada a distribuição do IC médio nesse periódico por ano. É possível perceber que o IC médio não variou muito de um ano para o outro. Também é destacado o número de artigos que possuem IC igual a 0, indicando que não houve colaboração. Do total de 283 artigos, a maioria (80; 28,27%) possui IC igual a 0 e cada um desses artigos foi citado 9 vezes em média. 22 artigos (7,77%) possuem IC igual a 0,10 (colaboração "local") e o número médio de citações foi 7,14. 43 artigos (15,19%) possuem IC igual a 1 (colaboração "internacional") e foram citados 8,86 vezes em média. A maioria dos artigos com IC igual a 1 foram publicados em 2012.

Tabela 5.2 - Distribuição do IC médio no periódico "Journal of Informetrics" por ano.

Ano	Artigos	IC	IC = 0	IC = 0,10	IC = 1
2007	30	0,38	12	0	5
2008	33	0,32	13	4	5
2009	32	0,38	9	3	5
2010	61	0,39	19	6	9
2011	59	0,41	12	3	6
2012	68	0,42	15	6	13
Total	283	0,39	80	22	43

Os 10 artigos mais citados (57,20 citações em média) desse periódico possuem IC igual a 0,29 e número médio de autores de 2,10 por artigo. Já os 10 artigos com mais autores (5 a 9 autores) possuem IC igual a 0,44 e o número médio de citações de 9,60 por artigo.

Analisando os 10 artigos com mais citações publicados em periódicos indexados na Scopus por autores com pelo menos um brasileiro, é possível perceber em quase todos os artigos (8) houve colaboração com outros países, como pode ser observado na Tabela 5.3. Desses artigos, em 6 deles (destacados em azul) é possível concluir isso verificando o IC.

Tabela 5.3 - Distribuição do IC dos artigos publicados em periódicos indexados na Scopus por autores do Brasil com mais citações.

n	Ano	Autores	Países	Citações	IC
1	2002	14	7	4.522	**0,75**
2	1988	9	2	4.354	0,24
3	1996	2	2	3.317	**1**
4	1988	1	1	3.166	0
5	2005	19	8	3.156	**0,64**
6	2002	12	8	2.980	**0,83**
7	2000	12	7	2.965	**0,79**
8	2008	24	9	2.727	**0,65**
9	1995	3	2	2.691	0,55
10	2002	3	1	2.661	0,10

Fazendo a mesma análise, só que considerando artigos publicados por autores com pelo menos um deles dos Estados Unidos, foi constatado que em todos os 10 artigos só há autores desse país. Isso ficou evidenciado com o cálculo do IC desses artigos, conforme apresentado na Tabela 5.4.

Tabela 5.4 - Distribuição do IC dos artigos publicados em periódicos indexados na Scopus por autores do Estados Unidos com mais citações.

n	Ano	Autores	Citações	IC
1	1976	1	107.071	0
2	1987	2	44.809	0,60
3	1975	3	35.988	0,35
4	1990	5	35.761	0,10
5	1996	3	32.496	0,35
6	2001	2	27.681	0,60
7	1983	1	21.436	0
8	1965	1	20.640	0
9	1977	2	16.458	0,10
10	1976	2	15.545	0,10

Também é possível analisar os artigos mais citados em uma determinada área. No caso da Matemática, percebe-se com o cálculo do IC que a maioria (7) das colaborações são locais, mesmo quando o número de autores é alto.

Tabela 5.5 - Distribuição do IC dos artigos publicados em periódicos indexados na Scopus na área de Matemática com mais citações.

n	Ano	Autores	Citações	IC
1	1994	3	39.774	0,55
2	1977	2	16.458	0,10
3	1999	2	12.597	0,10
4	1988	3	11.278	0,10
5	1951	1	9.246	0
6	1995	11	7.821	0,10
7	1995	2	7.696	0,10
8	1986	1	7.556	0,10
9	1998	5	7.540	0,10
10	1990	4	7.273	0,10

5.7. Vantagens e limitações

Uma das vantagens do IC é que ele é simples de calcular. Os dados para realizar o cálculo também podem ser obtidos de maneira simples, inclusive no próprio artigo. Ao contrário de outros índices, o IC é calculado para um único artigo, não sendo necessário que um conjunto de artigos seja considerado para realizar o seu cálculo. Para calcular o IC também não é necessário qualquer dado que seja contabilizado por alguma base de dados, como por exemplo, o número de citações de um artigo.

Uma limitação do índice proposto é que ele não dá indicativos sobre a quantidade de autores. Por exemplo, o IC pode ser igual a 1 quando houve a colaboração entre 10 países ou 2 países apenas. Isso também ocorre quando todos os autores são de uma mesma instituição. Entretanto, em ambos os casos o IC estabelece com precisão o tipo de colaboração que ocorreu, internacional ou local.

Artigos com muitos autores tornam o cálculo do IC trabalhoso quando realizado manualmente, pois imaginem, por exemplo, contar o número de autores, instituições, cidades e países de um artigo com mais de 1.000 autores.

Infelizmente, nem sempre os dados informados em artigo a respeito da afiliação dos autores são realmente corretos. Um pesquisador pode estar passando um tempo em outra instituição, de um outro país, e publicar um artigo com pesquisadores dessa instituição e ao informar a afiliação dos autores seja colocado apenas a instituição em que ele está no momento. Assim, uma colaboração internacional que realmente ocorreu não pode ser identificada.

5.8. Considerações finais

Conforme mencionado anteriormente, há diversos motivos que levam os pesquisadores a colaborarem. Porém, não é de nosso conhecimento a existência de algum indicador que meça como essa colaboração ocorre.

O IC é uma tentativa de medir a colaboração entre os autores de um artigo. Com esse índice é possível observar se a produção científica de um pesquisador é local, institucional, regional ou internacional, como mostrado com vários exemplos ilustrativos.

No próximo Capítulo são apresentados alguns estudos de casos, ilustrando como as ferramentas desenvolvidas podem ser utilizadas para realizar análises nas bases de dados consideradas neste trabalho.

6 ESTUDOS DE CASOS

Nos capítulos anteriores foram apresentadas todas as ferramentas desenvolvidas ao longo deste trabalho, bem como as bases de dados utilizadas. Neste Capítulo são apresentados alguns estudos de casos com análises realizadas.

Com as ferramentas desenvolvidas e com as bases de dados utilizadas, é possível analisar as informações técnico-científicas de um único pesquisador; grupo de pesquisadores; curso ou programa; uma ou mais instituições; áreas, periódicos; estados ou regiões; países e até mesmo toda a produção científica mundial indexada nas principais bases de dados do saber. Sendo assim, os estudos de casos aqui apresentados demonstram parte do potencial que pode ser explorado por demais pesquisadores.

Na Tabela 6.1 é apresentado um resumo dos estudos de casos deste Capítulo, destacando as análises realizadas e as bases de dados utilizadas, além de exemplos de questões que podem ser respondidas. Esse resumo serve como indicativo do que é necessário para que estudos similares sejam realizados e pode se constituir como base para novos estudos.

O mapas apresentados neste estudo foram gerados utilizando a ferramenta *GPS Visualizer* (acessível em http://www.gpsvisualizer.com) e as figuras apresentadas na forma de “nuvens de palavras” (*word clouds*) foram geradas utilizando a ferramenta *Wordle*™ (acessível em http://www.wordle.net/). Para gerar uma “nuvem de palavras” a ferramenta leva em consideração o número de vezes que cada palavra aparece. Essas duas ferramentas são de uso gratuito. O acesso ao fator de impacto (FI) de 2011 no JCR® também foi feito através do Portal de Periódicos da CAPES.

Tabela 6.1 - Resumo dos estudos de casos.

n	Tópico	Análise	Bases de dados	Questões
1	Instituição	Instituto Nacional de Pesquisas Espaciais	WoS e JCR®	Qual o perfil de uma instituição?
2	Grupo de pesquisadores	Pesquisa Operacional	PL, Bolsas em curso, Qualis Periódico da CAPES e Cursos de Pós-Graduação recomendados pela CAPES	Qual a contribuição de um grupo de pesquisadores para uma área do conhecimento?
3	Área	Pesquisa Operacional	WoS e JCR®	Qual o perfil de uma área?
4	Grande área	Química	PL, Bolsas em curso, Qualis Periódico da CAPES e JCR®	Qual o perfil de uma grande área do conhecimento?
5	Bases de dados	JCR® e Scopus	JCR® e Scopus	Como se dá a comparação de indicadores de bases distintas?
6	Periódico	Journal of Informetrics	JCR® e Scopus	Qual a trajetória de um periódico?
7	Área de atuação	Engenharia de Software	PL, WoS e JCR®	É possível definir qual é realmente a área de atuação de um pesquisador?

6.1. Instituição

Este estudo de caso tem como objetivo mapear a produção científica de uma instituição considerando as informações sobre os artigos publicados nos periódicos indexados na WoS. O mapeamento de uma instituição permite conhecer parte de sua história, assim como pode indicar o perfil de colaboração científica dos pesquisadores dessa instituição, dentre várias outras coisas.

O objeto desta Seção é o Instituto Nacional de Pesquisas Espaciais (INPE) que foi criado em 1961 e que tem reconhecimento internacional da comunidade científica. Uma das razões advém de sua produção científica qualificada. Alguns estudos já foram realizados considerando somente a pós-graduação do INPE (MOREIRA; VELHO, 2009; MOREIRA, 2009; MOREIRA; VELHO, 2010; MOREIRA; VELHO, 2012), porém, não se tem conhecimento de nenhum

estudo que tenha mapeado a totalidade da produção científica do INPE na WoS.

6.1.1. Coleta de dados

Inicialmente, foi realizada uma busca na base de dados WoS por artigos de autores com vinculação ao INPE e publicados em periódicos, de acordo com os critérios informados na Figura 6.1. O número de artigos encontrados foi 5.883. É importante destacar que a busca foi feita utilizando o rótulo “OG” (Organização Consolidada) ao invés do rótulo “OO” (Organização). A opção adotada é mais abrangente, pois os artigos foram buscados pelo nome preferencial da instituição e por suas variantes que foram identificadas e associadas a ela pela própria WoS. Porém, nem todas as organizações foram incluídas nessa lista. Evidentemente, o resultado da busca pode ser afetado pelo fato de se utilizar um ou outro rótulo.

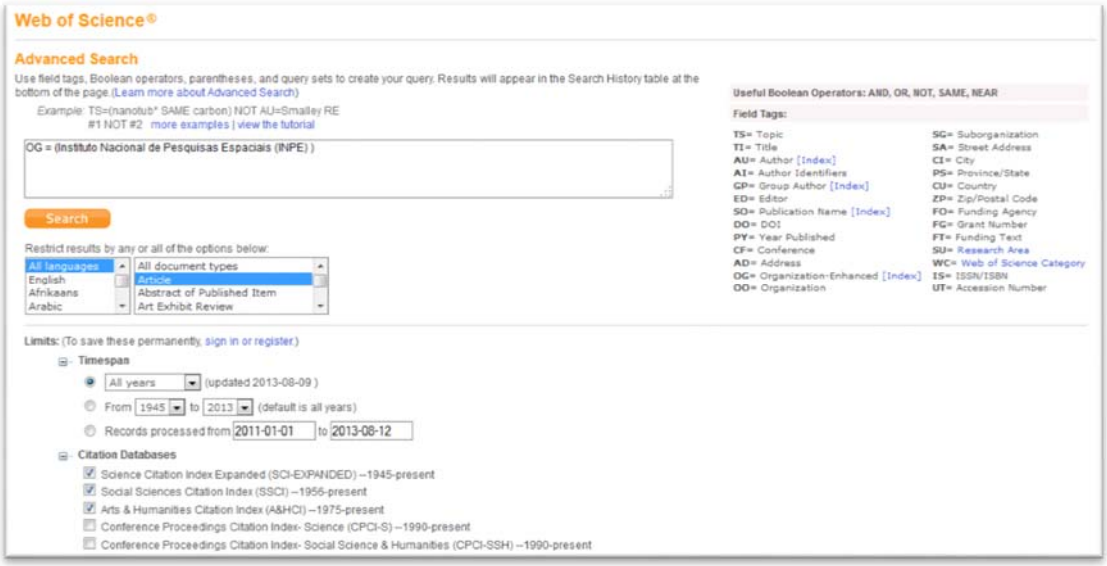


Figura 6.1 - Definição de critérios de busca na WoS por artigos de autores com vinculação ao INPE e publicados em periódicos.

A partir disso, foi possível baixar as informações desses artigos no formato “bibtex”, um dos formatos disponíveis na base WoS. Dessa forma, foi feita a extração automática e as informações foram armazenadas, também de forma

automática, em um banco de dados. Todo esse processo, brevemente aqui descrito, foi realizado no dia 12 de agosto de 2013.

6.1.2. Resultados e discussões

Dos 5.883 artigos publicados por pesquisadores do INPE em periódicos indexados na WoS, 1.049 (17,83%) ainda não foram citados sendo que quase metade desses artigos (474; 45,19%) foram publicados nos últimos três anos. Os outros 4.834 artigos (82,17%) foram citados por 77.385 artigos, com uma média de 13,15 citações por artigo. A Figura 6.2 apresenta o número de artigos publicados por autores com vinculação ao INPE (o qual passaremos a nos referir genericamente como pesquisadores) em periódicos indexados na WoS, desde 1968 até o dia 12 de agosto de 2013, destacando-se os artigos que ainda não foram citados.

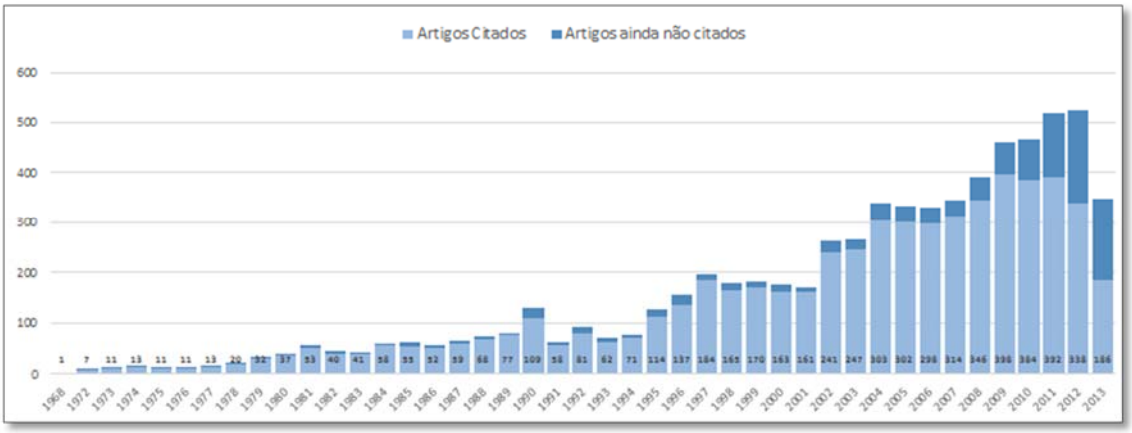


Figura 6.2 - Número de artigos publicados por pesquisadores do INPE em periódicos indexados na base de dados WoS.

Para publicar esses 5.883 artigos, os pesquisadores do INPE colaboraram com autores de outros 112 países. A Figura 6.3 apresenta a distribuição geográfica dos coautores que publicaram artigos com pesquisadores do INPE em periódicos indexados na base WoS. Para gerar esse mapa, foram considerados somente os 46 (41,07%) países que os pesquisadores do INPE colaboraram pelo menos 10 vezes ao longo dos anos considerados. Os 5 países que contaram com maior colaboração foram: Estados Unidos (1.185;

20,14%), Inglaterra (299; 5,08%), Alemanha (252; 4,28%), França (222; 3,77%) e Japão (212; 3,60%).

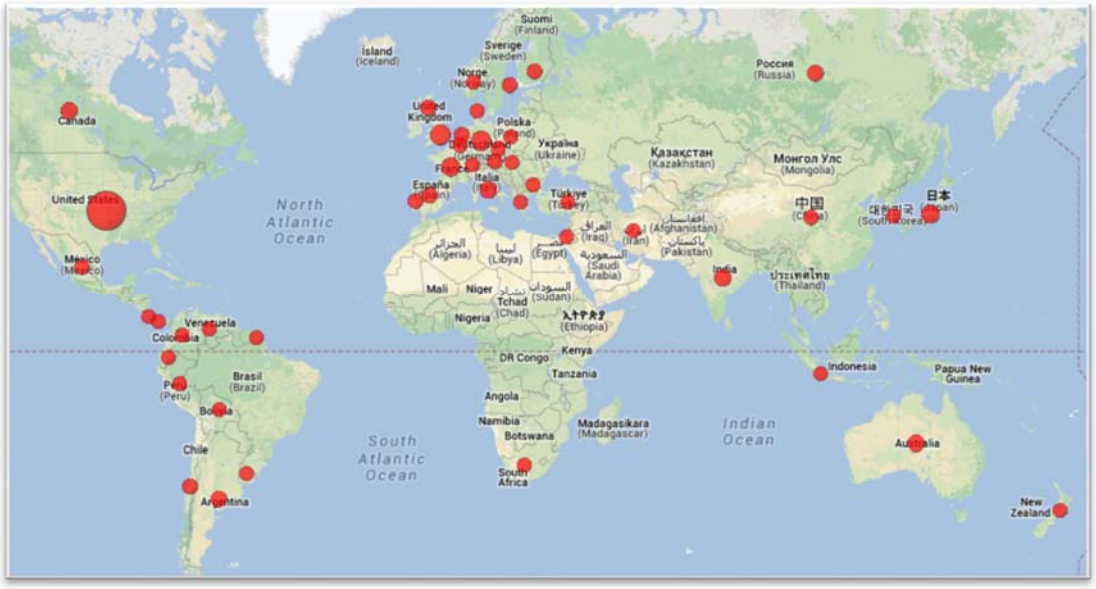


Figura 6.3 - Distribuição geográfica dos coautores que publicaram artigos com pesquisadores do INPE em periódicos indexados na base de dados WoS.

A Tabela 6.2 apresenta a distribuição dos 10 periódicos mais utilizados para publicação desses 5.883 artigos de acordo com a WoS. Desse total, somente em 65 artigos o ISSN não foi identificado. Para os demais (5.117; 87,95%), foi obtido o FI dos periódicos de acordo com o JCR® de 2011. Ao todo foram utilizados 889 periódicos diferentes com FI. Também é apresentada a média de citações por artigos publicados em cada um dos periódicos, sendo possível perceber que as médias variam bastante e que nem sempre, os periódicos mais utilizados são os mais citados. Nas citações também é apresentado entre parênteses o número de artigos que foram citados.

A Figura 6.4 apresenta as 30 palavras-chave mais utilizadas nos artigos publicados por pesquisadores do INPE em periódicos indexados na WoS. As palavras-chave foram informadas somente em 3.127 (53,15%) dos artigos, o que ocorreu a partir de 1990. Nesses artigos, foram encontradas 8.329 palavras-chave distintas, sendo que as 5 mais utilizadas foram: “Brazil” (150; 4,80%), “Amazônia” (130; 4,16%), “Amazon” (125; 4,00%), “Remote Sensing”

(97; 3,10%) e “Taxonomy” (88; 2,81%). Vale destacar que na WoS não é possível obter diretamente as palavras-chave mais utilizadas em um conjunto de artigos, somente as palavras-chave de um único artigo por consulta. Observe-se que a contagem manual para a obtenção deste resultado demandaria um tempo muito grande.

Tabela 6.2 - Distribuição dos periódicos indexados no JCR® de 2011 mais utilizados para publicação pelos pesquisadores do INPE de acordo com a base de dados WoS.

ISSN	Periódico	Artigos (A)	Citações (C)	C / A	FI 2011
0273-1177	ADVANCES IN SPACE RESEARCH	220	956 (160)	4,35	1,178
1364-6826	JOURNAL OF ATMOSPHERIC AND SOLAR-TERRESTRIAL PHYSICS	156	1.496 (133)	9,59	1,596
0148-0227	JOURNAL OF GEOPHYSICAL RESEARCH	154	3.649 (142)	23,69	3,021
0094-8276	GEOPHYSICAL RESEARCH LETTERS	124	2.812 (120)	22,68	3,792
0143-1161	INTERNATIONAL JOURNAL OF REMOTE SENSING	118	1.797 (112)	15,23	1,117
0004-637X	ASTROPHYSICAL JOURNAL	96	2.169 (93)	22,59	6,024
0992-7689	ANNALES GEOPHYSICAE	96	922 (85)	9,60	1,842
0004-6361	ASTRONOMY & ASTROPHYSICS	90	961 (84)	10,68	4,587
1175-5326	ZOOTAXA	77	152 (52)	1,97	0,927
0100-204X	PESQUISA AGROPECUARIA BRASILEIRA	76	228 (57)	3,00	0,756

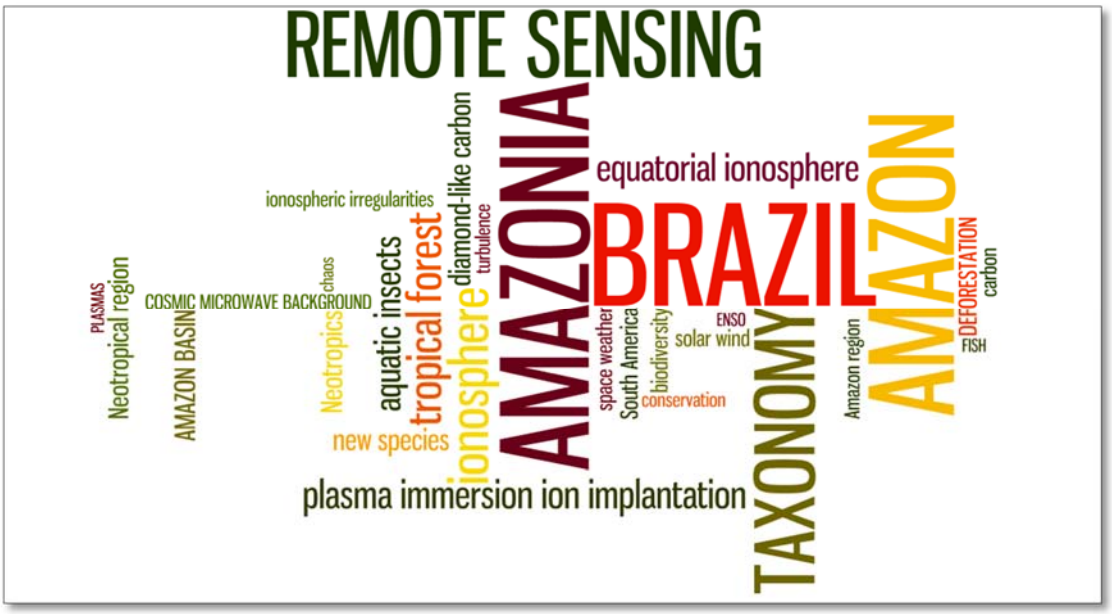


Figura 6.4 - Palavras-chave mais utilizadas por pesquisadores do INPE em artigos publicados em periódicos indexados na base de dados WoS.

Pode-se perceber que análises bem interessantes podem ser realizadas mesmo não utilizando todas as informações obtidas a partir dos registros dos artigos considerados neste estudo. Por exemplo, foi possível verificar que a grande maioria dos artigos publicados por pesquisadores do INPE já foi citado pelo menos uma vez; os pesquisadores do INPE já colaboraram com pesquisadores de mais de 100 países diferentes, o que contribui para aumentar a visibilidade e o reconhecimento do INPE na comunidade científica internacional.

6.2. Grupo de pesquisadores

Neste estudo de caso o objetivo é analisar o perfil dos bolsistas de Produtividade em Pesquisa (PQ) das áreas de Engenharia de Produção (EP) e de Engenharia de Transportes (ET) na subárea de Pesquisa Operacional (PO), utilizando informações extraídas dos seus currículos Lattes. Conhecer o perfil dos pesquisadores bolsistas PQ do CNPq das diversas áreas de conhecimento é de interesse para gestores de Ciência e Tecnologia, que passam a ter um melhor conhecimento do grupo de pesquisadores de cada uma destas áreas, serve como suporte para um melhor planejamento científico, identificação de regiões carentes, avaliar a maturidade de grupos e da área, mudanças decorrentes de políticas adotadas por agências de fomento, entre outros. A comunidade científica também se interessa em saber qual o perfil dos seus pares, além de ter indicadores quantitativos da produção científica e tecnológica qualificada dos que pertencem a este seleto grupo de bolsistas PQ do CNPq.

6.2.1. Coleta de dados

Neste estudo foi utilizada a relação de bolsistas PQ com bolsas ativas em 20 de abril de 2011, disponíveis no sítio do CNPq das áreas de EP e ET. Os bolsistas considerados foram apenas os listados com indicação de estarem em “Em folha de Pagamento”; os demais, por exemplo, com bolsas suspensas não foram considerados. O número de bolsistas encontrados nesta condição foram 137 pesquisadores da área de EP e 49 pesquisadores da área de ET. A partir

da identificação dos bolsistas, foram obtidos os currículos Lattes destes 186 pesquisadores. Para realizar essa tarefa, foi usada a linguagem “LattesMiner”.

Inicialmente, foram criados dois arquivos texto contendo o nome dos pesquisadores, conforme constava na relação de bolsistas PQ (Bolsas em Curso) das áreas de EP e ET, respectivamente. Apesar de ser possível informar todos os nomes em um mesmo arquivo, essa divisão permite diferenciar os pesquisadores das duas áreas. Em seguida, a linguagem LattesMiner identificou automaticamente o número (ID) de cada um dos pesquisadores, permitindo assim que os seus currículos fossem baixados e armazenados como arquivos HTML. Dessa forma, foi feita a extração automática dos dados, sendo os mesmos armazenados, também de forma automática, em um banco de dados. Todo esse processo, brevemente aqui descrito, foi realizado no dia 20 de abril de 2011 em menos de 1 hora.

Utilizando a linguagem LattesMiner, foram encontrados 4 homônimos de pesquisadores dentre os 137 nomes dos bolsistas de EP durante a etapa de busca por nome realizada pela linguagem LattesMiner. Nesse caso, é retornado o número (ID) de todos os homônimos identificados, sendo necessário que o usuário verifique o número (ID) correto referente ao pesquisador procurado. A seguir são apresentados os bolsistas PQ de EP que possuem homônimos e o respectivo número de homônimos identificados: Carlos Silva Oliveira (2), Edson Pinheiro de Lima (2), Paulo Henrique Siqueira (2) e Ricardo José Rabelo (2). Dentre os 49 bolsistas de ET, também foram encontrados 4 homônimos: João Carlos Souza (2), Paulo Cesar Marques da Silva (2), Renato da Silva Lima (2) e Yaeko Yamashita (2).

As seguintes informações foram extraídas dos currículos Lattes dos pesquisadores pela linguagem LattesMiner e armazenadas num banco de dados: dados pessoais, endereço profissional, formação acadêmica, participação em bancas examinadoras, produção bibliográfica em periódicos e congressos, orientações e áreas de atuação. A linguagem LattesMiner permite extrair outras informações, mas para este estudo apenas essas foram consideradas.

Para realizar este estudo os seguintes atributos foram considerados: gênero, categoria do bolsista, instituição de origem, tempo de conclusão do doutorado, distribuição geográfica, artigos completos publicados em periódicos, trabalhos completos publicados em anais de congressos, orientações concluídas de mestrado e doutorado, participação em bancas examinadoras de mestrado e doutorado, áreas de atuação, periódicos utilizados para publicação, classificação de periódicos segundo o Qualis Periódicos da CAPES em 2011 (QUALIS, 2013), cursos de Pós-Graduação recomendados e reconhecidos pela CAPES (CAPES, 2013), população segundo o censo de 2010 (IBGE, 2013) e o número de autores por artigo publicados em periódicos.

6.2.2. Resultados e discussões

Dos 186 bolsistas PQ das áreas de EP (137) e ET (49), 185 informaram sua(s) área(s) de atuação nos seus currículos Lattes. Desse total, 176 (95,14%) se declaram como atuantes na Grande Área de "Engenharias", conforme destacado na Tabela 6.3.

Tabela 6.3 - Distribuição dos bolsistas PQ das áreas de EP e ET segundo a Grande Área de atuação.

Grande Área	n	%
Ciências Biológicas	2	1,08
Ciências da Saúde	3	1,62
Ciências Exatas e da Terra	70	37,84
Ciências Humanas	8	4,32
Ciências Sociais Aplicadas	23	12,43
Engenharias	**176**	**95,14**
Outros	3	1,62

Fonte: Currículo Lattes dos bolsistas PQ (n = 185)

Na Tabela 6.4 é apresentado o número de bolsistas em cada uma das categorias. Seria natural esperar um decréscimo de bolsistas na medida em que se move da categoria 2 até 1A, mas isto não se observa. Estes valores podem ser reflexos da política de expansão do número de bolsas PQ pelo CNPq que não tem crescido de maneira gradual.

Tabela 6.4 - Distribuição dos bolsistas PQ das áreas de EP e ET segundo categoria.

Categoria	EP		ET		EP/ET	
	n	%	n	%	n	%
2F	3	75,00	1	25,00	4	2,15
2	85	75,89	27	24,11	112	60,22
1D	27	81,82	6	18,18	33	17,74
1C	5	62,50	3	37,50	8	4,30
1B	8	53,33	7	46,67	15	8,06
1A	9	64,29	5	35,71	14	7,53
Total	137	73,66	49	26,34	186	100

Fonte: Currículo Lattes dos bolsistas PQ (n = 186)

Dos 137 bolsistas de EP, 85 (62,04%) se declaram como sendo da Área de "Engenharia de Produção" e da Subárea de "Pesquisa Operacional" e dos 49 da ET, apenas 6 (12,24%) se declararam da mesma forma, totalizando 91 bolsistas da subárea de PO. No caso da EP, a subárea de PO é a subárea de atuação que mais bolsistas dizem atuar, enquanto na ET é apenas a quinta, como pode ser observado na Tabela 6.5. Na ET a subárea de atuação que mais bolsistas atuam é a de "Planejamento de Transportes".

Tabela 6.5 - Distribuição dos bolsistas PQ das áreas de EP e ET segundo a subárea de atuação.

	Grande Área	Área	Subárea	Total
EP	Engenharias	Engenharia de Produção	**Pesquisa Operacional**	**85**
	Engenharias	Engenharia de Produção	Gerência de Produção	51
	Engenharias	Engenharia de Produção	Não informada	25
	Ciências Exatas e da Terra	Matemática	Matemática Aplicada	22
	Ciências Exatas e da Terra	Ciência da Computação	Teoria da Computação	18
ET	Engenharias	Engenharia de Transportes	Planejamento de Transportes	38
	Engenharias	Engenharia de Transportes	Operações de Transportes	28
	Engenharias	Não informada	Não informada	14
	Engenharias	Engenharia Civil	Infra-Estrutura de Transportes	6
	Engenharias	Engenharia de Produção	**Pesquisa Operacional**	**6**

Fonte: Currículo Lattes dos bolsistas PQ (n = 185)

A seguir apenas os bolsistas que indicaram atuar na subárea de PO são considerados. Conforme já mencionado, 91 (48,92%) dos bolsistas se declaram como sendo da subárea de PO. Cabe alertar que desse total, 74 (81,32%) atualizaram o currículo Lattes no ano de 2011 e 82 (90,11%) atualizaram no período máximo de 6 meses. Portanto, nem todas as informações estão atualizadas o que pode ocasionar ligeiras variações nas observações feitas a seguir. A distribuição desses bolsistas segundo cada categoria é apresentada na Tabela 6.6. A maioria deles (75,83%) se concentra nas categorias 2 e 1D, muito semelhante a porcentagem referente aos 186 (77,96%) e igual em relação às categorias. O tempo médio da titulação (ano de conclusão do doutorado) é de 15,19 anos. É interessante notar que o tempo médio da categoria 1C (16,25 anos) é menor que o da categoria 1D (17,82 anos). Cabe ressaltar que nos casos em que o bolsista possui mais de um título de doutorado, foi considerada a data da primeira titulação. As categorias 1B e 2 têm um bolsista cada com mais de um título de doutorado.

Tabela 6.6 - Distribuição dos bolsistas PQ da subárea de PO segundo categoria.

Categoria	n	%	Tempo médio (anos)
2F	2	2,20	6,50
2	47	51,65	11,30
1D	22	24,18	17,82
1C	4	4,39	16,25
1B	7	7,69	19,86
1A	9	9,89	26,89
Total	91	100	15,19

Fonte: Currículo Lattes dos bolsistas PQ de PO (n = 91)

Desses 91 bolsistas, 65 são do gênero masculino e 26 do feminino, conforme ilustra a Tabela 6.7. Houve equilíbrio nas categorias 2F e 1C; sendo que nas outras categorias houve predomínio do gênero masculino, principalmente na categoria 1A, que não possuía à época nenhuma bolsista do gênero feminino.

Tabela 6.7 - Distribuição dos bolsistas PQ da subárea de PO por gênero segundo categoria.

Categoria	Masculino		Feminino	
	n	%	n	%
2F	1	50,00	1	50,00
2	33	70,21	14	29,79
1D	15	68,18	7	31,82
1C	2	50,00	2	50,00
1B	5	71,43	2	28,57
1A	9	100	0	0
Total	65	71,43	26	28,57

Fonte: Currículo Lattes dos bolsistas PQ de PO (n = 91)

A Tabela 6.8 apresenta a distribuição geográfica dos bolsistas que indicam atuar na subárea de PO, e que mostra que o Sudeste é a região do Brasil com o maior número de bolsistas, tendo o estado de São Paulo como principal destaque. No entanto, analisando o número de bolsistas por milhão de habitantes, o destaque é o estado do Rio de Janeiro. Na região Sudeste, apenas o estado do Espírito Santo está abaixo do nível nacional de bolsistas por milhão de habitantes.

Foi levantado da CAPES o número de cursos de Pós-Graduação (PG) em Engenharia de Produção e de Transportes, que estão em áreas distintas, respectivamente Engenharias III e Engenharias I, e observou-se que os estados de São Paulo e Rio de Janeiro possuem a grande maioria. O estado de Minas Gerais, no entanto, é o que se destaca dos demais pelo número de bolsistas por cursos de PG no estado, uma vez que possui 3,2 bolsistas por cursos de PG, enquanto o estado de São Paulo possui 1,52 e o Rio de Janeiro apenas 1,16. O estado do Espírito Santo possui um bolsista, porém não tem nenhum curso de PG. Na região Nordeste, o estado que mais se destaca é o de Pernambuco que possui 3 bolsistas por cursos de PG. Este estado também se destaca em termos de número de bolsistas por milhão de habitantes em relação a estados da região Sudeste como Minas Gerais, que possui 0.82 bolsistas por milhão de habitantes e São Paulo que possui 0,78 bolsistas por

milhão de habitantes. Além de Pernambuco, o estado do Rio Grande do Norte também está acima do nível nacional de bolsistas por milhão de habitantes. O estado do Ceará, apesar de possuir 3 cursos de PG, possui apenas um bolsista na subárea de PO. O estado da Bahia com 3 cursos de PG e o estado da Paraíba com 1, não possuem bolsistas de PO. Na região Sul, o estado que se destaca é o Paraná, porém está abaixo do nível nacional de bolsistas por milhão de habitantes. Na região Norte há apenas um curso de PG no estado do Amazonas e nenhum bolsista. Na região Centro-Oeste somente o Distrito Federal possui 1 bolsista.

Tabela 6.8 - Distribuição geográfica dos bolsistas PQ da subárea de PO.

Região	UF	PQ	%	População[1]	%	PQ/milhão	Cursos PG[2]		
							M	D	F
Norte		0	0	15.864.454	8,32	0	0	0	1
	AM	0	0	3.483.985	1,83	0	0	0	1
Nordeste		12	13,19	53.081.950	27,83	0,23	6	3	2
	PE	9	9,89	8.796.448	4,61	1,02	1	1	1
	RN	2	2,20	3.168.027	1,66	0,63	1	0	0
	CE	1	1,10	8.452.381	4,43	0,12	2	1	0
	BA	0	0	14.016.906	7,35	0	1	1	1
	PB	0	0	3.766.528	1,97	0	1	0	0
Sudeste		71	78,02	80.364.410	42,13	0,88	25	15	5
	SP	32	35,16	41.262.199	21,63	0,78	12	8	1
	RJ	22	24,18	15.989.929	8,38	1,38	10	5	4
	MG	16	17,58	19.597.330	10,27	0,82	3	2	0
	ES	1	1,10	3.514.952	1,84	0,28	0	0	0
Sul		7	7,69	27.386.891	14,36	0,26	9	3	2
	PR	4	4,39	10.444.526	5,48	0,38	3	1	0
	RS	3	3,30	10.693.929	5,61	0,28	4	1	1
	SC	0	0	6.248.436	3,28	0	2	1	1
Centro-Oeste		1	1,10	14.058.094	7,37	0,07	2	1	0
	DF	1	1,10	2.570.160	1,35	0,39	1	1	0
	GO	0	0	6.003.788	3,15	0	1	0	0
Brasil		**91**	**100**	**190.755.799**	**100**	**0,48**	**42**	**22**	**10**

Fontes: Currículo Lattes dos bolsistas PQ de PO (n = 91)

[1]IBGE (censo 2010)

[2]CAPES (Cursos recomendados e reconhecidos)

Legenda: M - Mestrado Acadêmico, D - Doutorado e F - Mestrado Profissional

É importante ressaltar que os dados apresentados na Tabela 6.8, referem-se apenas aos bolsistas que atuam na subárea de PO. Considerando os 186 bolsistas PQ das áreas de EP e ET, a distribuição geográfica é a seguinte: a região Norte não possui bolsistas; a região Nordeste possui 19 bolsistas (12 em Pernambuco, 3 no Rio Grande do Norte, 3 no Ceará e 1 na Bahia); a região Sudeste possui 135 bolsistas (62 em São Paulo, 51 no Rio de Janeiro, 20 em Minas Gerais e 2 no Espírito Santo); a região Sul possui 25 bolsistas (7 no Paraná, 10 no Rio Grande do Sul e 8 em Santa Catarina) e a região Centro-Oeste possui 7 bolsistas, todos do Distrito Federal.

A Tabela 6.9 apresenta o ranking das instituições com maior número de bolsistas PQ atuando na subárea de PO (foram listadas apenas as instituições com mais de um bolsista). É importante ressaltar que para elaborar esse ranking foi considerada a cidade em que a instituição se localiza, fornecida no currículo Lattes. Caso não fosse levada em consideração o ranking seria diferente. Por exemplo, a UFPE (Universidade Federal de Pernambuco) possui ao todo 9 bolsistas atuando em PO, porém 8 são de Recife e 1 de Caruaru. A USP (Universidade de São Paulo) também possui 9 bolsistas, porém 5 são da cidade de São Paulo e 4 da cidade de São Carlos. No caso de empate no número de bolsistas, foi considerada a categoria dos bolsistas.

Na Tabela 6.9 também é possível analisar as instituições segundo a categoria dos bolsistas. Nesse caso, a PUC-Rio (Pontifícia Universidade Católica do Rio de Janeiro) e o INPE (Instituto Nacional de Pesquisas Espaciais) se destacam, pois cada uma dessas instituições possui 2 bolsistas na categoria 1A.

Tabela 6.9 - Ranking das instituições com maior número de bolsistas PQ na subárea de PO.

Posição	Instituição	Cidade	UF	PQ	Categorias					
					1A	1B	1C	1D	2	2F
1ª	UFMG	Belo Horizonte	MG	9	-	-	1	2	6	-
2ª	UFPE	Recife	PE	8	1	-	-	2	5	-
3ª	PUC-Rio	Rio de Janeiro	RJ	5	2	-	-	2	1	-
4ª	UNICAMP	Campinas	SP	5	1	-	1	2	1	-
5ª	UFSCar	São Carlos	SP	5	1	-	-	1	3	-
6ª	USP	São Paulo	SP	5	-	2	-	2	1	-
	UFRJ	Rio de Janeiro	RJ	5	-	2	-	2	1	-
7ª	UFF	Niterói	RJ	5	-	-	-	2	3	-
8ª	USP	São Carlos	SP	4	1	-	-	1	2	-
9ª	UFRGS	Porto Alegre	RS	3	-	1	1	1	-	-
10ª	INPE	São José dos Campos	SP	2	2	-	-	-	-	-
11ª	ITA	São José dos Campos	SP	2	-	1	-	1	-	-
12ª	PUC Minas	Belo Horizonte	MG	2	-	1	-	-	1	-
13ª	UFPR	Curitiba	PR	2	-	-	1	-	1	-
14ª	UNESP	Guaratinguetá	SP	2	-	-	-	1	1	-
	UTFPR	Curitiba	PR	2	-	-	-	1	1	-
15ª	UFRN	Natal	RN	2	-	-	-	-	2	-
	UNIFEI	Itajubá	MG	2	-	-	-	-	2	-
	UFABC	Santo André	SP	2	-	-	-	-	2	-
	UNESP	São José do Rio Preto	SP	2	-	-	-	-	2	-
16ª	UFF	Volta Redonda	RJ	2	-	-	-	-	1	1
TOTAL				**76**	8	7	4	20	36	1

Fonte: Currículo Lattes dos bolsistas PQ de PO (n = 91)

A atuação acadêmica dos bolsistas das áreas de EP e ET é apresentada na Tabela 6.10, destacando os bolsistas que atuam na subárea de PO. No caso dos periódicos foram considerados apenas os artigos publicados e no caso dos congressos, apenas os trabalhos completos publicados em anais. Em relação às orientações foram consideradas apenas as concluídas, em nível de mestrado e doutorado. No caso das bancas, foram consideradas apenas as participações em dissertações e teses.

Tabela 6.10 - Atuação Acadêmica dos bolsistas PQ das áreas de EP e ET.

Tópicos	EP/ET	EP		ET		PO	
	n	n	%	n	%	n	%
Periódicos	5.176	4.131	79,81	1.045	20,19	2.672	51,62
Periódicos (2001-2010)	3.496	2.991	85,55	505	14,45	1.893	54,15
Congressos	11.687	8.072	69,07	3.615	30,93	4.882	41,77
Congressos (2001-2010)	8.134	5.850	71,92	2.284	28,08	3.488	42,88
Orientações	4.474	3.113	69,58	1.361	30,42	2.016	45,06
Orientações (2001-2010)	3.127	2.205	70,51	922	29,49	1.402	44,84
Bancas	7.040	4.945	70,24	2.095	29,76	3.015	42,83
Bancas (2001-2010)	5.526	3.842	69,53	1.684	30,47	2.282	41,30

Fonte: Currículo Lattes dos bolsistas PQ (n = 186) e de PO (n = 91)

Analisando a Tabela 6.10 é possível perceber um número maior de publicações em congressos em comparação a periódicos, tanto na área de EP quanto na área de ET, podendo o mesmo ser observado na subárea de PO. Há uma diferença em termos percentuais entre as áreas de EP e ET que é percebida no caso das publicações em periódicos, principalmente no período de 2001 a 2010, e que não se observa no caso de publicações em congressos, orientações concluídas e participações em bancas, em que as porcentagens aproximadas de 70% e 30%, respectivamente, são mantidas, inclusive quando considerado apenas o período de 2001 a 2010. No caso dos bolsistas que se dizem atuar em PO, as suas publicações em periódicos correspondem a mais de 51% do todo dos bolsistas, o que destaca os atuantes dessa subárea dos demais de EP e ET. Nos outros três tópicos considerados a situação se inverte, uma vez que a porcentagem relativa aos bolsistas que se dizem atuar em PO é menor do que 45,06% do total.

A Tabela 6.11 apresenta uma lista dos 20 bolsistas que dizem atuar na área de PO que mais publicaram em periódicos no período de 2001 a 2010. No caso de empate para a lista ordenada foi considerado o ano de conclusão do doutorado.

Tabela 6.11 - Ranking dos bolsistas PQ da subárea de PO que mais publicaram em periódicos no período de 2001 a 2010.

Nome	Cat.	Ano	Instituição	Cidade	UF	P	C	O	B
João Carlos Correia Baptista Soares de Mello	1D	2002	UFF	Niterói	RJ	131 **(1°)**	187 **(1°)**	23 (17°)	49 (13°)
Luiz Flavio Autran Monteiro Gomes	2	1976	IBMEC	Rio de Janeiro	RJ	101 (2°)	88 (8°)	88 **(1°)**	110 **(1°)**
Eliane Gonçalves Gomes	1D	2003	EMBRAPA	Brasília	DF	82 (3°)	103 (5°)	-	9 (66°)
Reinaldo Morabito Neto	1A	1992	UFSCar	São Carlos	SP	71 (4°)	53 (26°)	26 (14°)	91 (2°)
José Luis Duarte Ribeiro	1C	1989	UFRGS	Porto Alegre	RS	61 (5°)	117 (2°)	76 (2°)	30 (32°)
Luiz Antonio Nogueira Lorena	1A	1985	INPE	São José dos Campos	SP	60 (6°)	56 (23°)	13 (43°)	-
Adiel Teixeira de Almeida	1A	1994	UFPE	Recife	PE	56 (7°)	113 (3°)	55 (5°)	37 (23°)
Lidia Ângulo Meza	2	2002	UFF	Volta Redonda	RJ	49 (8°)	76 (10°)	1 (83°)	22 (46°)
Petr Iakovlevitch Ekel	1B	1980	PUC Minas	Belo Horizonte	MG	44 (9°)	33 (44°)	19 (25°)	56 (6°)
Marcos Pereira Estellita Lins	1B	1993	UFRJ	Rio de Janeiro	RJ	41 (10°)	33 (43°)	31 (12°)	33 (29°)
Flavio Cesar Faria Fernandes	2	1991	UFSCar	São Carlos	SP	41 (11°)	32 (45°)	14 (38°)	-
Flávio Sanson Fogliato	1D	1997	UFRGS	Porto Alegre	RS	36 (12°)	99 (7°)	55 (4°)	59 (5°)
Maria Teresinha Arns Steiner	1C	1995	UFPR	Curitiba	PR	35 (13°)	65 (19°)	36 (9°)	81 (3°)
Moacir Godinho Filho	2	2004	UFSCar	São Carlos	SP	33 (14°)	35 (36°)	6 (61°)	27 (37°)
Nair Maria Maia de Abreu	1B	1984	UFRJ	Rio de Janeiro	RJ	33 (15°)	19 (58°)	9 (56°)	51 (10°)
Marcos Nereu Arenales	1A	1984	USP	São Carlos	SP	32 (16°)	15 (65°)	24 (16°)	-
Rosangela Helena Loschi	1C	1998	UFMG	Belo Horizonte	MG	30 (17°)	14 (68°)	14 (36°)	53 (9°)
Frederico Rodrigues Borges da Cruz	1D	1997	UFMG	Belo Horizonte	MG	30 (18°)	27 (51°)	11 (46°)	43 (18°)
Denis Borenstein	1B	1995	UFRGS	Porto Alegre	RS	29 (19°)	41 (32°)	42 (6°)	41 (21°)
Annibal Parracho Sant´anna	2	1977	UFF	Niterói	RJ	28 (20°)	56 (25°)	34 (11°)	10 (63°)
TOTAL						**1.023**	**1.262**	**577**	**802**

Fonte: Currículo Lattes dos bolsistas PQ (n = 186) e de PO (n = 91)

Além disso, esses 20 bolsistas também foram classificados de acordo com o número de artigos completos publicados em anais de congressos (C), o número de orientações concluídas em nível de mestrado e doutorado (O) e o número de participações em bancas examinadoras, também em nível de mestrado e doutorado (B).

O ranking apresentado leva em conta apenas a quantidade, não considerando nenhum critério qualitativo. Também vale ressaltar que os 20 bolsistas listados são todos da área de EP. O bolsista da ET mais bem classificado de acordo com o número de publicações em periódicos ocupa a 47ª posição, com 14 publicações, tendo obtido o título de doutor no ano de 1997.

A Tabela 6.12 apresenta a distribuição dos 5 periódicos mais utilizados para publicação destes 20 bolsistas de PO que mais publicaram no período de 2001 a 2010, levantados segundo o ISSN indicado dos periódicos. Dos 1.023 artigos analisados, 1.007 (98,44%) possuíam ISSN. Dessa forma, foi consultado o estrato do periódico (o maior e menor estrato e a área de avaliação dos mesmos) de acordo com o QUALIS Periódicos da CAPES.

É possível perceber que há uma prevalência de publicação em periódicos nacionais, com destaque para a revista “Pesquisa Operacional”. Também é possível perceber uma grande variação quanto aos estratos definidos pelas áreas de avaliação. Por exemplo, a revista “Gestão & Produção” é considerada “A2” pela área de Administração, Ciências e Turismo e “C” pela Computação. O total de publicações nesses 5 periódicos dos 20 bolsistas listados é de 324, o que corresponde a 31,67% do total (1.023).

Tabela 6.12 - Distribuição dos periódicos mais utilizados para publicação pelos bolsistas PQ da subárea de PO que mais publicaram em periódicos no período de 2001 a 2010.

ISSN	Título	Total	Estrato	Área de Avaliação
0101-7438	Pesquisa Operacional (Impresso)	114	A2	Administração, Ciências Contábeis e Turismo
			C	Ensino de Ciências e Matemática
0104-530X	Gestão & Produção (UFSCAR. Impresso)	69	A2	Administração, Ciências Contábeis e Turismo
			C	Ciência da Computação
1678-2399	Relatórios de Pesquisa em Engenharia de Produção (UFF)	52	B5	Engenharias III
			C	Administração, Ciências Contábeis e Turismo
0103-6513	Produção (São Paulo. Impresso)	50	A2	Administração, Ciências Contábeis e Turismo
			C	Ciência da Computação
0874-5161	Investigação Operacional	39	B3	Engenharias III
			B5	Engenharias IV

Fontes: Currículo Lattes dos 20 bolsistas PQ de PO que mais publicaram em periódicos (n = 20)

Qualis Periódicos da CAPES

A Tabela 6.13 apresenta a média de autores por artigos dos 20 bolsistas de PO que mais publicaram em periódicos no período de 2001 a 2010. Foi mantida a mesma ordem de classificação definida na Tabela 6.11 para facilitar a identificação dos bolsistas. Porém, outra classificação é apresentada de acordo com a média calculada pelo número de autores no período de 2001 a 2010 pelo número de artigos publicados por cada um dos bolsistas no mesmo período. Apesar de ter sido agrupado em um período de 10 anos, é possível fazer esse tipo de análise ano a ano ou em outros períodos. A média de autores por artigo desses 20 bolsistas é de 3.08 e 9 (45%) dos 20 bolsistas apresentam média superior a essa. Os 1.023 artigos considerados correspondem a 68,75% do total (1.488).

Tabela 6.13 - Média de autores por artigos dos bolsistas PQ da subárea de PO que mais publicaram em periódicos no período de 2001 a 2010.

Nome	Artigos Total	Artigos 2001-2010	Autores 2001-2010	Média Autores/Artigos	
João Carlos Correia Baptista Soares de Mello	136	131	467	3.56	(5°)
Luiz Flavio Autran Monteiro Gomes	294	101	281	2.78	(13°)
Eliane Gonçalves Gomes	83	82	312	3.80	(3°)
Reinaldo Morabito Neto	99	71	171	2.41	(16°)
José Luis Duarte Ribeiro	97	61	170	2.79	(12°)
Luiz Antonio Nogueira Lorena	73	60	153	2.55	(15°)
Adiel Teixeira de Almeida	63	56	134	2.39	(17°)
Lidia Ângulo Meza	51	49	193	3.94	(2°)
Petr Iakovlevitch Ekel	109	44	184	**4.18**	**(1°)**
Marcos Pereira Estellita Lins	52	41	140	3.41	(7°)
Flavio Cesar Faria Fernandes	58	41	96	2.34	(18°)
Flávio Sanson Fogliato	48	36	83	2.31	(19°)
Maria Teresinha Arns Steiner	46	35	122	3.49	(6°)
Moacir Godinho Filho	37	33	93	2.82	(11°)
Nair Maria Maia de Abreu	43	33	122	3.70	(4°)
Marcos Nereu Arenales	45	32	95	2.97	(10°)
Rosangela Helena Loschi	33	30	99	3.30	(8°)
Frederico Rodrigues Borges da Cruz	35	30	95	3.17	(9°)
Denis Borenstein	40	29	76	2.62	(14°)
Annibal Parracho Sant´anna	46	28	61	2.18	(20°)
TOTAL	**1.488**	**1.023**	**3.147**	**3.08**	

Fonte: Currículo Lattes dos 20 bolsistas PQ de PO que mais publicaram em periódicos (n = 20)

É importante destacar que os dados levantados para este estudo e o levantamento do perfil foram realizados no período de 20 de abril a 08 de maio de 2011. Isso somente foi possível porque foi utilizada a linguagem LattesMiner, que automatiza quase todo o trabalho e que vem sendo desenvolvida há alguns anos. A grande maioria das informações aqui apresentadas em tabelas foi obtida através de consultas SQL, o que também foi possível porque a linguagem LattesMiner extrai as informações dos currículos Lattes e as armazena, automaticamente, em um banco de dados. Dessa forma, é simples obter outras informações e realizar análises mais aprimoradas e/ou aprofundadas. Por exemplo, levantamento dos bolsistas e

quantidade de publicação em periódicos levando em conta os estratos dos periódicos em que publicaram; levantamento dos bolsistas e quantidade de citações dos seus artigos em periódicos indexados na WoS etc.

Através do conhecimento do perfil dos pesquisadores da subárea de PO, podem ser definidas, de maneira mais eficaz, por exemplo, estratégias para incentivar a produção científica e acompanhar os seus resultados. Além disso, pode tornar mais transparentes as avaliações feitas dos pesquisadores contemplados com recursos de agências de fomento como as bolsas PQ e outros auxílios, bem como contextualizar os pesquisadores que hoje compõem a subárea de PO no Brasil.

6.3. Área

No estudo de caso anterior, foi realizado um estudo sobre o perfil dos bolsistas de Produtividade em Pesquisa do CNPq com enfoque na área de PO utilizando informações extraídas dos seus currículos Lattes. A PL, no entanto, não permite realizar um estudo mais completo das áreas, pois podem existir pesquisadores atuando no Brasil que não possuem currículo Lattes. Além disso, a PL não permite identificar de maneira precisa quais pesquisadores realmente atuam numa determinada área.

Neste estudo de caso é apresentado o mapeamento da produção científica brasileira na área de Pesquisa Operacional (PO) considerando-se as informações sobre os artigos publicados em periódicos indexados na WoS e classificados na categoria "*Operations Research & Management Science*" (OR&MS). Também foi realizada uma comparação da produção científica brasileira com a produção científica mundial na área de PO e uma análise das citações recebidas por toda a produção científica brasileira na área de PO. Conhecer o perfil de uma área permite, por exemplo, definir estratégias de incentivo para a área.

Pode-se dizer que a PO no Brasil teve um primeiro esforço coordenado de pesquisadores no ano de 1968, com a realização do I Simpósio de Pesquisa

Operacional realizado no Instituto Tecnológico de Aeronáutica (ITA), em São José dos Campos, São Paulo. A Sociedade Brasileira de Pesquisa Operacional (SOBRAPO) foi fundada no ano seguinte, em 1969. Desde esta época, os pesquisadores brasileiros da área de PO têm publicado em diversos periódicos importantes contribuindo assim para a divulgação da produção científica brasileira nessa área.

6.3.1. Coleta de dados

Inicialmente, foi verificada qual das 249 categorias definidas pela WoS era diretamente relacionada com a área de PO. A categoria OR&MS foi escolhida, permitindo obter a produção científica brasileira em artigos publicados nos periódicos indexados na WoS e classificados nessa categoria. Foi então realizada uma busca por artigos publicados em periódicos de acordo com os critérios informados na Figura 6.5. Os principais critérios definidos foram a categoria OR&MS e o país como "Brazil". A cobertura da WoS selecionada para essa busca foi apenas o Índice Expandido de Citações Científicas ("*Science Citation Index Expanded*™ - *SCI-EXPANDED*), que engloba mais de 8.000 periódicos da área de Ciências desde o ano de 1945.

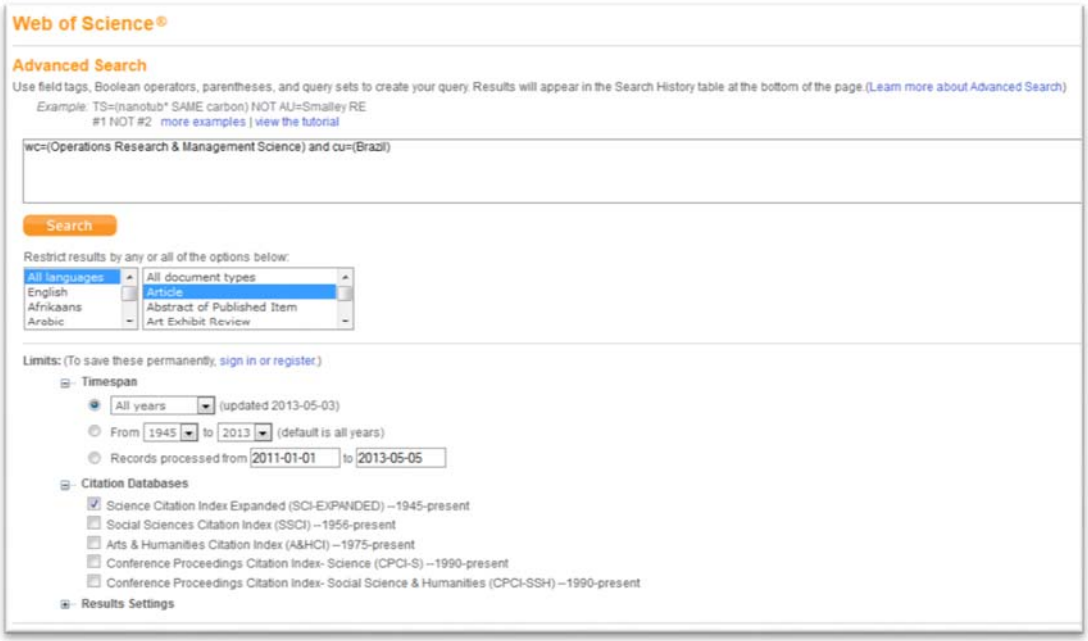


Figura 6.5 - Definição de critérios de busca na base de dados WoS para artigos brasileiros publicados em periódicos e classificados na categoria OR&MS.

O número de artigos encontrados nessa categoria foi 1.515. A partir disso, foi possível baixar as informações desses artigos no formato bibtex disponível na WoS. Dessa forma, foi feita a extração automática e as informações foram armazenadas, também de forma automática, em um banco de dados. Todo esse processo, brevemente aqui descrito, foi realizado no dia 5 de maio de 2013 em poucos minutos.

6.3.2. Resultados e discussões

O total de 1.515 artigos publicados em periódicos classificados na categoria OR&MS coloca o Brasil na décima nona (19ª) posição entre os países que mais publicaram nessa categoria de acordo com a WoS, conforme destacado na Tabela 6.14.

Nessa busca foi verificado que 169 países possuem pelo menos um artigo publicado em algum periódico indexado na WoS e classificado na categoria OR&MS. Se considerarmos toda a produção científica mundial, a categoria OR&MS ocupa a centésima quinta (105ª) posição de um total de 235 categorias distintas classificadas na WoS e é responsável por quase meio por cento (0,46%) de toda a produção. Esse percentual é muito semelhante ao percentual brasileiro (0,42%) nessa categoria, ocupando a centésima décima segunda (112ª) posição dentre as 226 categorias identificadas nas publicações de pesquisadores brasileiros.

O número total de artigos publicados em periódicos indexados na WoS na categoria OR&MS é 124.969. Porém, quando consideramos o número de artigos por país, um determinado artigo pode ser contabilizado por mais de um país. Isso se deve ao fato de que a WoS faz essa contagem considerando todas as afiliações informadas nos artigos (ALMEIDA; GUIMARÃES, 2013).

Tabela 6.14 - Ranking da produção científica mundial de acordo com artigos publicados em periódicos indexados na base de dados WoS e classificados na categoria OR&MS.

Posição	País	Artigos	%
1ª	Estados Unidos	43.306	34,65
2ª	China	10.244	8,20
3ª	Inglaterra	8.290	6,63
4ª	Canadá	8.124	6,50
5ª	Taiwan	6.795	5,44
6ª	França	5.775	4,62
7ª	Alemanha	4.532	3,63
8ª	Itália	4.351	3,48
9ª	Japão	4.210	3,37
10ª	Holanda	4.005	3,20
11ª	Espanha	3.766	3,01
12ª	Coréia do Sul	3.252	2,60
13ª	Índia	3.151	2,52
14ª	Austrália	2.968	2,37
15ª	Israel	2.592	2,07
16ª	Turquia	2.347	1,88
17ª	Bélgica	1.863	1,49
18ª	Singapura	1.822	1,46
19ª	**Brasil**	**1.515**	1,21
20ª	Grécia	1.391	1,11

Fonte: Web of Science (05/05/2013)

O primeiro artigo da área de PO em nível mundial de acordo com os periódicos atualmente indexados na WoS foi de Bernard Osgood Koopman, da *Columbia University* (KOOPMAN, 1952). Esse artigo foi publicado em 1952 no *Journal of the Operations Research Society of America*. Também de acordo com a WoS, o artigo de Storn e Price (1997) foi o mais citado da categoria OR&MS, com 2.664 citações até o dia da extração.

Considerando somente os 1.515 artigos publicados com pelo menos um autor brasileiro, de acordo com as informações obtidas na WoS, o primeiro artigo brasileiro da área de PO foi publicado por Bitran e Novaes (1973) e o artigo mais citado possui 459 citações. Esse artigo foi publicado em 1999 no periódico Systems & Control Letters (OLIVEIRA et al., 1999).

A Tabela 6.15 apresenta o número de artigos publicados em periódicos indexados na WoS e classificados na categoria OR&MS, desde 1973 até o dia 5 de maio de 2013. Também é apresentado o número de citações recebidas por esses artigos e a média de autores.

Do total de 1.515 artigos publicados em periódicos, 374 (24,69%) artigos ainda não foram citados. Porém, 197 (52,67%) desses artigos foram publicados em 2012 ou em 2013. Desconsiderando esses artigos, a porcentagem de artigos não citados é pequena (11,68%). Os outros 1.141 artigos (75,31%) foram citados por 12.856 artigos, com uma média de 11,27 citações por artigo. Nos últimos anos percebe-se que a média de autores por artigo está aumentando. Porém, esse aumento não é muito significativo comparando com a média do período que é de 2,68 autores por artigo. O número de artigos também está aumentando nos últimos anos, sendo que em 2011 o aumento foi considerável (31,79%).

Para publicar esses 1.515 artigos, os pesquisadores brasileiros da área de PO colaboraram com autores de outros 48 países. A Figura 6.6 apresenta a distribuição geográfica dos coautores que publicaram artigos com pesquisadores brasileiros em periódicos indexados na WoS e classificados na categoria OR&MS. São destacados na cor “cinza” os países (13) em que houve apenas uma única colaboração. Os 5 países que os pesquisadores brasileiros mais colaboraram foram: Estados Unidos (160; 10,56%), França (100; 6,60%), Inglaterra (65; 4,29%), Canadá (49; 3,23%) e Chile (38; 2,51%).

Tabela 6.15 - Número de artigos brasileiros publicados em periódicos indexados na base de dados WoS e classificados na categoria OR&MS.

Ano	Artigos	% Artigos Citados	Citações	Citações / Artigos	Autores / Artigos
1973	1	100	28	28,00	2,00
1974	3	100	34	11,33	1,00
1975	1	100	4	4,00	3,00
1976	2	50,00	18	9,00	1,50
1977	2	100	65	32,50	1,00
1978	1	100	87	87,00	2,00
1979	3	100	18	6,00	1,67
1980	1	100	42	42,00	1,00
1981	2	100	4	2,00	1,50
1982	5	80	231	46,20	2,00
1983	9	66,67	26	2,89	1,56
1984	6	100	150	25,00	2,17
1985	3	33,33	1	0,33	2,00
1986	6	100	133	22,17	2,17
1987	7	57,14	178	25,43	1,71
1988	7	85,71	146	20,86	1,86
1989	10	100	250	25,00	2,00
1990	14	92,86	223	15,93	1,79
1991	20	95,00	416	20,80	2,00
1992	7	85,71	25	3,57	1,86
1993	11	63,64	85	7,73	1,91
1994	15	86,67	373	24,87	2,40
1995	23	100	237	10,30	1,91
1996	16	87,50	236	14,75	2,63
1997	27	81,48	279	10,33	2,48
1998	40	95,00	711	17,78	2,08
1999	42	90,48	1.075	25,60	2,43
2000	53	96,23	922	17,40	2,21
2001	57	85,96	667	11,70	2,42
2002	51	96,08	676	13,25	2,57
2003	44	88,64	588	13,36	2,70
2004	39	94,87	374	9,59	2,77
2005	71	97,18	906	12,76	2,54
2006	70	97,14	820	11,71	2,61
2007	70	90,00	592	8,46	2,60
2008	103	91,26	725	7,04	2,80
2009	122	90,16	687	5,63	2,71
2010	138	78,99	465	3,37	3,05
2011	151	57,62	219	1,45	3,04
2012	199	30,65	135	0,68	3,01
2013*	63	6,35	5	0,08	3,21
Total	**1.515**	**75,31**	**12.856**	**8,49**	**2,68**

* Até 05 de maio de 2013

Figura 6.6 - Distribuição geográfica dos coautores que publicaram artigos com pesquisadores brasileiros em periódicos indexados na base de dados WoS e classificados na categoria OR&MS.

A distribuição geográfica dos autores brasileiros não foi apresentada porque não é possível identificar corretamente os endereços dos autores a partir das informações obtidas na WoS. O endereço informado na WoS é obtido a partir dos artigos dos pesquisadores e não há uma padronização para isso. Nota-se que em cada artigo o endereço é informado de uma forma, o que dificulta ou até mesmo impossibilita a extração automática dessa informação.

A Tabela 6.16 apresenta a distribuição dos 20 periódicos mais utilizados, desde 1973, para a publicação desses 1.515 artigos em periódicos indexados na WoS e classificados na categoria OR&MS. Com o ISSN foi obtido o FI dos periódicos de acordo com o JCR® de 2011. Nesse período foram utilizados 82 periódicos diferentes com FI. Também é apresentada a média de citações por artigos publicados em cada um dos periódicos. É possível perceber que as médias variam bastante de um periódico para outro e que nem sempre, os periódicos mais utilizados são os mais citados.

Tabela 6.16 - Distribuição dos periódicos mais utilizados pelos pesquisadores brasileiros para publicação considerando a categoria OR&MS na base de dados WoS.

ISSN	Periódico	Artigos (A)	Citações (C)	C / A	FI 2011
0377-2217	European Journal of Operational Research* (6/77)	148	1.224	8,27	1.815
0957-4174	Expert Systems with Applications[3] (5/77)	145	347	2,39	2.203
0305-0548	Computers & Operations Research[3] (10/77)	107	844	7,89	1.720
0025-5610	Mathematical Programming[3] (11/77)	78	1.598	20,49	1.707
0022-3239	Journal of Optimization Theory and Applications[2] (28/77)	74	757	10,23	1.062
0254-5330	Annals of Operations Research* (41/77)	69	466	6,75	0.840
0167-6911	Systems & Control Letters[2] (23/77)	60	1.445	24,08	1.222
0925-5273	International Journal of Production Economics[3] (8/77)	55	449	8,16	1.760
0926-6003	Computational Optimization and Applications[2] (20/77)	54	487	9,02	1.350
0160-5682	Journal of the Operational Research Society* (35/77)	50	377	7,54	0.971
0020-7543	International Journal of Production Research[3] (25/77)	48	266	5,54	1.115
0020-7721	International Journal of Systems Science[3] (33/77)	42	270	6,43	0.991
0233-1934	Optimization[2] (58/77)	31	162	5,23	0.500
0969-6016	International Transactions in Operational Research* (49/77)	30	23	0,77	0.648
0399-0559	RAIRO-Operations Research* (77/77)	30	63	2,10	0.220
0028-3045	Networks[2] (34/77)	29	299	10,31	0.983
0951-8320	Reliability Engineering & System Safety[2] (7/77)	27	159	5,89	1.770
0925-5001	Journal of Global Optimization[2] (24/77)	27	202	7,48	1.196
0167-6377	Operations Research Letters* (55/77)	22	60	2,73	0.537
1055-6788	Optimization Methods & Software[3] (48/77)	21	109	5,19	0.651

Fontes: Web of Science (05/05/2013) e JCR® (2011)

[2,3] Número de categorias em que o periódico está classificado no JCR® (2011, *Science Edition*)

* Periódico classificado somente na categoria OR&MS

( ) Posição de acordo com o FI do periódico em 2011 dentre os 77 periódicos classificados na categoria OR&MS

No JCR® (*Science Edition*) de 2011 há 77 periódicos classificados na categoria OR&MS, sendo que 20 (25,97%) desses periódicos estão classificados somente nessa categoria. O número de categorias em que os outros periódicos (57) estão classificados varia de 2 a 4. Na Tabela 6.16 é apresentada a posição em que cada periódico se encontra, de acordo com o FI em 2011, dentre os 77 periódicos da categoria OR&MS.

Na Tabela 6.17 é apresentada a mesma distribuição só que em nível mundial. É possível observar que dentre os 20 periódicos mais utilizados por pesquisadores brasileiros, 13 (65%) também são os mais utilizados por pesquisadores em nível mundial na categoria OR&MS. Além disso, os 2 periódicos mais utilizados são exatamente os mesmos e ambos estão muito bem ranqueados de acordo com o FI em 2011, sexto (6º) e quinto (5º) respectivamente, entre os 77 periódicos classificados na categoria OR&MS. Na Tabela 6.16 é destacado em "azul" os periódicos ranqueados na mesma posição nas duas distribuições e na cor "verde" os que estão entre os 20, porém em posições diferentes.

Atualmente, os periódicos na WoS são classificados em 249 categorias distintas. Os 1.515 artigos classificados na categoria OR&MS também estão classificados em 25 dessas categorias. Na Figura 6.7 são apresentadas as 20 principais categorias relacionadas com a categoria OR&MS na WoS de acordo com os artigos publicados por brasileiros em periódicos indexados nessa categoria.

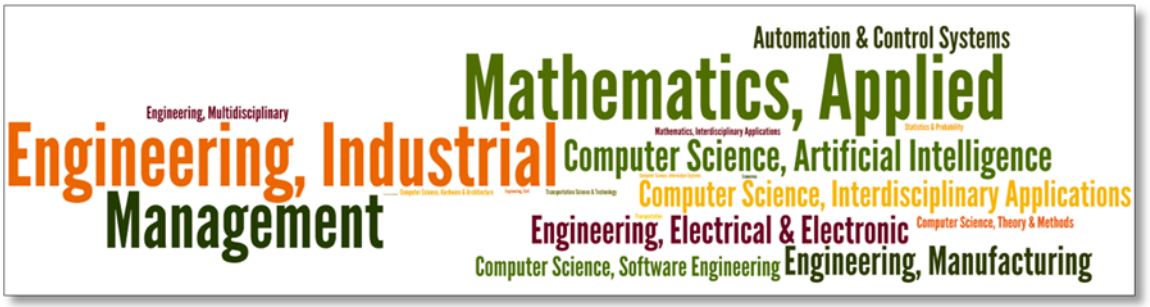


Figura 6.7 - Principais categorias relacionadas com a categoria OR&MS na base de dados WoS de acordo com a produção científica brasileira.

Tabela 6.17 - Distribuição dos periódicos mais utilizados pelos pesquisadores em nível mundial para publicação considerando a categoria OR&MS na base de dados WoS.

ISSN	Periódico	Artigos	%	FI 2011
0377-2217	European Journal of Operational Research* (6/77)	9.090	7,27	1.815
0957-4174	Expert Systems with Applications[3] (5/77)	7.833	6,27	2.203
0020-7543	International Journal of Production Research[3] (25/77)	6.871	5,50	1.115
0020-7721	International Journal of Systems Science[3] (33/77)	5.265	4,21	0.991
0022-3239	Journal of Optimization Theory and Applications (28/77)	4.799	3,84	1.062
0030-364X	Operations Research[2] (13/77)	4.476	3,58	1.665
0160-5682	Journal of the Operational Research Society* (35/77)	4.435	3,55	0.971
0925-5273	International Journal of Production Economics[3] (8/77)	4.153	3,32	1.760
0305-0548	Computers & Operations Research[3] (10/77)	3.827	3,06	1.720
0167-6911	Systems & Control Letters[2] (23/77)	3.493	2,80	1.222
0951-8320	Reliability Engineering & System Safety[2] (7/77)	3.350	2,68	1.770
0025-1909	Management Science* (9/77)	2.847	2,28	1.733
0025-5610	Mathematical Programming[3] (11/77)	2.832	2,27	1.707
0033-524X	Quality Progress	2.430	1,94	-
0254-5330	Annals of Operations Research* (41/77)	2.391	1,91	0.840
0894-069X	Naval Research Logistics* (31/77)	2.348	1,88	1.038
0167-6377	Operations Research Letters* (55/77)	2.289	1,83	0.537
0740-817X	IIE Transactions[2] (39/77)	2.200	1,76	0.856
0167-9236	Decision Support Systems[3] (12/77)	1.988	1,59	1.687
0925-7535	Safety Science[2] (19/77)	1.709	1,37	1.402

Fontes: Web of Science (05/05/2013) e JCR® (2011)

[2,3] Número de categorias em que o periódico está classificado no JCR® (2011, *Science Edition*)

* Periódico classificado somente na categoria OR&MS

( ) Posição de acordo com o FI do periódico em 2011 dentre os 77 periódicos classificados na categoria OR&MS

As 5 principais categorias relacionadas com a categoria OR&MS de acordo com a produção científica brasileira em artigos publicados em periódicos indexados na WoS são: "*Mathematics, Applied*" (342; 22,57%), "*Engineering, Industrial*" (333; 21,98%); "*Management*" (296; 19,54%), "*Computer Science, Artificial Intelligence*" (165; 10,89%) e "*Engineering, Manufacturing*" (152; 10,03%).

Outro tópico considerado neste estudo foram as palavras-chave mais utilizadas nos artigos publicados por pesquisadores brasileiros em periódicos indexados

na WoS e classificados na categoria OR&MS. A Figura 6.8 apresenta as 50 palavras-chave mais utilizadas nos 1.515 artigos analisados. Ao todo foram utilizadas 3.589 palavras-chave distintas, sendo que as 5 mais utilizadas foram: "*Heuristics*" (48), "*Nonlinear programming*" (40), "*Metaheuristics*" (33), "*Integer Programming*" (31) e "*Combinatorial Optimization*" (31). É possível identificar os autores e os respectivos grupos de pesquisa de acordo com as palavras-chave, caso haja interesse, por exemplo, quando se busca especialistas em determinadas áreas do conhecimento.

Figura 6.8 - Palavras-chave mais utilizadas na produção científica brasileira em artigos publicados em periódicos indexados na WoS e classificados na categoria OR&MS.

A WoS também permite identificar quais e quantos artigos publicados em periódicos indexados citaram algum dos 1.515 artigos publicados por pesquisadores brasileiros e classificados na categoria OR&MS. De acordo com a WoS, esses artigos foram citados por outros 9.142 artigos, desconsiderando-se autocitações. Esse número diminui para 6.938 considerando-se apenas citações em artigos publicados em periódicos indexados. Considerando o país de afiliação informado nesses artigos pelos seus autores, foi possível obter a distribuição geográfica desses autores, conforme ilustra a Figura 6.9.

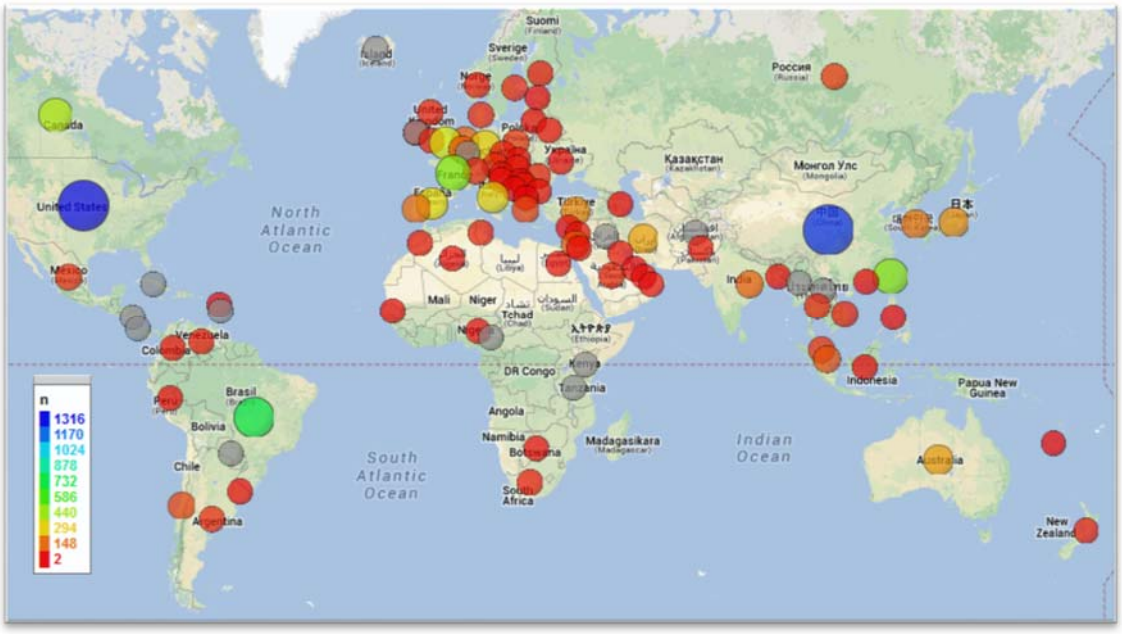


Figura 6.9 - Distribuição geográfica dos autores que citaram artigos publicados por pesquisadores brasileiros em periódicos indexados na base de dados WoS e classificados na categoria OR&MS, desconsiderando as autocitações.

Os 1.515 artigos publicados por pesquisadores brasileiros em periódicos indexados na WoS e classificados na categoria OR&MS foram citados por autores de 98 países diferentes, sendo que os 5 países mais identificados foram: Estados Unidos (1.316; 18,97%), China (1.257; 18,12%); Brasil (742; 10,69%), França (481; 6,93%) e Taiwan (467; 6,73%). É importante ressaltar novamente que foram desconsideradas todas as autocitações.

A partir dos 6.938 artigos que citaram algum dos 1.515 artigos publicados por pesquisadores brasileiros em periódicos indexados na WoS e classificados na categoria OR&MS, foram identificadas 150 categorias distintas em que os periódicos utilizados para publicar esses artigos estão classificados na WoS. Na Figura 6.10 são apresentadas as 10 categorias mais identificadas.

Figura 6.10 - Categorias dos artigos que citaram algum artigo de pesquisador brasileiro classificado na categoria OR&MS na base de dados WoS.

Como era de se esperar, a principal categoria é a própria categoria OR&MS, que foi identificada em 2.757 (39,74%) artigos. Dentre as categorias mais identificadas também estão: "*Mathematics, Applied*" (1.697; 24,46%), "*Automation & Control Systems*" (897; 12,93%), "*Engineering, Industrial*" (888; 12,80%) e "*Management*" (783; 11,29%). Dentre essas 4 categorias, 3 também estão entre as 5 mais identificadas nos 1.515 artigos analisados.

Também a partir dos 6.938 artigos que citaram algum dos 1.515 artigos publicados por pesquisadores brasileiros em periódicos indexados na WoS e classificados na categoria OR&MS, foram identificadas 13.643 palavras-chave distintas, sendo que as 50 mais utilizadas são apresentadas na Figura 6.11. As 5 palavras-chave mais utilizadas foram: "*Scheduling*" (221), "*Heuristics*" (161), "*Optimization*" (161), "*Genetic Algorithm*" (130) e "*Integer Programming*" (116).

Figura 6.11 - Palavras-chave mais utilizadas nos artigos que citaram artigos de pesquisadores brasileiros publicados em periódicos indexados na base de dados WoS e classificados na categoria OR&MS.

Dentre as 10 palavras-chave mais utilizadas nos artigos que citaram artigos de pesquisadores brasileiros, 7 (70%) também estão entre as 10 palavras-chave mais utilizadas na produção científica brasileira na WoS na categoria OR&MS. Isso demonstra que a produção científica brasileira está fortemente relacionada com a produção cientifica mundial nessa categoria.

Outra questão importante levantada em relação as citações de artigos de pesquisadores brasileiros é o fato de que 286 (18,88%) dos 1.515 artigos analisados contribuíram para 70% (8.999) do total de 12.856 citações; 409 (27,00%) artigos contribuíram para 80% (10.285) das citações e 602 (39,74%) artigos contribuíram para 90% das citações.

6.4. Grande área

Este estudo de caso tem como objetivo avaliar o perfil dos bolsistas PQ do CNPq da área de Química no Brasil de acordo com as informações contidas na PL, considerando a produção científica desses bolsistas nos últimos 10 anos (2002 a 2011).

Santos et al. (2010) analisaram o perfil dos 604 bolsistas PQ da área de Química com bolsa vigente em março de 2009, com base nas informações da PL. Nesse trabalho, pioneiro na área de Química, é possível ter uma visualização do perfil desses bolsistas, sua distribuição geográfica, por instituições acadêmicas, por gênero, por idade científica (número de anos decorridos desde o ano de publicação do primeiro artigo em periódico indexado) e por subáreas de atuação. Também foram considerados os índices numéricos de produtividade extraídos dos currículos Lattes, tais como índice de orientação (IO) (SANTOS et al., 2010), índice H (HIRSCH, 2005), número de artigos e somatório dos impactos.

O diferencial deste estudo está no fato que todo o processo de aquisição e extração dos dados foi feito automaticamente utilizando a linguagem LattesMiner, gastando um tempo bem menor. Este estudo também permite comparar dados mais recentes com o que foi observado há alguns anos, além de considerar atributos não utilizados no trabalho anterior.

6.4.1. Coleta de dados

Neste estudo foi utilizada a relação de bolsistas PQ, com bolsas ativas em 3 de dezembro de 2012, disponível no sítio do CNPq (Bolsas em Curso) da área de Química. Os bolsistas considerados foram apenas os listados com indicação de

estarem em “Em folha de Pagamento”; os demais, por exemplo, com bolsas suspensas não foram considerados. O número de bolsistas encontrados nesta condição foi 695.

A partir da identificação dos bolsistas, foram obtidos os seus currículos Lattes utilizando a linguagem LattesMiner. Inicialmente, foi criado um arquivo texto com o nome dos bolsistas PQ da área de Química. Em seguida, a linguagem LattesMiner obteve automaticamente o número identificador (ID) de cada um deles, permitindo com isso que os seus currículos fossem baixados e armazenados como arquivos HTML. Foi feita a extração automática dos dados, sendo os mesmos armazenados, também de forma automática, em um banco de dados.

As seguintes informações foram extraídas automaticamente: dados pessoais, endereço profissional, formação acadêmica, produção bibliográfica em periódicos, índice H e número de citações na WoS, nomes em citações bibliográficas, orientações concluídas, contatos (é considerado um contato todo “link” identificado no currículo Lattes de um bolsista para outro currículo Lattes), idiomas e áreas de atuação. Todo esse processo foi realizado no dia 3 de dezembro de 2012 em menos de 3 h.

Os seguintes atributos foram considerados neste estudo: gênero, categoria do bolsista, instituição de origem, ano de conclusão do doutorado, distribuição geográfica, artigos completos publicados em periódicos, índice H, número de citações na WoS, idiomas, tempo decorrido após conclusão do doutorado, orientações concluídas de iniciação científica, mestrado e doutorado, áreas de atuação, periódicos utilizados para publicação, classificação de periódicos segundo o Qualis Periódicos da CAPES de 2012, fator de impacto dos periódicos de acordo com o JCR® de 2011, somatório do fator de impacto dos periódicos, número de autores por artigo publicados em periódicos, relacionamentos entre os bolsistas e a população segundo o censo do IBGE (IBGE, 2013) de 2010.

6.4.2. Resultados e discussões

Dos 695 bolsistas PQ da área de Química, 674 (97,0%) atualizaram o currículo Lattes em 2012, sendo que 535 (77,0%) deles nos últimos 3 meses considerando a data de coleta dos dados. 690 (99,3%) dos bolsistas declaram a área de Química como a sua principal área de atuação. Os demais bolsistas (5), 2 declararam atuar na área de Física, 1 em Bioquímica, 1 em Farmácia e o outro não informou. Dentre as mais de 200 diferentes subáreas de atuação informadas nos currículos Lattes dos bolsistas, as que mais se destacaram foram: físico-química (40,1%), química orgânica (38,8%), química analítica (30,6%) e química inorgânica (25,2%).

Na Tabela 6.18 é apresentada a distribuição dos bolsistas em cada uma das categorias e por gênero. Pode-se observar que a maioria é do sexo masculino (67,9%) e se concentra na categoria 2 (63,2%). Também é possível observar que há um predomínio masculino em todas as categorias, principalmente nas categorias SR e 1A. Isso é esperado tendo em vista o predomínio masculino na titulação e na carreira universitária há algumas décadas passadas.

Tabela 6.18 - Distribuição dos bolsistas PQ da área de Química segundo categoria e gênero.

Categoria	n	%	Tempo médio decorrido após o doutorado (anos)	Masculino	%	Feminino	%
SR	7	1,0	43,4	6	85,7	1	14,3
1A	45	6,5	31,4	42	93,3	3	6,7
1B	46	6,6	27,8	38	82,6	8	17,4
1C	63	9,0	23,3	45	71,4	18	28,6
1D	93	13,4	19,6	66	71,0	27	29,0
2	439	63,2	14,7	273	62,2	166	37,8
2F	2	0,3	7,5	2	100	0	0
Total	**695**	**100**	**18,4**	**472**	**67,9**	**223**	**32,1**

O tempo médio decorrido após conclusão do doutorado dos bolsistas é de 18,4 anos, variando de 3 a 54 anos e a grande maioria (45,2%) com tempo de 11 a 20 anos. Até 10 anos (146), a grande maioria (95,9%) é da categoria 2. Na

faixa de 11 a 20 anos (314), 69,4% são da categoria 2 e 18,5% da categoria 1D. Na faixa de 21 a 30 anos (167), 43,7% são da categoria 2 e 16,8% da categoria 1C. Na faixa de 31 a 40 anos (52), 26,9% são da categoria 1A e 25,0% são da categoria 1B. E entre os bolsistas com tempo maior que 40 anos (16), 50,0% são da categoria 1A e 25,0% da categoria SR.

Dos 695 bolsistas, 692 (99,6%) informaram o seu endereço profissional. A Figura 6.12 apresenta a distribuição geográfica deles segundo seus endereços profissionais e o número de bolsistas por milhão de habitantes (entre [ ]). O mapa foi gerado utilizando a ferramenta *GPS Visualizer*, sendo destacados na cor “cinza” todos os estados em que o número de bolsistas é menor do que 10. O Sudeste é a região do Brasil com o maior número de bolsistas tendo o estado de São Paulo como principal destaque, com mais de um terço dos bolsistas. Há bolsistas em quase todos os estados do país. As exceções ocorrem na região Norte, em que os estados do Acre, Amapá, Roraima e Tocantins não possuem bolsistas e na região Centro-Oeste, em que o estado de Mato Grosso não possui bolsista.

Analisando o número de bolsistas por milhão de habitantes, o grande destaque é o estado do Rio Grande do Sul (6,8) que possui quase o dobro de bolsistas por milhão de habitantes quando comparado com a média do país que é de 3,6. Ainda na região Sul, o estado de Santa Catarina (4,8) também está acima da média. Na região Sudeste, é interessante observar que todos os estados, com exceção do Espírito Santo, estão acima da média nacional de bolsistas por milhão de habitantes. Na região Nordeste, os estados de Sergipe (5,3) e da Paraíba (3,7) também estão acima da média nacional. Na região Centro-Oeste, o Distrito Federal (5,1) também se encontra nessa condição.

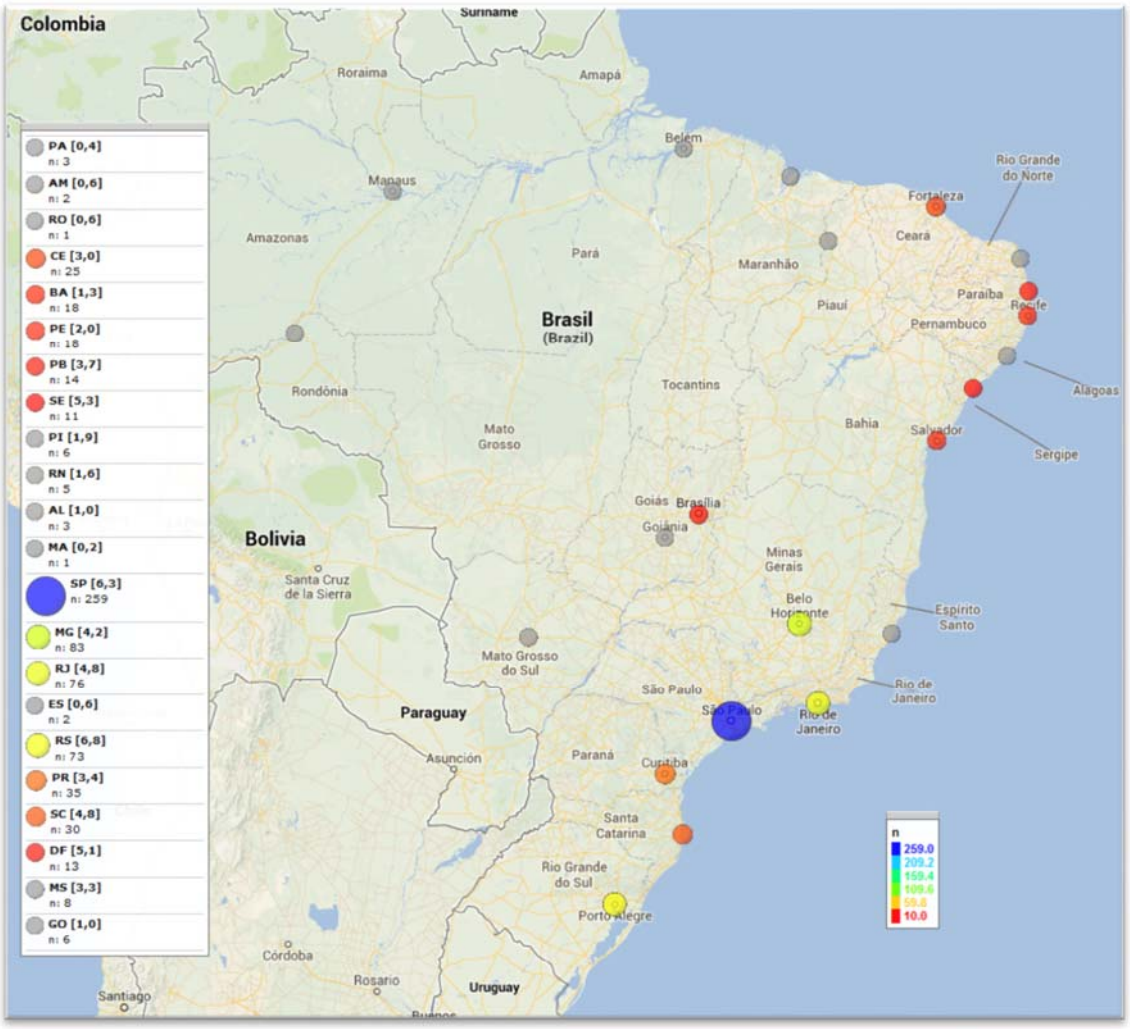


Figura 6.12 - Distribuição geográfica dos bolsistas PQ da área de Química.

Os bolsistas da categoria SR estão concentrados na região Sudeste, sendo que apenas 1 (14,3%) do total de 7 está na região Sul. Os bolsistas da categoria 1A estão um pouco mais distribuídos. Apesar da grande maioria (71,1%) estar na região Sudeste, também há bolsistas nas regiões Sul (17,8%) e Nordeste (11,1%). Os bolsistas da categoria 1B também se concentram na região Sudeste (76,1%). O mesmo ocorre com os bolsistas da categoria 1C (69,8%) e com os bolsistas da categoria 1D (60,2%). Já os bolsistas da categoria 2 estão distribuídos em 20 estados mais o Distrito Federal e a maioria (56,3%) também se encontra na região Sudeste. E os bolsistas da categoria 2F estão na região Nordeste (50,0%) e Sul (50,0%).

No endereço profissional os bolsistas também informam a instituição em que trabalham. São 77 instituições diferentes sendo que as 5 mais informadas são: Universidade de São Paulo (104), Universidade Estadual de Campinas (56), Universidade Federal de Minas Gerais (43), Universidade Federal do Rio de Janeiro (42) e Universidade Estadual Paulista Júlio de Mesquita Filho (40). Todas essas instituições estão na região Sudeste e representam pouco mais de 40% dos bolsistas PQ da área de Química. Na região Sul, a Universidade Federal do Rio Grande do Sul está na sexta posição com 37 bolsistas. Na região Nordeste, a Universidade Federal do Ceará é a instituição melhor colocada ocupando a nona posição com 23 bolsistas. Na região Centro-Oeste, a instituição melhor colocada é a Universidade de Brasília que ocupa a décima quarta posição com 13 bolsistas. E na região Norte, a Universidade Federal do Pará é a instituição melhor colocada ocupando a trigésima sexta posição com 3 bolsistas. É interessante destacar também que mais da metade (55,6%) dos bolsistas da categoria 1A trabalha na Universidade de São Paulo ou na Universidade Estadual de Campinas.

Dos 695 bolsistas, 688 (99,0%) informaram ter conhecimento de idiomas (total de 19 diferentes), sendo que Inglês (98,8%), Espanhol (77,8%) e Francês (49,9%) foram os três principais. Dos 687 bolsistas que declararam ter conhecimento no idioma “Inglês”, 358 (52,1%) declararam ler, escrever, falar e compreender bem. Desse total, 203 (56,7%) são da categoria 2 e 32 (71,1%) dos 45 bolsistas da categoria 1A também declaram ter esse conhecimento.

O número de artigos publicados em periódicos no período de 2002 a 2011 foi de 32.873 artigos, conforme destacado na Tabela 6.19. Dividindo esse período em dois períodos de 5 anos, percebe-se que houve um aumento de 23,9% no número de artigos publicados em periódicos entre o período de 2002 a 2006 e o período de 2007 a 2011. A média de artigos por ano de cada um dos 695 bolsistas foi de 4,7.

Tabela 6.19 - Número de artigos publicados em periódicos pelos bolsistas PQ da área de Química no período de 2002 a 2011.

Período	Artigos	PQ	Artigos / PQ
2002	2.608	585	4,5
2003	2.493	596	4,2
2004	2.797	633	4,4
2005	3.310	648	5,1
2006	3.474	666	5,2
2002 a 2006	**14.682**	**692**	**21,2**
2007	3.476	667	5,2
2008	3.683	670	5,5
2009	3.645	664	5,5
2010	3.756	664	5,7
2011	3.631	659	5,5
2007 a 2011	**18.191**	**695**	**26,2**
2002 a 2011	**32.873**	**695**	**47,3**

Do total de publicações em periódicos no período de 2002 a 2011, 31.876 (97,0%) artigos possuíam ISSN. A Tabela 6.20 apresenta a lista dos periódicos com mais de 200 artigos no período de 2002 a 2011 mais utilizados para publicação nesse período, obtidos a partir dos ISSN indicados. Nesse período, os bolsistas publicaram em 1.979 periódicos diferentes. Assim também foi possível obter o estrato do periódico de acordo com o Qualis Periódicos da CAPES de 2012, considerando a área de avaliação da "Química". Também foi possível obter o fator de impacto dos periódicos correspondentes no JCR® de 2011.

É possível perceber que há uma prevalência de publicação em periódicos nacionais, que ocupam as duas primeiras posições da lista. O primeiro periódico da lista é "Química Nova", que teve em média 187,2 artigos publicados nele por ano. Dos 695 bolsistas, 489 (70,4%) publicaram pelo menos uma vez nesse periódico no período de 2002 a 2011. E o segundo periódico da lista é o "*Journal of the Brazilian Chemical Society*", com média de

180,3 artigos publicados nele por ano. Dos 695 bolsistas, 498 (71,7%) publicaram nesse periódico. Dos periódicos da lista, apenas 2 (Revista Brasileira de Farmacognosia e Eclética Química) não possuem fator de impacto no JCR® de 2011, sendo que o primeiro fazia parte do JCR® quando a maioria dos artigos considerados foi publicado.

Tabela 6.20 - Distribuição dos periódicos mais utilizados para publicação pelos bolsistas PQ da área de Química no período de 2002 a 2011.

ISSN	Periódico	n	PQ	n / PQ	Qualis	FI
0100-4042	Química Nova	1.872	489	3,8	B2	0,763
0103-5053	Journal of the Brazilian Chemical Society	1.803	498	3,6	A2	1,434
0039-9140	Talanta (Oxford)	610	168	3,6	A2	3,794
0040-4039	Tetrahedron Letters	534	117	4,6	B1	2,683
1600-5368	Acta Crystallographica. Section E	361	69	5,2	B5	0,347
0102-695X	Revista Brasileira de Farmacognosia	347	87	4,0	B5	-
0003-2670	Analytica Chimica Acta	318	127	2,5	A1	4,555
0013-4686	Electrochimica Acta	313	89	3,5	A2	3,832
0584-8547	Spectrochimica Acta. Part B, Atomic Spectroscopy	285	56	5,1	A2	2,876
0022-2860	Journal of Molecular Structure	279	107	2,6	B2	1,634
0026-265X	Microchemical Journal	256	90	2,8	B1	3,048
0021-9797	Journal of Colloid and Interface Science	254	97	2,6	A2	3,070
0100-4670	Eclética Química (Unesp)	251	95	2,6	B5	-
0021-9673	Journal of Chromatography	251	77	3,3	A1	4,531
1040-0397	Electroanalysis (New York, N.Y.)	246	69	3,6	B1	2,872
0277-5387	Polyhedron	237	100	2,4	B1	2,057
1089-5639	The Journal of Physical Chemistry. A	234	79	3,0	B1	2,946
0020-1693	Inorganica Chimica Acta	233	92	2,5	B2	1,846
0926-860X	Applied Catalysis. A, General	229	74	3,1	A2	3,903
1388-6150	Journal of Thermal Analysis and Calorimetry	221	56	3,9	B2	1,604
0009-2614	Chemical Physics Letters	221	75	2,9	B1	2,337
1381-1169	Journal of Molecular Catalysis. A, Chemical	212	82	2,6	B1	2,947
0021-8995	Journal of Applied Polymer Science	205	81	2,5	B3	1,289
0968-0896	Bioorganic & Medicinal Chemistry	201	79	2,5	B1	2,921

Vale salientar que no total informado de artigos dos bolsistas, alguns desses são contabilizados mais de uma vez devido a coautorias, sendo esta a forma

correta e usual quando se considera a produção cientifica de pesquisadores de mais de uma instituição (ALMEIDA; GUIMARÃES, 2013).

A média de autores nos artigos publicados no periódico “Química Nova” foi de 4,5 e no periódico “*Journal of the Brazilian Chemical Society*” foi de 5,3. Entre os periódicos com mais de 200 artigos no período de 2002 a 2011, o que tem a maior média de autores (8,0) é o “*Bioorganic & Medicinal Chemistry*”. Existe uma diferença nas médias de autores por artigo publicado em periódicos diferentes. A razão destas diferenças mereceria uma análise mais profunda por especialistas da área.

A Tabela 6.21 apresenta alguns índices numéricos de produtividade dos bolsistas no período de 2002 a 2011 de acordo com as suas categorias e a Tabela 6.22 pelo tempo decorrido após a conclusão do doutorado, sendo que para cada índice é apresentado o valor mínimo, médio e máximo. É importante observar que quando se consideram valores médios por categoria é necessário levar em conta que há uma quantidade muito maior de bolsistas da categoria 2 em comparação com as demais.

Analisando o número de artigos publicados em periódicos no período de 2002 a 2011, percebe-se que os bolsistas da categoria 1A foram os que mais publicaram nesse período, sendo que um desses bolsistas publicou 332 artigos, com uma média de 33,2 artigos por ano e com tempo decorrido após a conclusão do doutorado de 21 a 30 anos. Os bolsistas com tempo decorrido após a conclusão do doutorado até 10 anos publicaram 4.587 (14,0%) artigos, com média de 31,4 artigos por bolsista nesse período. Os bolsistas na faixa de 11 a 20 anos publicaram 13.917 (42,3%) artigos, com média de 44,3. Na faixa de 21 a 30 anos, os bolsistas publicaram 9.838 (29,9%) artigos, com média de 58,9. Na faixa de 31 a 40 anos foram publicados 3.225 (9,8%) artigos, com média de 62,0. E os bolsistas com tempo decorrido após a conclusão do doutorado acima de 40 anos publicaram 1.306 (4,0%) artigos, com média de 81,6 artigos por bolsista. É interessante observar que esses bolsistas, proporcionalmente, foram os que mais publicaram.

Tabela 6.21 - Índices numéricos de produtividade dos bolsistas PQ da área de Química por tópico no período de 2002 a 2011.

Tópicos		SR	1A	1B	1C	1D	2	2F
Número de artigos publicados em periódicos	Min	29	31	23	24	18	9	9
	Média	**78,1**	**99,1**	**80,8**	**65,4**	**49,2**	**35,1**	**20,5**
	Max	179	332	174	187	173	184	32
Média de autores de artigos publicados em periódicos	Min	3,6	3,0	3,0	3,3	2,6	2,2	5,3
	Média	**4,8**	**4,8**	**5,2**	**5,2**	**5,1**	**5,4**	**5,8**
	Max	6,6	7,5	7,7	7,8	7,5	10,4	6,2
Somatório do fator de impacto dos periódicos	Min	22,5	32,4	37,1	57,6	13,1	5,0	17,0
	Média	**110,1**	**251,3**	**170,1**	**141,3**	**100,9**	**68,9**	**39,6**
	Max	233,6	1.143,2	335,5	296,3	230,6	284,7	62,3
Fator de impacto por artigo publicado em periódico	Min	0,5	1,0	1,2	1,1	0,1	0,3	1,9
	Média	**1,4**	**2,5**	**2,1**	**2,2**	**2,0**	**2,0**	**1,9**
	Max	3,2	4,7	3,5	3,6	3,5	6,1	1,9
Número de citações na WoS	Min	763	479	447	11	207	22	48
	Média	**2.299,0**	**3.523,4**	**1.584,8**	**1.140,3**	**855,2**	**397,6**	**369,5**
	Max	4.362	13.368	3.346	2.389	5.933	1.964	691
Número de citações no ISI por artigo publicado em periódico	Min	4,5	5,4	4,4	0,2	3,4	1,0	5,3
	Média	**44,9**	**40,3**	**21,6**	**20,2**	**19,5**	**12,1**	**13,5**
	Max	123,0	121,7	45,2	60,3	109,9	64,1	21,6
Índice H	Min	13	13	14	1	9	4	4
	Média	**22,7**	**29,3**	**21,6**	**18,4**	**15,7**	**11,1**	**9,5**
	Max	33	51	33	28	28	26	15
Índice de Orientação (IO)	Min	7,0	13,5	10,5	12,0	5,3	0,5	5,0
	Média	**30,9**	**52,6**	**48,5**	**40,8**	**33,7**	**21,6**	**6,5**
	Max	57,5	124,5	96,0	121,0	80,8	101,3	8,0

Tabela 6.22 - Índices numéricos de produtividade dos bolsistas PQ da área de Química por tempo de titulação do doutorado no período de 2002 a 2011.

Tempo decorrido após conclusão do doutorado (anos)		0 a 10	11 a 20	21 a 30	31 a 40	> 40
Número de artigos publicados em periódicos	Min	9	10	9	27	23
	Média	**31,4**	**44,3**	**58,9**	**62,0**	**81,6**
	Max	97	184	332	135	225
Média de autores de artigos publicados em periódicos	Min	2,2	2,2	2,9	2,8	3,3
	Média	**5,5**	**5,3**	**5,2**	**4,8**	**4,5**
	Max	10,4	8,6	8,7	7,7	6,6
Somatório do fator de impacto dos periódicos	Min	17,0	5,0	19,5	23,1	22,5
	Média	**71,6**	**91,7**	**121,1**	**124,6**	**162,2**
	Max	284,7	643,3	1.143,2	289,5	480,6
Fator de impacto por artigo publicado em periódico	Min	0,7	0,1	0,7	0,5	0,6
	Média	**2,3**	**2,1**	**2,1**	**2,0**	**2,0**
	Max	6,1	4,4	4,7	3,6	3,3
Número de citações na WoS	Min	45	11	59	273	251
	Média	**363,0**	**659,7**	**1.153,1**	**1.578,7**	**2.593,4**
	Max	1.763	5.933	13.368	5.360	8.022
Número de citações no ISI por artigo publicado em periódico	Min	1,2	0,2	1,9	5,4	4,5
	Média	**11,9**	**15,0**	**19,3**	**26,8**	**32,2**
	Max	37,7	109,9	121,7	123,0	86,3
Índice H	Min	4	1	5	9	8
	Média	**10,9**	**13,3**	**16,9**	**19,4**	**24,0**
	Max	26	37	51	37	50
Índice de Orientação (IO)	Min	0,5	3,8	5,3	7,5	7,0
	Média	**10,9**	**29,4**	**39,5**	**39,8**	**33,0**
	Max	40,5	121,0	124,5	96,0	57,0

A média de autores por artigo em publicações em periódicos no período de 2002 a 2011 é 5,3. A maior média foi de 10,4 autores por artigo de um bolsista da categoria 2 e com tempo decorrido após a conclusão do doutorado até 10 anos. A menor média foi de 2,2, também de um bolsista da categoria 2. É interessante observar que a média de autores por artigo diminui à medida que o tempo decorrido após a conclusão do doutorado desses aumenta. Dos 32.873 artigos publicados em periódicos nesse período, em 8.352 (25,4%)

tinham como primeiro autor um dos bolsistas e em 10.191 (31,0%) deles um bolsista como último autor.

Todos os 695 bolsistas possuem pelo menos um artigo publicado em periódico indexado no JCR® no período de 2002 a 2011, sendo possível calcular o somatório do FI de todos os periódicos que esses bolsistas publicaram. Dos 1.979 periódicos utilizados, 1.225 (61,9%) possuem FI no JCR® de 2011. Do total de 32.873 publicações nesses periódicos, 27.697 (84,3%) possuem FI. O maior valor encontrado foi de um bolsista da categoria 1A, com somatório de 1.143,2 com o FI de suas publicações variando de 0,493 a 40,197. O menor valor foi de 5,0 de um bolsista da categoria 2 e a média geral foi de 98,6. Exceto pela categoria SR, observa-se que o valor médio diminui significativamente de uma categoria para outra, com os maiores valores nas principais categorias. Observa-se como esperado que quanto maior o tempo decorrido após a conclusão do doutorado, maior o valor médio do somatório do FI.

Dividindo o somatório do FI dos periódicos pelo número de artigos publicados no período de 2002 a 2011, é possível calcular o FI por artigo publicado do bolsista. É interessante observar que os bolsistas da categoria 2 publicam em periódicos em que a média do FI desses periódicos é maior que a média dos bolsistas da categoria SR, sendo que o mesmo ocorre com os bolsistas da categoria 1C também em relação aos da categoria 1B. É curioso observar que quanto maior o tempo decorrido após a conclusão do doutorado, menor o valor médio do FI por artigo publicado do bolsista. Isso dá um indicativo que os bolsistas com menor tempo de doutoramento estão direcionando suas publicações para periódicos com maior FI, possivelmente, motivados pela classificação desses no Qualis/CAPES da área de avaliação da “Química”. Todavia, é imperativo notar que uma simples comparação do valor médio do FI com o tempo decorrido após a conclusão do doutorado é insuficiente para inferir o aumento ou o declínio na produtividade da publicação de artigos em periódicos indexados no JCR®, uma vez que são considerados períodos de gerações distintas.

Em relação ao número de citações na WoS dos artigos dos bolsistas no período de 2002 a 2011, percebe-se que há uma grande variação. O maior número de citações na WoS foi de um bolsista da categoria 1A, com 13.368 citações. O menor valor foi de um bolsista da categoria 1C, com 11 citações. Dos 695 bolsistas, 680 (97,8%) informaram o número de citações na WoS em seus currículos Lattes, com valor médio de 821,9 nesse período. Considerando o tempo decorrido após a conclusão do doutorado, o valor médio aumenta significativamente em cada faixa. É interessante observar que o maior valor ocorreu na faixa de 21 a 30 anos, tendo ocorrido o mesmo em relação ao somatório do FI.

Dividindo o total de citações pelo número de artigos publicados é possível calcular o número de citações na WoS por artigo de cada bolsista. Percebe-se que o número médio de citações por artigo está fortemente correlacionado com a categoria do bolsista, pois os valores são maiores nas principais categorias. Considerando o tempo decorrido após a conclusão do doutorado, percebe-se também que quanto maior o tempo de doutorado maior o número médio de citações por artigo. Analisando estes dados com os de valor médio dos fatores de impacto dos periódicos em que o bolsista publica temos uma indicação interessante: aparentemente, a maioria das citações de artigos publicados pelos bolsistas com mais anos de doutorado não vem de artigos publicados em periódicos com fatores de impacto elevados.

O índice H foi informado em 676 (97,3%) dos currículos Lattes dos bolsistas, com valor médio de 14,3. O maior valor informado foi de um bolsista da categoria 1A, com índice H de 51. O menor índice H informado foi 1 de um bolsista da categoria 1C. Percebe-se que o maior valor médio também ocorre na faixa de 21 a 30 anos.

Outro atributo considerado neste estudo foi o número de orientações concluídas de iniciação científica, mestrado e doutorado. Dos 695 bolsistas, 694 (99,9%) informaram as orientações concluídas no período de 2002 a 2011. Na modalidade iniciação científica, 648 (93,4%) bolsistas orientaram em média 12,1 alunos nesse período. Na modalidade mestrado, 646 (93,1%) orientaram

em média 6,3 alunos e na modalidade doutorado, 477 (68,7%) orientaram 5,8 alunos nesse período.

Outra forma de analisar as orientações é utilizando o IO (SANTOS et al., 2010). O IO é calculado pelo somatório do número de alunos de cada modalidade de orientação multiplicado por seus respectivos pesos: iniciação científica (0,5), mestrado (1,5) e doutorado (3,0). Às co-orientações foram atribuídos pesos na metade do valor das orientações concluídas como orientador principal. O bolsista com maior IO no período de 2002 a 2011 é da categoria 1A, com IO de 124,5 e tempo decorrido após a conclusão do doutorado de 21 a 30 anos. A média do IO dos bolsistas nesse período foi de 28,8.

Considerando o tempo decorrido após a conclusão do doutorado, foi possível verificar que os bolsistas na faixa de 31 a 40 anos (7,5%) foram os que mais contribuíram para a formação de recursos humanos, com a média do IO de 39,8. O IO também foi significativo para os bolsistas PQ com o tempo de doutorado de 21 a 30 anos (24,0%), pois sua média do IO foi de 39,5. Apenas após os 40 anos (2,3%) de conclusão do doutorado observa-se uma queda no IO, com a média de 33,0. Isso talvez se deva ao fato que não havia muitos programas de doutorado antigamente e mais recentemente, os bolsistas nesta faixa estão se aposentando e não orientam mais. Os bolsistas na faixa de 11 a 20 anos (45,2%) possuem média de 29,4. E os bolsistas com até 10 anos (21,01%) possuem uma média bem menor (10,9). Esse resultado era de se esperar, pois recém-doutores não atuam imediatamente na pós-graduação.

A linguagem LattesMiner permite identificar os relacionamentos nas orientações tidas entre os bolsistas. A Figura 6.13 ilustra a rede de orientações concluídas de mestrado e doutorado entre os bolsistas de acordo com as suas categorias. Essa rede foi gerada a partir do sistema SUCUPIRA, que foi desenvolvido utilizando a linguagem LattesMiner. As cores dos vértices representam a categoria do bolsista, sendo que “roxo” é para a categoria SR, “azul” para a categoria 1A, “verde” para a categoria 1B, “amarelo” para 1C, “laranja” para 1D, “vermelho” para a categoria 2 e “rosa” para a categoria 2F. A rede apresenta a relação orientador-orientado por categoria, do centro para as

extremidades. Estas identificam as orientações concluídas de mestrado (M) e doutorado (D) entre os bolsistas. As cores das arestas representam o número de orientações entre as categorias, indicando quão intenso é um relacionamento.

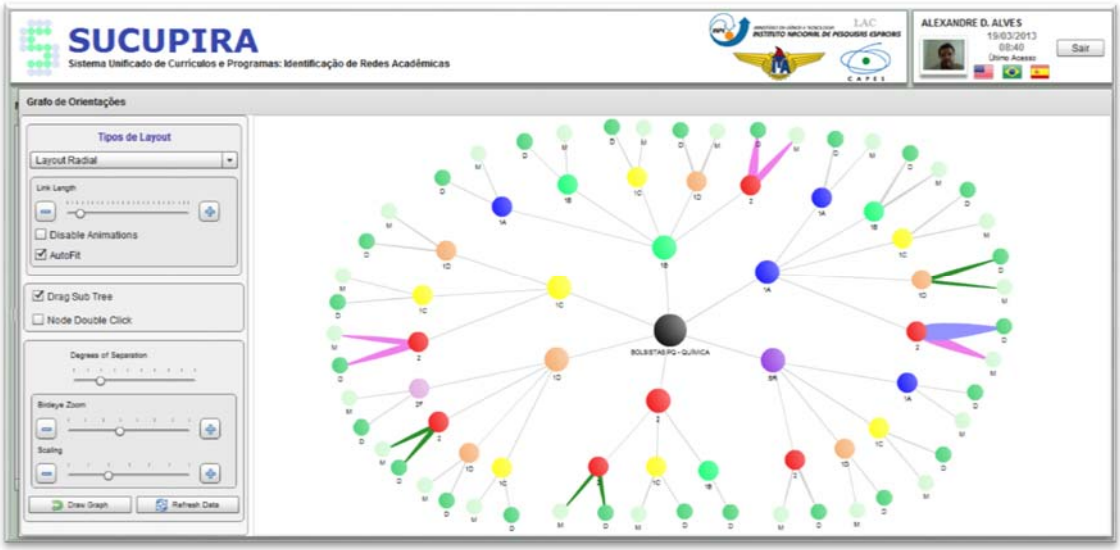


Figura 6.13 - Rede de orientações concluídas de mestrado (M) e doutorado (D) entre os bolsistas PQ da área de Química de acordo com a categoria.

Há relacionamentos nessa rede que chamam a atenção e estão destacados em “azul”, representando que o relacionamento ocorreu mais de 50 vezes; em “rosa”, representando que o relacionamento ocorreu entre 25 e 50 vezes e em “verde escuro”, representando que o relacionamento ocorreu entre 10 e 25 vezes. O principal relacionamento de orientador-orientado ocorreu entre os bolsistas da categoria 1A e os da categoria 2 nas orientações concluídas de doutorado, sendo que essa relação ocorreu 81 vezes.

A linguagem LattesMiner também permite identificar os contatos (todos os “links” identificados para outros currículos Lattes) contidos no currículo Lattes de um determinado pesquisador. Todo contato contém o ID do pesquisador, o que permite identificar os relacionamentos entre os pesquisadores. Deve-se salientar que nem todo relacionamento de um pesquisador possui “link” para outro no currículo Lattes. Estes relacionamentos não “certificados” no currículo Lattes não são contabilizados.

A Figura 6.14 ilustra a rede de contatos nos artigos publicados em periódicos no período de 2002 a 2011 entre os bolsistas por categoria. Ao todo foram identificados 27.328 contatos nesta condição, sendo que 5.672 (20,8%) ocorreram entre os bolsistas da categoria 2 com eles mesmos. Esse relacionamento é destacado em “azul”, representando que o relacionamento ocorreu mais de 5.000 vezes. Também são destacados os relacionamentos em “rosa”, representando que o relacionamento ocorreu entre 1.000 e 5.000 vezes e em “verde escuro”, representando que o relacionamento ocorreu entre 100 e 1.000 vezes.

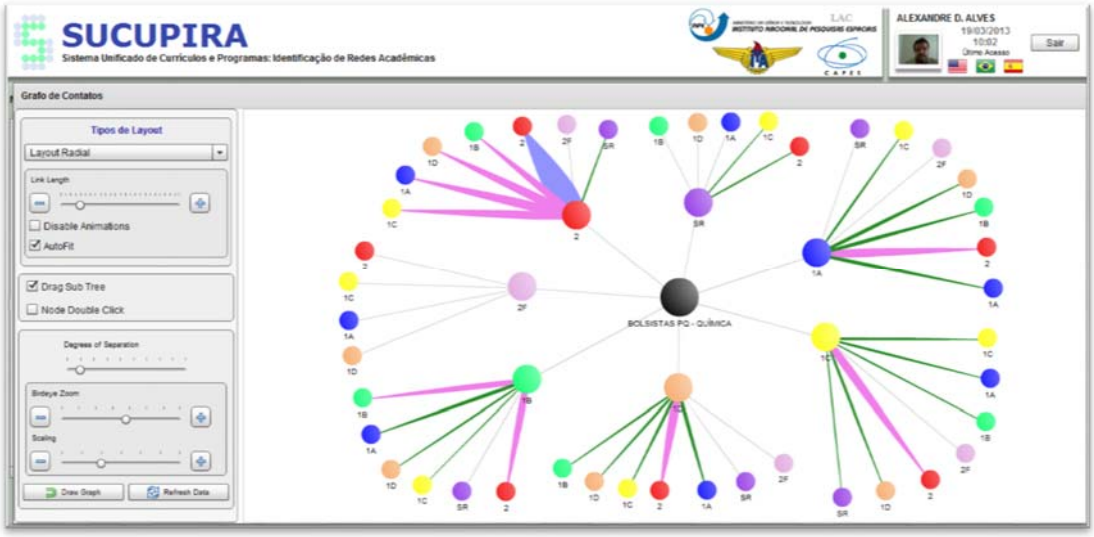


Figura 6.14 - Rede de contatos identificados nos artigos publicados em periódicos no período de 2002 a 2011 entre os bolsistas PQ da área de Química de acordo com a categoria.

Também é importante destacar que os bolsistas de todas as categorias se relacionam entre eles em praticamente todas as categorias, o que denota que a desejável cooperação acadêmica é alcançada pela área de Química.

Este estudo apresentou um perfil dos bolsistas PQ do CNPq da área de Química baseado em informações extraídas dos currículos Lattes de dezembro de 2012. Houve um aumento em torno de 15% no número de bolsistas desde o estudo realizado por Santos et al. (2010) com os pesquisadores com bolsas ativas em 2009. A grande maioria dos bolsistas ainda é do sexo masculino, uma vez que a porcentagem continua praticamente a mesma, em torno de

68%. A porcentagem de bolsistas da categoria 1 era de 36,8 e agora está em 35,5. A porcentagem de bolsistas da categoria 2 era de 62,3 e agora está em 63,2. A região Sudeste possuía 63,7% do total de bolsistas e atualmente é de 60,7%. São Paulo continua sendo o estado com o maior número de bolsistas. Entretanto, a porcentagem diminuiu de 41,2 para 37,3. A média nacional do número de bolsistas por milhão de habitantes aumentou de 3,2 para 3,6 (aumento de 12,5%). O estado com a maior razão é o Rio Grande do Sul com 6,8 bolsistas por milhão de habitantes (o estado de São Paulo tem 6,3). Nesse ponto houve uma mudança, pois no estudo de Santos et al. (2010) o estado de São Paulo possuía a maior razão com 6,0 bolsistas por milhão de habitantes e o estado do Rio Grande era o segundo (5,1).

É interessante destacar que as cinco instituições com mais bolsistas continuam sendo as mesmas e todas da região Sudeste, possuindo juntas 285 bolsistas. Porém, a porcentagem diminuiu de 45,9 para 41,0. O predomínio ainda continua sendo da Universidade de São Paulo com 15,0% dos bolsistas, tendo diminuído a porcentagem que era de 17,4. Outra questão interessante é que a Universidade de São Paulo e a Universidade Estadual de Campinas, em conjunto, continuam tendo mais da metade dos bolsistas da categoria 1A.

Em relação às publicações em periódicos, nota-se que não houve uma queda significativa nos valores das médias quando consideramos os bolsistas por categoria, exceto em uma. No estudo realizado por Santos et al. (2010) a média de publicações por ano em cada categoria era a seguinte: SR (8,0), 1A (9,0), 1B (8,3), 1C (6,9), 1D (5,6) e 2 (4,2). A média neste estudo foi a seguinte: SR (7,8), 1A (9,9), 1B (8,1), 1C (6,5), 1D (4,9) e 2 (3,5). Comparando, percebe-se que houve um aumento (10,0%) na categoria 1A e uma queda mais acentuada na categoria 2 (16,7%).

No período de 2002 a 2011, os artigos dos bolsistas foram publicados em periódicos de 149 categorias diferentes de um total de 226 constantes no JCR® de 2011. As categorias mais utilizadas foram: “Química Multidisciplinar” em 5.247 (18,9%) artigos, “Físico-Química” em 4.641 (16,8%) artigos, “Química Analítica” em 3.951 (14,3%) artigos, “Química Orgânica” em 2.718 (9,8%)

artigos e “Farmacologia e Farmácia” em 1.761 (6,4%) artigos. A categoria “Química Inorgânica e Nuclear” ocupa a sétima posição com 1.727 (6,2%) artigos publicados nessa categoria. Vale lembrar que um mesmo periódico pode ser classificado em mais de uma categoria do JCR®.

É interessante observar que a distribuição dos artigos entre categorias não é igual às subáreas de atuação informadas pelos bolsistas, pois 40,1% declaram atuar em “Físico-Química”, 38,8% em “Química Orgânica”, 30,6% em “Química Analítica” e 25,2% em “Química Inorgânica”. Percebe-se, então, que os bolsistas publicam mais em “Química Analítica” do que em “Química Orgânica” apesar de declararem justamente o contrário. Isso decorre da atuação interdisciplinar dos bolsistas PQ da área de Química.

Considerando o somatório do FI por ano, percebe-se também que houve em geral uma queda nos valores médios de acordo com a categoria dos bolsistas. No estudo realizado por Santos et al. (2010) o valor médio do somatório por ano em cada categoria era a seguinte: SR (13,5), 1A (19,8), 1B (17,1), 1C (12,8), 1D (11,1) e 2 (7,3). Neste estudo, o valor médio do somatório por ano em cada categoria é o seguinte: SR (11,0), 1A (25,1), 1B (17,0), 1C (14,1), 1D (10,1) e 2 (6,9). Houve um aumento no valor médio nas categorias 1A e 1C. O maior valor médio por ano no estudo realizado por Santos et al. (2010) era 57 e de um bolsista da categoria 1B. Neste estudo, o maior valor foi de 114,3 de um bolsista da categoria 1A.

No caso do índice H, como era de se esperar, ocorreu um aumento em todas as categorias dos bolsistas. No estudo realizado por Santos et al. (2010) o valor médio do índice H era o seguinte: SR (17,5), 1A (24), 1B (17,4), 1C (13,5) e 1D (12,3). Para a categoria 2 não foi informado o valor médio do índice H. No nosso estudo, o valor médio do índice H foi o seguinte: SR (22,7), 1A (29,3), 1B (21,6), 1C (18,4) e 1D (15,7). O maior índice H era 47 e passou a ser 51.

Em relação ao IO, verificou-se que os bolsistas com tempo decorrido após a conclusão do doutorado de 21 a 30 anos e de 31 a 40 anos foram os que mais contribuíram para a formação de recursos humanos. Segundo Santos et al.

(2010), essa faixa ficava entre os 10 e 30 anos. Percebe-se que há uma mudança no perfil dos bolsistas considerando-se diferentes janelas de tempo, pois o estudo realizado por Santos et al. (2010) considerou os pesquisadores com bolsas PQ ativas em 2009.

Uma última informação que merece ser destacada é o fato de que 34 (4,9%) dos bolsistas PQ da área de Química serem membros titulares da Academia Brasileira de Ciências, o que representa 7,6% do total de membros da Academia.

6.5. Bases de dados

Neste estudo de caso é aplicado a Lei de Benford para analisar dados das bases de dados JCR® e Scopus considerando o número de artigos publicados em periódicos indexados nessas bases. Também foi feita uma análise investigando a conformidade com a Lei de Benford do número de artigos publicados de acordo com o país de origem e a categoria dos periódicos indexados nas duas bases de dados.

6.5.1. Coleta de dados

Neste estudo foram utilizados dados disponíveis no JCR® nas edições "*Science*" e "*Social Sciences*" no período de 2007 a 2011. Todos os periódicos indexados no JCR® com pelo menos um artigo publicado foram incluídos. Também foi considerado o país de origem e a categoria do periódico.

Inicialmente, foi identificado o primeiro digito significativo do número de artigos publicados de cada periódico indexado no JCR®, para cada ano e edição, separadamente, para calcular a frequência de cada dígito e comparar com o número previsto pela Lei de Benford.

Em seguida, foi realizado o teste do qui-quadrado:

$$x^2(n-1) = \sum_{i=1}^{n} \frac{\left(N_o(d) - N_e(d)\right)^2}{N_e(d)}, \qquad (2)$$

para testar a *Hipótese Nula*, H_0 de que a distribuição observada do primeiro dígito significativo (*d*), em cada caso considerado, é o mesmo que o número esperado (N_e) com base na Lei de Benford.

Para n = 9 tem-se n – 1 = 8 graus de liberdade, e $x^2(8) = 15{,}507$ para um nível de confiança de 95%. Este é o valor crítico para a aceitação ou rejeição da *Hipótese Nula*, ou seja, se o valor calculado de x^2 for menor que o valor crítico, então H_0 é aceita e conclui-se que os dados estão em conformidade com a Lei de Benford, caso contrário, rejeita-se H_0.

Alternativamente, pode-se testar cada uma das nove proporções separadamente. O estatística Z é um teste para verificar se a proporção observada para um dígito difere significativamente do valor esperado com base na Lei de Benford (NIGRINI, 2012). A fórmula da estatística Z leva em conta o valor absoluto da diferença entre os valores observados e esperados, a cardinalidade do conjunto de dados e o valor da proporção esperada é dado pela seguinte equação:

$$Z = \frac{|P_o - P_e| - \left(\frac{1}{2N}\right)}{\sqrt{\frac{P_e(1 - P_e)}{N}}}, \qquad (3)$$

em que P_o denota o valor da proporção observada, P_e o valor da proporção esperada e *N* o número total de observações. O termo (1 / 2*N*) é de correção de continuidade e é considerado apenas quando é menor que o outro termo no numerador. Para o nível de significância de 5%, o nível de corte é 1,96. Quando a estatística Z excede 1,96 então a diferença entre os valores das proporções observadas e esperadas é significativa no nível de 0,05, o que significa que existe apenas uma probabilidade de 5% que a diferença seja devido ao acaso.

Também foram utilizados dados disponíveis na Scopus. Utilizando a Scopus foi testado o número de artigos publicados em periódicos de alguns países e categorias do JCR®. Da mesma forma, todos os periódicos indexados na

Scopus com pelo menos um artigo publicado foram considerados. Somente periódicos presentes em ambas as bases de dados foram considerados.

Utilizando a distribuição binomial, a raiz quadrada do erro médio, $\Delta[N(d)]$:

$$\Delta N(d) = \sqrt{NP(d)\big(1 - P(d)\big)}\,, \qquad (4)$$

também foi calculada, em que *N* é o número total de pontos considerados e *P(d)* é a previsão pela Lei de Benford.

6.5.2. Resultados e discussões

Campanario e Coslado (2011) observaram que o número de artigos publicados, citações recebidas e o fator de impacto de periódicos indexados no JCR® Edição "*Science*" no período de 1998 a 2007 nem sempre estão em conformidade com a Lei de Benford. Um resumo dessa análise é apresentada na Tabela 6.23.

Tabela 6.23 - Valores x^2 para o número de artigos publicados, citações recebidas e fator de impacto dos periódicos indexados no JCR® Edição "*Science*" no período de 1998 a 2007 (CAMPANARIO; COSLADO, 2011).

Ano	Artigos	Citações	Fator de Impacto
1998	27,8*	15,1	6,6
1999	27,4*	7,1	11,3
2000	16,2*	4,5	22,2*
2001	38,1*	5,2	20,2*
2002	57,9*	3,1	24,9*
2003	43,5*	3,5	12,5
2004	31,3*	3,0	16,7*
2005	41,5*	11,2	16,3*
2006	27,8*	9,7	39,3*
2007	31,3*	8,4	40,4*

* Denota diferença significativa entre os valores observadores e os esperados para $p = 0{,}05$

Pode-se observar que os valores de x^2 para o número de artigos são maiores que o valor crítico (15,507) em todos os anos, ou seja, todos os valores não estão em conformidade com a Lei de Benford.

Estendemos essa análise e os dados dos anos seguintes foram investigados. Foi analisado o número de artigos publicados em periódicos indexados no JCR® Edição "*Science*" no período de 2007 a 2011 e o resultado é apresentado na Tabela 6.24. Apesar do valor de x^2 para 2007 já ter sido calculado por eles, calculou-se novamente para verificar a compatibilidade dos nossos resultados com os deles. Foi observada uma pequena diferença, provavelmente, devido ao fato de considerarmos um número maior de periódicos, com a atualização do JCR®.

Tabela 6.24 - Frequência de ocorrência de *d* como primeiro dígito significativo, obtido a partir do número de artigos publicados em periódicos indexados no JCR® Edição "*Science*" no período de 2007 a 2011.

Ano	*d*	1	2	3	4	5	6	7	8	9	Total	x^2
2007	*No*	1.730	1.047	840	643	515	450	389	354	307	6.275	31,6*
	Ne	1.888,8	1105,0	783,7	608,0	497,0	419,8	364,0	321,3	287,4		
	ΔN(d)	36,34	30,17	26,19	23,43	21,39	19,79	18,52	17,46	16,56		
	valor Z	4,36**	1,90	2,12**	1,47	0,82	1,49	1,33	1,86	1,17		
2008	*No*	1.790	1.090	841	670	595	415	411	343	332	6.487	41,1*
	Ne	1.952,6	1.142,4	810,2	628,6	513,8	434,0	376,2	332,1	297,1		
	ΔN(d)	36,94	30,68	26,63	23,83	21,75	20,12	18,82	17,75	16,84		
	valor Z	4,39**	1,69	1,13	1,71	3,72**	0,93	1,82	0,60	2,06**		
2009	*No*	2.018	1.204	955	766	631	494	430	402	316	7.216	34,3*
	Ne	2.172,0	1.270,7	901,3	699,2	571,5	482,8	418,5	369,5	330,5		
	ΔN(d)	38,96	32,36	28,08	25,13	22,94	21,23	19,86	18,72	17,76		
	valor Z	3,95**	2,05**	1,88	2,63**	2,58**	0,49	0,56	1,73	0,77		
2010	*No*	2.172	1.321	1.065	798	720	554	472	406	367	7.875	44,9*
	Ne	2.370,4	1.386,8	983,6	763,1	623,7	526,8	456,7	403,2	360,7		
	ΔN(d)	40,71	33,80	29,34	26,25	23,96	22,17	20,74	19,56	18,55		
	valor Z	4,87**	1,93	2,75**	1,31	4,00**	1,19	0,71	0,14	0,33		
2011	*No*	2.246	1.371	1.070	880	694	591	495	428	353	8.128	43,2*
	Ne	2.446,5	1.431,3	1.015,2	787,6	643,7	543,8	471,4	416,2	372,3		
	ΔN(d)	41,35	34,34	29,81	26,67	24,35	22,53	21,07	19,87	18,85		
	valor Z	4.84**	1,74	1,81	3,44**	2,05**	2,06**	1,10	0,59	0,98		

* Denota diferença significativa entre os valores observadores e os esperados para *p* = 0,05

** Denota diferença significativa entre a proporção observada e a esperada no nível de 0,05

Os valores de x^2 em todos os anos são significativamente maiores do que o valor crítico. Além disso, observa-se que os valores de *Z* para o dígito 1 é maior do que o nível de corte (1,96) em todos os anos. O mesmo ocorreu com o dígito 5, exceto em 2007.

Campanario e Coslado (2011) consideraram apenas os periódicos indexados no JCR® Edição "*Science*" e nós estendemos o cálculo para JCR® Edição "*Social Sciences*". O resultado é apresentado na Tabela 6.25. Como pode ser observado, o resultado é ainda pior: todos os anos não estão em conformidade com a Lei de Benford e os valores de *Z* são maiores do que o nível de corte em quase todos os dígitos. Campanario e Coslado (2011) mencionaram no seu estudo que não têm explicação para essas diferenças.

Tabela 6.25 - Frequência de ocorrência de *d* como primeiro dígito significativo, obtido a partir do número de artigos publicados em periódicos indexados no JCR® Edição "*Social Sciences*" no período de 2007 a 2011.

Ano	*d*	1	2	3	4	5	6	7	8	9	Total	x^2
2007	*No*	406	492	336	227	125	98	55	51	43	1.833	263,1*
	Ne	551,7	322,8	228,9	177,6	145,2	122,6	106,3	93,9	84,0		
	ΔN(d)	19,64	16,31	14,15	12,66	11,56	10,70	10,01	9,43	8,95		
	valor Z	7,40**	10,35**	7,52**	3,86**	1,70	2,26**	5,08**	4,48**	4,51**		
2008	*No*	438	492	392	242	134	91	58	48	57	1.952	287,2*
	Ne	587,6	343,7	243,8	189,2	154,6	130,6	113,2	99,9	89,4		
	ΔN(d)	20,27	16,83	14,61	13,07	11,93	11,04	10,33	9,74	9,24		
	valor Z	7,36**	8,78**	10,10**	4,00**	1,68	3,55**	5,30**	5,28**	3,45**		
2009	*No*	507	563	425	262	158	118	75	54	53	2.215	289,3*
	Ne	666,7	390,1	276,7	214,6	175,4	148,2	128,5	113,4	101,4		
	ΔN(d)	21,59	17,93	15,56	13,92	12,71	11,76	11,00	10,37	9,84		
	valor Z	7,38**	9,62**	9,50**	3,36**	1,33	2,53**	4,81**	5,67**	4,87**		
2010	*No*	622	722	487	302	180	141	88	72	54	2.668	364,5*
	Ne	803,1	469,8	333,2	258,5	211,3	178,5	154,8	136,6	122,2		
	ΔN(d)	23,69	19,67	17,08	15,28	13,95	12,91	12,07	11,38	10,80		
	valor Z	7,62**	12,79**	8,97**	2,81**	2,21**	2,87**	5,49**	5,62**	6,26**		
2011	*No*	674	794	523	328	205	137	91	76	63	2.891	409,8*
	Ne	870,2	509,1	361,1	280,1	229,0	193,4	167,7	148,0	132,4		
	ΔN(d)	24,66	20,48	17,78	15,90	14,52	13,43	12,57	11,85	11,24		
	valor Z	7,94**	13,89**	9,07**	2,98**	1,61	4,17**	6,06**	6,03**	6,12**		

* Denota diferença significativa entre os valores observadores e os esperados para *p* = 0,05

** Denota diferença significativa entre a proporção observada e a esperada no nível de 0,05

Mir (2012) observou que os dados das três principais denominações cristãs seguem a Lei de Benford. No entanto, quando o cristianismo é considerado como um único grupo religioso, a distribuição aderente dos dígitos significativos dos dados desvia das previsões da Lei de Benford. Inspirado por esta observação, analisamos os periódicos de acordo com seu país de origem e sua categoria no JCR®.

A Tabela 6.26 apresenta o número total de países que estão em conformidade (SIM) ou não (NÃO) com a Lei de Benford considerando os valores de x^2 para o número de artigos publicados em periódicos indexados no JCR® Edição "*Science*" no período de 2007 a 2011, com destaque para os três países com os maiores valores de x^2 que não estão em conformidade com a Lei de Benford e seu respectivo número de periódicos e artigos considerados em cada ano.

Tabela 6.26 - Total de países que estão em conformidade (SIM) ou não (NÃO) com a Lei de Benford considerando os valores x^2 para o número de artigos publicados em periódicos indexados no JCR® Edição "*Science*" no período de 2007 a 2011.

Ano	SIM (%)	NÃO (%)	País	# de Periódicos (# de artigos)	x^2
2007	56 (81,16)	13 (18,84)	Turquia Eslováquia Croácia	7 (464) 10 (540) 10 (578)	41,9* 27,1* 22,7*
2008	61 (84,72)	11 (15,28)	Ucrânia Estados Unidos Uruguai	4 (309) 2.461 (405.322) 1 (8)	44,6* 26,9* 18,8*
2009	71 (92,21)	6 (7,79)	Polônia Estados Unidos Finlândia	101 (7.642) 2.551 (413.409) 13 (927)	26,2* 23,7* 22,9*
2010	68 (81,93)	15 (18,07)	Polônia Turquia Singapura	120 (8.936) 47 (3.396) 50 (4.790)	31,2* 29,8* 23,7*
2011	67 (82,72)	14 (17,28)	Polônia Turquia Suíça	124 (9.721) 52 (3.953) 170 (26.609)	40,5* 29,5* 25,3*

* Denota diferença significativa entre os valores observadores e os esperados para p = 0,05

É possível observar que a maioria dos países está em conformidade com a Lei de Benford. “Polônia” e “Turquia” são os países que apareceram mais vezes na lista dos três principais países que não estão em conformidade com a Lei de Benford. No caso da “Turquia” é interessante notar que o número de periódicos indexados no JCR® aumentou muito de um ano para o outro. Além disso, pode-se observar que os valores de x^2 diminuíram a medida que o número de periódicos e artigos aumentou. É importante observar que o número de periódicos indexados no JCR® é muito pequeno para alguns países, não sendo suficiente para o uso de teste do qui-quadrado para a aderência dos dados à Lei de Benford. De acordo com Nigrini (2012), a regra para a Lei de Benford para o teste do qui-quadrado para primeiro dígito significativo é que o número esperado de observações para cada célula deve ser pelo menos 5, por isso, o número de observações deve ser pelo menos 100 (100 vezes 0,0458, que está perto o suficiente para 5).

O resultado é muito semelhante para os periódicos indexados no JCR® Edição “*Social Sciences*”. Apenas alguns países não estão em conformidade com a Lei de Benford, como mostra a Tabela 6.27. No entanto, os valores de x^2 são muito menores do que os valores apresentados quando periódicos foram considerados como um único grupo. É interessante observar que os “Estados Unidos” e “Inglaterra” não estão em conformidade com a Lei de Benford em todos os anos.

Outras análises realizadas consideraram a categoria dos periódicos no JCR® Edição “*Science*” no período de 2007 a 2011. O resultado é apresentado na Tabela 6.28. É possível verificar que o percentual de categorias que estão em conformidade com a Lei de Benford é maior em comparação com a porcentagem de países que estão em conformidade com a Lei de Benford em quase todos os anos, exceto em 2009. “*Mathematics*” e “*Nursing*” apareceram mais vezes na lista das três principais categorias que não estão em conformidade com a Lei de Benford.

Tabela 6.27 - Total de países que estão em conformidade (SIM) ou não (NÃO) com a Lei de Benford considerando os valores x^2 para o número de artigos publicados em periódicos indexados no JCR® Edição "*Social Sciences*" no período de 2007 a 2011.

Ano	SIM (%)	NÃO (%)	País	# de Periódicos (# de artigos)	x^2
2007	38 (92,68)	3 (7,32)	Estados Unidos Inglaterra Holanda	999 (44.124) 464 (23.285) 116 (6.979)	134,5* 123,2* 18,2*
2008	42 (95,45)	2 (4,55)	Estados Unidos Inglaterra	1.042 (46.559) 484 (25.237)	157,2* 135,8*
2009	45 (88,24)	6 (11,76)	Estados Unidos Inglaterra Turquia	1.067 (48.548) 545 (27.829) 7 (274)	199,2* 89,1* 25,1*
2010	47 (90,38)	5 (9,62)	Estados Unidos Inglaterra Finlândia	1.199 (53.586) 716 (35.160) 2 (87)	187,8* 124,5* 19,1*
2011	44 (83,02)	9 (16,98)	Estados Unidos Inglaterra Finlândia	1.254 (57.695) 828 (40.470) 2 (114)	185,0* 153,5* 23,0*

* Denota diferença significativa entre os valores observadores e os esperados para $p = 0,05$

Tabela 6.28 - Total de categorias de periódicos que estão em conformidade (SIM) ou não (NÃO) com a Lei de Benford considerando os valores x^2 para o número de artigos publicados em periódicos indexados no JCR® Edição "*Science*" no período de 2007 a 2011.

Ano	SIM (%)	NÃO (%)	Categoria	# de Periódicos (# de artigos)	x^2
2007	156 (90,70)	16 (9,30)	*Statistics & Probability* *Mathematics* *History & Philosophy of Science*	90 (6.512) 199 (16.141) 35 (1.007)	31,0* 25,1* 22,4*
2008	151 (87,28)	22 (12,72)	*Mathematics, Interdisciplinary Applications* *Nursing* *Mathematics*	74 (6.103) 61 (3.706) 208 (17.228)	27,5* 25,9* 24,9*
2009	157 (90,75)	16 (9,25)	*Nursing* *Entomology* *Statistics & Probability*	72 (4.232) 72 (4.988) 100 (6.844)	32,1* 26,1* 25,4*
2010	159 (91,38)	15 (8,62)	*Nursing* *Mathematics* *Mathematics, Applied*	88 (5.246) 269 (20.049) 232 (20.998)	31,1* 29,8* 27,7*
2011	157 (89,20)	19 (10,80)	*Mathematics, Applied* *Mathematics* *Nursing*	240 (21.860) 281 (20.961) 98 (5.601)	36,0* 31,9* 29,1*

* Denota diferença significativa entre os valores observadores e os esperados para $p = 0,05$

Para os periódicos indexados no JCR® Edição "*Social Sciences*", o resultado é significativamente pior em comparação com os resultados do país de origem do periódico, como mostra a Tabela 6.29. Em alguns casos, o número de periódicos em conformidade com a Lei de Benford foi menor do que o número de periódicos em desacordo. "*Sociology*" é a categoria que não está em conformidade com a Lei de Benford em todos os anos.

Tabela 6.29 - Total de categorias de periódicos que estão em conformidade (SIM) ou não (NÃO) com a Lei de Benford considerando os valores x^2 para o número de artigos publicados em periódicos indexados no JCR® Edição "*Social Sciences*" no período de 2007 a 2011.

Ano	SIM (%)	NÃO (%)	Categoria	# de Periódicos (# de artigos)	x^2
2007	38 (69,09)	17 (30,91)	*Sociology* *Economics* *Political Science*	94 (3.099) 191 (9.245) 89 (3.672)	46,9* 44,2* 36,1*
2008	30 (53,57)	26 (46,43)	*Economics* *Law* *Sociology*	206 (10.724) 101 (3.049) 98 (3.342)	48,3* 41,3* 39,7*
2009	32 (58,18)	23 (41,82)	*Sociology* *Law* *Economics*	111 (3.581) 113 (3.309) 246 (11.856)	49,5* 46,6* 40,0*
2010	26 (46,43)	30 (53,57)	*Sociology* *Education & Educational Research* *Political Science*	128 (4.159) 180 (6.862) 140 (5.078)	59,1* 50,8* 46,5*
2011	26 (46,43)	30 (53,57)	*Sociology* *Economics* *Political Science*	132 (4.553) 314 (15.327) 145 (5.097)	66,6* 51,7* 46,0*

* Denota diferença significativa entre os valores observadores e os esperados para p = 0,05

É interessante observar que os valores de x^2 observados para os periódicos no JCR® Edição "*Social Sciences*" são sempre maiores que os apresentados para os periódicos indexados no JCR® Edição "*Science*".

Também comparamos o número de artigos publicados informados no JCR® e na Scopus. A comparação foi limitada a periódicos de alguns países e de algumas categorias. Entretanto, para realizar a comparação, foram considerados apenas periódicos que estavam presentes em ambas as bases de dados.

A análise realizada mostrou que existem alguns casos em que os dados de periódicos da Scopus estão em conformidade com a Lei de Benford, mas os dados correspondentes do JCR® não. Também foi observado o oposto, isto é, em que dados de periódicos do JCR® estão em conformidade com a Lei de Benford, mas dados correspondentes da Scopus não. Na Tabela 6.30 são apresentados esses resultados com 8 exemplos.

Os exemplos apresentados na Tabela 6.30 foram cuidadosamente escolhidos de modo que o número total de periódicos fosse maior do que 100. Em cada exemplo, o número de periódicos indexados nas duas bases é apresentado. Além desse valor, também é apresentado entre parênteses o número de periódicos indexados no JCR®. As colunas “Min” e “Max” indicam o mínimo e máximo de artigos publicados em periódicos indexados no JCR® e na Scopus, respectivamente, de acordo com o país de origem ou de uma categoria considerada. Os valores de x^2 também são apresentados e os valores que não estão em conformidade com a Lei de Benford são realçados. Também são apresentados os dígitos (*d*) com diferenças significativas de acordo com o teste com a estatística *Z*. Observa-se que há dois exemplos em conformidade com a Lei de Benford, de acordo com o teste do qui-quadrado, mas com um dígito com diferença significativa de acordo com o seu valor de Z.

Tabela 6.30 - Comparação do número de artigos publicados em periódicos indexados no JCR® e na Scopus e sua conformidade com a Lei de Benford.

2008	JCR® Edição "*Science*"		País: Suíça		Periódicos: 145 (152)
	Artigos	**Min**	**Max**	x^2	***valor Z* (*d*)**
JCR®	22.735	1	1.960	20,1*	3,70** (5)
Scopus	21.353	1	1.885	11,3	
2007	JCR® Edição "*Social Sciences*"		**País:** Holanda		**Periódicos**: 116 (116)
	Artigos	**Min**	**Max**	x^2	***valor Z* (*d*)**
JCR®	6.979	7	325	18,2*	2,70** (2)
Scopus	6.784	3	318	8,6	
2008	**JCR® Edição "*Science*"**		**País:** Japão		**Periódicos**: 165 (175)
	Artigos	**Min**	**Max**	x^2	***valor Z* (*d*)**
JCR®	21.409	9	1.963	6,8	
Scopus	20.045	5	1.948	17,6*	3,03** (4); 2,22** (9)
2011	JCR® Edição "*Social Sciences*"		**País:** Alemanha		**Periódicos**: 109 (118)
	Artigos	**Min**	**Max**	x^2	***valor Z* (*d*)**
JCR®	3.226	2	141	12,0	2,09** (2)
Scopus	3.020	2	127	18,7*	2,02** (1); 2,09** (2); 1,98** (7)
2011	**JCR® Edição "*Science*"**		**Categoria:** *Endocrinology & Metabolism*		**Periódicos**: 115 (122)
	Artigos	**Min**	**Max**	x^2	***valor Z* (*d*)**
JCR®	15.281	5	704	16,2*	2,01** (3); 2,54** (6)
Scopus	13.164	2	639	7,5	
2011	JCR® Edição "*Social Sciences*"		**Categoria:** *Business*		**Periódicos**: 110 (113)
	Artigos	**Min**	**Max**	x^2	***valor Z* (*d*)**
JCR®	4.819	9	273	21.6*	2,62** (1); 3,04** (2)
Scopus	4.757	6	320	14.0	
2011	**JCR® Edição "*Science*"**		**Categoria:** *Computer Science, Information Systems*		**Periódicos**: 132 (135)
	Artigos	**Min**	**Max**	x^2	***valor Z* (*d*)**
JCR®	9.232	4	564	15,4	2,47** (8)
Scopus	8.389	4	520	16,3*	2,66** (7)
2010	JCR® Edição "*Social Sciences*"		**Categoria:** *Public, Environmental & Occupational Health*		**Periódicos**: 112 (116)
	Artigos	**Min**	**Max**	x^2	***valor Z* (*d*)**
JCR®	9.215	15	485	9,7	
Scopus	8.635	5	511	15,7*	2,43** (3)

* Denota diferença significativa entre os valores observadores e os esperados para $p = 0,05$

** Denota diferença significativa entre a proporção observada e a esperada no nível de 0,05

A não conformidade com a Lei de Benford identificada com a análise realizada neste estudo pode ser um indício de dados incompletos (por exemplo, Karamourzov (2012) observou que há uma pequena fração (menos de 8%) dos periódicos da Rússia indexados pelo JCR® em 2010; Michels e Schmoch (2012) observaram o aumento constante de publicações nos últimos anos que foram também indexadas na WoS e Scopus), erros de dados, inconsistências ou anomalias, e/ou conformidade a uma grande lei de potência exponencial, ocorrendo com os dados do JCR® e/ou Scopus, tendo em vista as diferenças significativas observadas. Estas indicações já foram mencionadas em trabalhos anteriores em que foram observadas não conformidades (por exemplo, Nigrini (2012)).

Acreditamos que a principal contribuição deste estudo é alertar sobre essas diferenças e, talvez, fornecer um instrumento exploratório para identificar onde, possivelmente, algumas anomalias de dados podem estar ocorrendo, independentemente de qual base de dados seja a correta.

6.6. Periódico

O objetivo deste estudo foi analisar e mapear o conteúdo do periódico "*Jornal of Informetrics*" (JOI) até o final de 2012 considerando os dados obtidos na base de dados Scopus, apresentando informações que não podem ser obtidas diretamente na Scopus ou em qualquer outra base de dados.

JOI é um periódico trimestral revisado por pares que abrange a pesquisa em cientometria e informetria. O periódico foi fundado em 2007 e é publicado pela Elsevier. Recentemente, um estudo sobre este periódico foi realizado por Egghe (2012). Esse estudo analisou a coautoria, os países dos autores, as decisões editoriais, tempo de produção e editorial, fator de impacto e aspectos de download dos artigos. Nesse estudo foram considerados 239 artigos publicados até o final de 2011.

6.6.1. Coleta de dados

Até 2012, 290 artigos foram publicados no JOI. No entanto, realizando uma consulta na Scopus apenas 289 artigos foram encontrados. Isso ocorreu porque o artigo "*Object-relational data modelling for informetric databases*" (YU et al., 2008) aparecia na Scopus como "no prelo" e o artigo é de 2008. Neste estudo esse artigo também foi considerado. É importante mencionar que, neste estudo, considerou-se apenas "artigos" publicados no JOI, não levando em conta outro tipo de documento.

Inicialmente, foram baixadas as páginas Web geradas dinamicamente pela Scopus para cada um dos artigos com todos os dados disponíveis. As páginas Web foram armazenadas como arquivos HTML e o nome do arquivo foi salvo com o número de identificação (ID) de cada artigo na Scopus. O próximo passo foi extrair os dados e armazená-los em um banco de dados. Os seguintes dados foram extraídos: número de identificação (ID) do artigo, título, tipo de documento, tipo de fonte, ISSN, volume, número, mês, ano, páginas, total de referências, total de citações e palavras-chave. Também foi extraído o número de identificação (ID) de cada um dos autores em cada artigo. Além disso, foram obtidos o número de identificação (ID) de todas as referências de cada um dos artigos e o número de identificação (ID) de todos os artigos que citaram algum artigo do JOI. No caso das referências estava disponível apenas o número de identificação (ID) dos documentos indexados pela Scopus.

O passo seguinte foi baixar e extrair os dados desses documentos (aqui foi considerado qualquer tipo de documento e não apenas artigos). Em adição, também foi obtido o número de identificação (ID) dos periódicos que os artigos foram publicados. Finalmente, foram baixadas as páginas Web geradas para cada um dos autores de acordo com o seu número de identificação (ID) e também para cada um dos periódicos. Para autores foram extraídos os seguintes dados: nome, instituição, número de identificação (ID) da instituição, cidade, país, total de documentos, total de citações, índice H e as suas áreas de estudo. Para periódicos os seguintes dados foram extraídos e também

armazenados em um banco de dados: título, ISSN, editor e suas áreas temáticas.

Todo esse processo brevemente descrito aqui foi realizado nos dias 22 e 23 de dezembro de 2012. Isso só foi possível porque foi utilizada a linguagem ScopusMiner que permite extrair automaticamente dados da Scopus e armazenar em um banco de dados.

6.6.2. Resultados e discussões

De acordo com dados da Scopus, o JOI publicou 290 artigos até o final de 2012. O número de artigos publicados por edição é apresentado na Tabela 6.31. O número de autores e citações por edição também é apresentado. É possível observar que o número médio de autores está aumentando nas últimas edições. Além disso, o número médio de autores por artigo também aumentou. No estudo realizado por Egghe (2012), o número médio de autores por artigo era 2,276 e, neste estudo é igual a 2,36 (aumento de 3,69%). O número de países dos autores de acordo com a afiliação também está aumentando nos últimos anos. Isso é importante porque mostra que o JOI obteve uma maior inserção internacional ou ainda porque a área de Cientometria/Bibliometria começa a se consolidar na comunidade científica.

De um total de 290 artigos publicados no JOI, 54 (18,62%) artigos ainda não foram citados. É interessante observar que em algumas edições todos os artigos foram citados. O número total de citações foi 2.458 de 1.421 documentos (não restrito a artigos). O número total de citações em documentos com ISSN foi 2.356 (95,85%) de 398 periódicos distintos. 583 (24,75%) dessas citações foram em 218 artigos publicados no JOI. Também é interessante observar que o número de países dos autores que citaram algum artigo publicado no JOI é muito maior (57,50%) que o número de países dos autores que publicaram no JOI.

Tabela 6.31 - Número de artigos publicados, autores e citações por edição do JOI.

Ano	Ed.	n	Autores	Média	Países	Citações	Média	Países	Artigos	%
2007	1	10	21	2,10	9	242	24,20	41	10	100
	2	8	15	1,88	6	202	25,25	36	8	100
	3	7	20	2,86	8	204	29,14	35	7	100
	4	6	10	1,67	5	44	7,33	18	6	100
2008	1	5	11	2,20	6	25	5,00	9	4	80
	2	6	9	1,50	6	69	11,50	23	6	100
	3	9	18	2,00	8	97	10,78	23	8	88,89
	4	13	25	1,92	11	201	15,46	32	12	92,31
2009	1	8	24	3,00	9	97	12,13	38	8	100
	2	6	16	2,67	7	63	10,50	28	6	100
	3	8	20	2,50	7	130	16,25	27	8	100
	4	10	22	2,20	8	102	10,20	28	10	100
2010	1	14	31	2,21	13*	142	10,14	33	14	100
	2	8	18	2,25	7	39	4,88	16	7	87,50
	3	23	46	2,00	14*	322	14,00	41	23	100
	4	19	43	2,26	16	94	4,95	25	18	94,74
2011	1	18	59	3,28	13*	159	8,83	26	16	88,89
	2	7	17	2,43	7	26	3,71	10	6	85,71
	3	17	40	2,35	11	68	4,00	26	14	82,35
	4	19	46	2,42	16	69	3,63	25	17	89,47
2012	1	15	37	2,47	17*	30	2,00	19	10	66,67
	2	17	43	2,53	13	22	1,29	13	10	58,82
	3	10	24	2,40	13	6	0,60	5	4	40
	4	27	70	2,59	15	5	0,19	5	4	14,81
Total		290	685	2,36	40**	2.458	8,48	63	236	81,38

* 1 autor não informou o país

** 3 autores não informaram o país

O número médio de citações do JOI é 8,48 por artigo. O FI (4,229) do JOI é alto e o segundo na sua categoria, sua meia-vida (2,6 anos) é baixa e o índice de imediatez (0,771) é alto. Além desses fatores, que podem ser obtidos diretamente no JCR®, esse número significativo de citações provavelmente seja decorrente do baixo tempo editorial (tempo entre a primeira submissão e a aceitação final), como pode ser observado na Figura 6.15. O tempo editorial médio por edição, que não é obtido no JCR®, é de cerca de 14 semanas e está diminuindo uma vez que no estudo realizado por Egghe (2012) esse tempo era

de 18 semanas. O tempo editorial foi calculado verificando manualmente os dados em cada um dos artigos publicados no JOI.

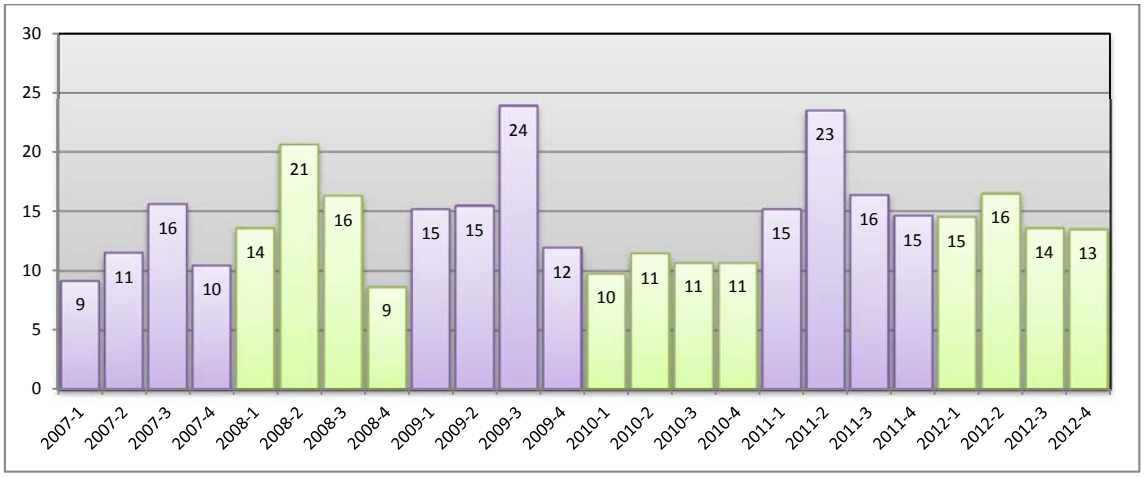


Figura 6.15 - Tempo editorial dos artigos publicados no JOI por edição em semanas.

De um total de 685 autores do JOI, 415 são autores distintos e 409 (98,55%) têm a cidade e o país de sua afiliação informados na Scopus. Esses autores estão distribuídos em 180 locais diferentes, conforme ilustra a Figura 6.16. De acordo com a cidade e o país dos autores foram obtidas a latitude e a longitude de cada local automaticamente (utilizando a ferramenta *GPS Visualizer*) considerando as coordenadas geográficas do Yahoo! e os mapas foram apresentados utilizando o *Google Maps*.

“Madri” é a cidade com mais autores (35) e “Pequim” é a segunda com mais autores (27). “Roma” (24) e “Amsterdã” (21) são as outras cidades com mais de 20 autores. Pode-se observar que a maioria das cidades estão localizadas na Europa e nos Estados Unidos. De um total de 180 locais, 79 (43,89%) têm apenas um autor que publicou no JOI.

Figura 6.16 - Distribuição geográfica dos autores que publicaram artigos no JOI de acordo com a cidade e país de sua afiliação.

Também foi analisada a distribuição geográfica dos autores considerando o índice H deles de acordo com dados da Scopus, conforme ilustra a Figura 6.17. De um total de 415 autores distintos, 394 (94,94%) têm o índice H informado na Scopus. Entretanto, 5 (1,27%) deles não têm a cidade e o país de sua afiliação informado na Scopus. Neste caso, há 177 locais diferentes e em 75 (42,37%) deles a soma do índice H é menor do que 10.

Figura 6.17 - Distribuição geográfica dos autores que publicaram artigos no JOI considerando o seu índice H e o número de autores em cada cidade e país de sua afiliação.

“Madri” também é a cidade com maior soma do índice H dos autores (189). De um total de 35 autores, 20 (57,14%) deles têm o índice H informado na Scopus. A média da soma do índice H desses autores é 9,45. “Pequim” é a cidade com mais autores com índice H. Entretanto, é somente a quarta cidade com maior soma do índice H, atrás de “Amsterdã” (soma do índice H igual a 127 e 7 autores) e “Lovaina” (soma do índice H igual a 121 e 8 autores). “Canonsburg” é a cidade com o maior valor da média da soma do índice H (55) e apenas um autor.

De um total de 290 artigos publicados no JOI, em apenas um não foi informado nenhuma palavra-chave. Nesses artigos 921 palavras-chave distintas foram informadas e as 50 palavras-chave mais utilizadas nos artigos são apresentadas na Figura 6.18. “*h-index*” é a palavra-chave mais utilizada nos artigos publicados no JOI e foi utilizada em 39 (13,49%) desses artigos. Além disso, há outras variações (“*Hirsch index*” (13) e “*h index*” (10)). Considerando-se todas as variações, essa palavra-chave foi utilizada em 62 (21,45%) artigos. “*Bibliometrics*” (28) e “*Citation analysis*” (25) também foram citadas por um número significativo de artigos. 23 palavras-chave foram citadas em 3 artigos publicados no JOI. Entretanto, apenas 9 foram apresentadas na “nuvem de palavras” devido ao limite de 50 palavras-chave.

Figura 6.18 - Palavras-chave mais utilizados nos artigos publicados no JOI.

Considerando o número de citações de cada artigo, é possível verificar as palavras-chave mais utilizadas nestes artigos citados, conforme apresentado na Tabela 6.32. Neste caso, 235 artigos publicados no JOI foram citados e 762 palavras-chave distintas foram utilizadas. "*h-index*" também é a palavra-chave mais utilizada de acordo com o número de citações. De um total de 39 artigos que utilizaram essa palavra-chave, 33 são artigos que foram citados pelo menos uma vez, num total de 423 citações.

Tabela 6.32 - Ranking das palavras-chave mais utilizadas nos artigos publicados no JOI de acordo com o número de citações.

Palavra-chave	Citações	n	Total	Palavra-chave	Citações	n	Total
h-index (1ª)	423	33	39	*Ranking* (11ª)	152	8	8
g-index (2ª)	183	19	19	*Journal Impact Factor* (12ª)	144	7	7
Citations (3ª)	178	13	14	*Research evaluation* (13ª)	121	13	15
Citation analysis (4ª)	177	20	25	*Web of Science* (14ª)	113	6	7
Bibliometrics (5ª)	176	24	28	***Pareto distribution*** (15ª)	97	2	2
Scopus (5ª)	176	6	7	*Peer review* (16ª)	92	9	11
Hirsch index (7ª)	174	10	13	***Physical Review*** (17ª)	87	2	2
Normalization (8ª)	171	6	6	***Stochastic model*** (18ª)	85	3	3
Impact factor (9ª)	169	12	13	*Source normalization* (19ª)	84	2	2
Citation (10ª)	165	6	7	*Informetric process* (20ª)	83	2	2

Pode ser interessante saber as palavras-chave utilizadas nos artigos mais citados, pois isso pode indicar temas com maior visibilidade e interesse da comunidade leitora daquele veículo. Também é interessante observar que algumas palavras-chave (em cor azul e negrito) apesar de, aparentemente, serem menos aderentes ao escopo do periódico, foram relativamente bastante citadas.

De um total de 415 autores distintos dos artigos publicados no JOI, apenas um não tem a área de estudo informada na Scopus. 27 áreas distintas foram obtidas e são apresentadas na Figura 6.19. A grande maioria dos autores é classificada em mais de uma área. 290 (69,88%) dos 415 autores são da área de "*Computer Science*". "*Mathematics*" (69,64%) e "*Decision Sciences*" (69,64%) são outras áreas que os autores também são classificados de maneira significativa.

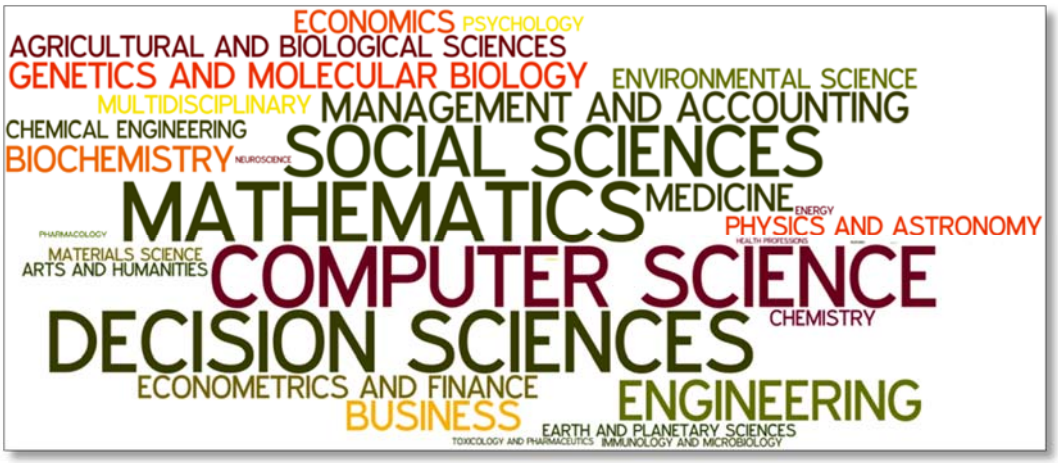


Figura 6.19 - Áreas de estudo dos autores dos artigos publicados no JOI.

O número de referências dos artigos publicados no JOI é apresentado na Tabela 6.33. O número dessas referências indexadas na Scopus e o número de autocitações também são apresentados. O número médio de referências por edição variou pouco nos últimos dois anos. O mesmo ocorre com a média de referências indexadas na Scopus e com o número de países de autores distintos. O número de países dos autores distintos das referências (71) é maior que o número de países dos autores (40) e que o número de países dos autores das citações (63). Embora o percentual de autocitações esteja aumentando nos últimos anos, ele ainda é baixo (5,07%), mesmo considerando apenas referências indexadas na Scopus (7,95%).

De um total de 290 artigos publicados no JOI, apenas 3 (1,03%) não têm pelo menos uma referência indexada na Scopus. De um total de 5.875 referências indexadas na Scopus, há 3.096 referências distintas das quais 2.940 são de

artigos publicados em outros periódicos. Neste caso, não foram consideradas as referências do próprio JOI.

Tabela 6.33 - Número de referências dos artigos publicados no JOI, número de referências indexadas na Scopus e número de autocitações por edição.

Ano	Ed.	Referências	Média	Scopus	Média	Países	Autocitações	% Refs.	% Scopus
2007	1	224	22,40	120	12,00	28	-	-	-
	2	158	19,75	66	8,25	20	1	0,63	1,52
	3	159	22,71	67	9,57	18	-	-	-
	4	181	30,17	73	12,17	20	-	-	-
2008	1	135	27,00	74	14,80	19	4	2,96	5,41
	2	171	28,50	88	14,67	23	3	1,75	3,41
	3	199	22,11	124	13,78	22	3	1,51	2,42
	4	374	28,77	235	18,08	31	20	5,35	8,51
2009	1	226	28,25	99	12,38	26	3	1,33	3,03
	2	247	41,17	138	23,00	26	5	2,02	3,62
	3	528	66,00	240	30,00	24	14	2,65	5,83
	4	263	26,30	175	17,50	30	7	2,66	4,00
2010	1	431	30,79	293	20,93	33	24	5,57	8,19
	2	226	28,25	147	18,38	34	8	3,54	5,44
	3	686	29,83	471	20,48	39	24	3,50	5,10
	4	712	37,47	502	26,42	44	25	3,51	4,98
2011	1	667	37,06	433	24,06	38	45	6,75	10,39
	2	252	36,00	165	23,57	28	14	5,56	8,48
	3	555	32,65	349	20,53	38	34	6,13	9,74
	4	636	33,47	432	22,74	46	52	8,18	12,04
2012	1	559	37,27	391	26,07	42	39	6,98	9,97
	2	576	33,88	402	23,65	38	42	7,29	10,45
	3	286	28,60	214	21,40	30	21	7,34	9,81
	4	766	28,37	577	21,37	40	79	10,31	13,69
Total		9.217	31,78	5.875	20,26	71	467	5,07	7,95

Nos artigos citados nos artigos publicados no JOI há 4.293 autores distintos. Dentre eles, 4.129 (96,18%) autores têm a cidade e o país de sua afiliação informados na Scopus. A distribuição geográfica dos autores dos artigos citados (referências) nos artigos publicados no JOI de acordo com a cidade e o país de sua afiliação é apresentada na Figura 6.20. Esses autores estão distribuídos em 890 locais diferentes. “Amsterdã” é a cidade com mais artigos

citados no JOI e “Budapeste” é a segunda. Pode-se observar que aqui também os autores estão concentrados na Europa e nos Estados Unidos.

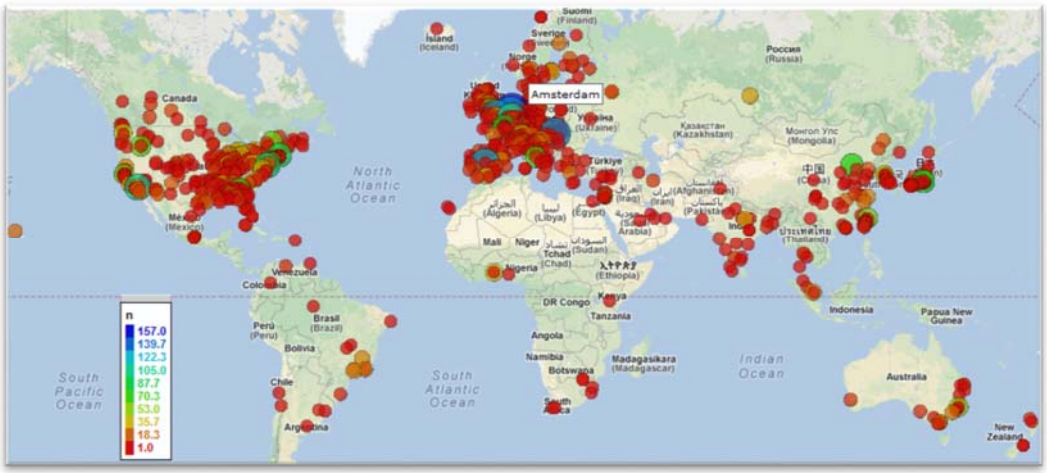


Figura 6.20 - Distribuição geográfica dos autores dos artigos citados nos artigos publicados no JOI.

De um total de 5.875 referências indexadas na Scopus, 4.684 (79,73%) são classificadas como “Periódico” no tipo de fonte e como “Artigo” no tipo de documento e o ISSN também estão disponíveis. Desse total, há 562 periódicos distintos classificados em 211 áreas distintas e representam 2.411 artigos distintos. O número de referências das principais áreas é apresentado na Tabela 6.34.

Tabela 6.34 - Número de referências por área dos periódicos citados nos artigos publicados no JOI.

Área	n	%	Periódicos	Artigos
Computer Science: Computer Science Applications	1.879	40,12	35	808
Social Sciences: Library and Information Sciences	1.609	34,35	38	776
Social Sciences	1.385	29,57	61	671
Social Sciences: Law	1.178	25,15	4	525
Computer Science: Information Systems	934	19,94	35	427
Decision Sciences: Management Science and Operations Research	828	17,68	16	352
Mathematics: Statistics and Probability	626	13,36	29	269
Computer Science: Artificial Intelligence	613	13,09	27	265
Computer Science: Software	609	13,00	25	255
Computer Science: Computer Networks and Communications	591	12,62	17	239

"*Computer Science: Computer Science Applications*" é a área dos periódicos que mais foram citados nos artigos publicados no JOI. Entretanto, "*Economics, Econometrics and Finance: Economics and Econometrics*" é a área que foi mais citada por periódicos distintos. Essa área foi referenciada por 66 (11,74%) periódicos distintos representando 142 (3,03%) das referências e 116 (4,81%) dos artigos distintos citados nas referências. Essa área ocupa a vigésima segunda posição e, portanto, não é listada na Tabela 6.34.

De um total de 2.411 artigos distintos citados nos artigos publicados no JOI, 2.234 (92,66%) artigos são de periódicos indexados no JCR® e 4.409 (94,13%) das 4.684 referências em periódicos e 473 (84,16%) de 562 periódicos diferentes também são indexados no JCR®. Os periódicos indexados no JCR® com mais artigos citados nos artigos publicados no JOI são apresentados na Tabela 6.35. "Scientometrics" é o periódico indexado no JCR® com mais artigos citados (23,05%) e com mais referências citadas (26,45%) nos artigos publicados no JOI. JOI é o terceiro periódico indexado no JCR® com mais artigos citados. É interessante observar que os periódicos indexados no JCR® com mais artigos citados nos artigos publicados no JOI são de 3 países, com destaque para periódicos da "Holanda".

Tabela 6.35 - Periódicos indexados no JCR® com mais artigos citados nos artigos publicados no JOI.

ISSN	Periódico	País	FI	Artigos	Referências
0138-9130	Scientometrics	Holanda	1,966	515	1.166
1532-2882	Journal of the Am. Soc. for Inf. Sci. and Tec.	Estados Unidos	2,081	215	558
1751-1577	Journal of Informetrics	Holanda	4,229	156	467
0048-7333	Research Policy	Holanda	2,520	104	178
0306-4573	Information Processing & Management	Estados Unidos	1,119	65	145
1539-3755	Physical Review E	Estados Unidos	2,255	45	66
0022-0418	Journal of Documentation	Inglaterra	1,058	37	61
0165-5515	Journal of Information Science	Inglaterra	1,299	36	65
0036-8075	Science	Estados Unidos	31,201	33	123
0958-2029	Research Evaluation	Inglaterra	0,845	32	49

Além do FI também é possível obter as categorias de um periódico no JCR® uma vez que um periódico pode ser classificado em mais de uma categoria. Assim, é possível obter o número de referências por categoria do JCR® dos artigos citados nos artigos publicados no JOI, conforme apresentado na Tabela 6.36.

Tabela 6.36 - Número de referências por categoria do JCR® em 2011 dos periódicos dos artigos citados nos artigos publicados no JOI.

Categoria	Edição	Referências	Periódicos	JCR®	%	FI Mediana
Information Science & Library Science	*Social*	2.652	37	83	44,58	0,641
Computer Science, Interdisciplinary Applications	*Science*	1.186	10	99	10,10	1,271
Computer Science, Information Systems	*Science*	865	26	135	19,26	0,898
Multidisciplinary Sciences	*Science*	387	11	56	19,64	0,499
Management	*Social*	314	40	168	23,81	1,183
Planning & Development	*Social*	218	8	54	14,81	0,925
Economics	*Social*	133	60	321	18,69	0,778
Physics, Multidisciplinary	*Science*	124	15	84	17,86	0,983
Business	*Social*	106	26	113	23,01	1,135
Physics, Mathematical	*Science*	85	7	55	12,73	1,211

JOI está classificado na categoria "*Information Science & Library Science*", que é a categoria mais utilizada (entre 141) nas referências dos artigos publicados no JOI. Em 2011, o FI do JOI aumentou de 3,119 para 4,229 reclassificando o JOI da terceira para a segunda posição dentre 83 periódicos nesta categoria. Além disso, foram citados quase metade (44,58%) dos periódicos desta categoria. Também é interessante observar que o FI do JOI é maior que o FI da mediana dos periódicos das principais categorias apresentadas na Tabela 6.36.

O número de citações dos artigos publicados no JOI vem aumentando significativamente a cada ano, como pode ser observado na Tabela 6.37. De um total de 2.458 citações, 1.978 (80,47%) delas são classificadas como "Periódico"" no tipo de fonte e como "Artigo" no tipo de documento e o ISSN também estão disponíveis. Essas citações representam 1.091 artigos distintos

publicados em periódicos distintos (308 de um total de 415) e com autores de 60 países distintos. Considerando apenas estes artigos o número de autocitações por ano é maior que quando considerado por edição, uma vez que por edição foram considerados todos os tipos de citações. Embora nenhum artigo tenha sido publicado em 2006, um artigo (no prelo) do JOI obteve uma citação. O número de periódicos que cita algum artigo publicado na JOI também está aumentando. Em 2012, por exemplo, o aumento foi de 59,38%.

Tabela 6.37 - Número de citações e autocitações por ano dos artigos publicados no JOI.

Ano	n	Citações	Artigos	Periódicos	Países	Autocitações	%	Artigos	%
2006	-	1	1	1	1	-	-	-	-
2007	31	9	8	6	8	1	11,11	1	12,50
2008	33	103	58	18	20	30	29,13	17	29,31
2009	32	190	128	54	34	29	15,26	15	11,72
2010	64	337	211	80	44	81	24,04	40	18,96
2011	61	541	273	96	44	152	28,10	50	18,32
2012	69	749	391	153	49	181	24,17	59	15,09
2013*	-	48	21	7	18	41	85,42	15	71,43
Total	**290**	**1.978**	**1.091**	**308****	**60****	**515**	**26,04**	**197**	**18,06**

* Dados parciais ** Distintos

Atualmente, o artigo publicado no JOI com o maior número de citações é "*A systematic analysis of Hirsch-type indices for journals*" (SCHUBERT; GLÄNZEL, 2007) com 77 citações na Scopus em 16/01/2013. Este artigo também é o mais citado de acordo com dados da WoS com 72 citações.

De um total de 1.091 artigos distintos de periódicos que citaram algum artigo publicado no JOI, há 1.602 autores distintos. Dentre eles, 1.566 (97,75%) autores distintos têm a cidade e o país de sua afiliação informados na Scopus. Desse total, 1.439 autores distintos citaram algum artigo publicado no JOI em algum periódico que não é o JOI. Esses autores estão distribuídos em 543 locais diferentes, como ilustra a Figura 6.21. "Madrid" (98) e "Granada" (75) são as cidades com mais autores nessa condição. De um total de 543 locais, 208 (38,31%) locais têm apenas um autor que citou algum artigo publicado no JOI.

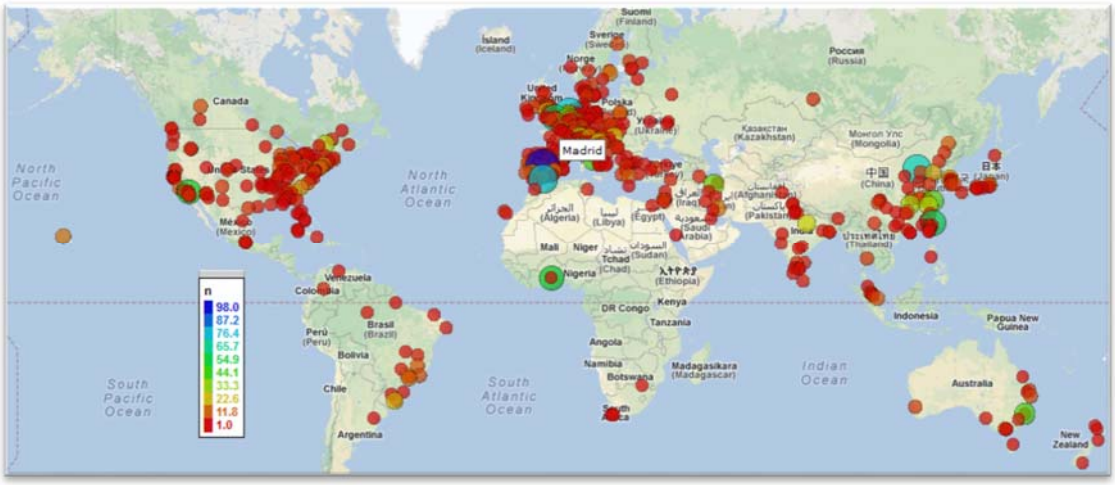


Figura 6.21 - Distribuição geográfica dos autores dos artigos que citaram algum artigo publicado no JOI em algum periódico diferente do JOI de acordo com a cidade e o país de sua afiliação.

Entretanto, se forem considerados apenas os 284 autores distintos que citaram algum artigo publicado no JOI em algum outro artigo publicado no JOI, a distribuição geográfica é muito diferente, como pode ser observado na Figura 6.22. Neste caso, “Roma” (21) e “Madri” (19) são as cidades com mais autores entre os 135 locais diferentes. Assim, é possível verificar a cobertura das citações dos artigos publicados no JOI.

Figura 6.22 - Distribuição geográfica dos autores dos artigos que citaram algum artigo publicado no JOI em algum outro artigo publicado no JOI de acordo com a cidade e o país de sua afiliação.

De um total de 1.091 artigos distintos de periódicos que citaram algum artigo publicado no JOI, apenas em 3 não foi possível identificar o número de identificação (ID) da fonte. Esse identificador é utilizado para obter as áreas do periódico. Assim, foi possível obter o número de citações por área dos periódicos dos artigos que citaram algum artigo publicado no JOI. 188 áreas distintas foram identificadas e as áreas com o maior número de citações são apresentadas na Tabela 6.38.

Tabela 6.38 - Número de citações por área dos periódicos dos artigos que citaram algum artigo publicado no JOI.

Área	n	%	Periódicos	Artigos
Computer Science: Computer Science Applications	1.028	13,62	23	482
Social Sciences: Library and Information Sciences	656	8,69	49	373
Decision Sciences: Management Science and Operations Research	559	7,41	10	225
Mathematics: Statistics and Probability	555	7,35	10	231
Mathematics: Modeling and Simulation	528	7,00	7	204
Mathematics: Applied Mathematics	523	6,93	7	203
Social Sciences	463	6,13	23	256
Social Sciences: Law	422	5,59	5	226
Computer Science: Information Systems	419	5,55	20	208
Computer Science: Software	317	4,20	14	151

"*Computer Science: Computer Science Applications*" também é a área dos periódicos que mais foi citada nos artigos que citaram algum artigo publicado no JOI. Quase metade (44,34%) dos artigos que citaram algum artigo publicado no JOI são dessa área. Entretanto, "*Social Sciences: Library and Information Sciences*" é a área com mais periódicos distintos e ocupa a segunda posição na Tabela 6.38. Essa área foi citada por 49 (16,12%) periódicos distintos (entre 304) representando 656 (8,69%) das citações e 373 (34,31%) dos artigos distintos citados nas citações. É interessante observar que essas duas áreas são as áreas do periódico "Scientometrics", que é o periódico indexado no JCR® que citou mais artigos publicados no JOI, conforme apresentado na Tabela 6.39. Além disso, "*Social Sciences: Library and Information Sciences*" também é a área do JOI de acordo com a classificação da Scopus.

Tabela 6.39 - Periódicos indexados no JCR® que citaram mais artigos publicados no JOI.

ISSN	Periódico	País	FI	Artigos	Citações
0138-9130	Scientometrics	Holanda	1,966	221	416
1751-1577	Journal of Informetrics	Holanda	4,229	197	515
1532-2882	Journal of the Am. Soc. for Inf. Sci. and Tec.	Estados Unidos	2,081	133	296
1932-6203	**PLOS ONE**	Estados Unidos	4,092	31	50
0958-2029	Research Evaluation	Inglaterra	0,845	26	48
1468-4527	**Online Information Review**	Estados Unidos	0,939	16	34
0306-4573	Information Processing & Management	Estados Unidos	1,119	16	24
0378-4371	**Physica A:Statistical Mech. and its Applications**	Holanda	1,373	16	19
0165-5515	Journal of Information Science	Inglaterra	1,299	12	23
1539-3755	Physical Review E	Estados Unidos	2,255	8	8

É interessante observar que os periódicos indexados no JCR® que citaram mais artigos publicados no JOI também são de 3 países e também com destaque para periódicos da "Holanda". O mesmo ocorre com os periódicos indexados no JCR® com mais artigos citados nos artigos publicados no JOI. Além disso, há 3 periódicos indexados no JCR® que são muito citados nos artigos publicados no JOI (Tabela 6.35) e não estão na lista entre os periódicos indexados no JCR® que citaram muitos artigos publicados no JOI: "Research Policy", "Journal of Documentation" e "Science". Há também 3 periódicos que citaram muitos artigos publicados no JOI e não estão entre os periódicos indexados no JCR® que são muito citados nos artigos publicados no JOI: "PLOS ONE", "Online Information Review" and "Physica A:Statistical Mech. and its Applications" (na cor azul e em negrito).

O número de citações por categoria do JCR® em 2011 dos artigos que citaram artigos publicados no JOI é apresentado na Tabela 6.40. 111 categorias distintas foram identificadas e "*Information Science & Library Science*" é a categoria do JCR® com maior número de citações. "*Planning & Development*", "*Physics, Mathematical*" são as categorias que estão entre as mais citadas nos artigos publicados no JOI e não estão entre as principais categorias dos artigos

que citaram algum artigo publicado no JOI. Por outro lado, "*Biology*" e "*Operation Research & Manage Science*" (na cor azul e em negrito) são as categorias que estão entre as categorias dos artigos que citaram algum artigo publicado no JOI e não estão entre as categorias mais citadas nos artigos publicados no JOI. Também é possível observar que JOI tem inserção interdisciplinar uma vez que as categorias que citaram algum artigo publicado no JOI são de áreas muito diferentes.

Tabela 6.40 - Número de citações por categoria do JCR® em 2011 dos periódicos dos artigos que citaram artigos publicados no JOI.

Categoria	Edição	Citações	Periódicos	JCR®	%	FI Mediana
Information Science & Library Science	*Social*	1.434	34	83	40,96	0,641
Computer Science, Interdisciplinary Applications	*Science*	429	10	99	10,10	1,271
Computer Science, Information Systems	*Science*	402	15	135	11,11	0,898
Biology	*Science*	52	3	85	3,53	1,540
Physics, Multidisciplinary	*Science*	38	10	84	11,90	0,983
Management	*Social*	29	16	168	9,52	1,183
Operation Research & Manage Science	*Science*	20	12	77	15,58	0,856
Economics	*Social*	20	8	321	2,49	0,778
Multidisciplinary Sciences	*Science*	19	8	56	14,29	0,499
Business	*Social*	18	6	113	5,31	1,135

Foi analisado também o número de citações dos artigos que citaram algum artigo publicado no JOI, conforme apresentado na Tabela 6.41. O número de referências e autores desses 1.091 artigos também são apresentados. Em 2006, por exemplo, é possível observar que o único artigo que citou algum artigo publicado no JOI já recebeu um número significativo de citações. É interessante notar que o número médio de autores também aumentou nos últimos anos.

Tabela 6.41 - Número de citações dos artigos que citaram algum artigo publicado no JOI.

Ano	Artigos	Citações	Média	Referências	Média	Autores	Média
2006	1	144	144,00	34	34,00	3	3,00
2007	8	306	38,25	149	18,63	24	3,00
2008	58	1.116	19,24	1.904	32,83	197	3,40
2009	128	1.323	10,34	4.481	35,01	423	3,30
2010	211	1.464	6,94	8.275	39,22	764	3,62
2011	273	1.051	3,85	10.598	38,82	1.387	5,08
2012	391	406	1,04	15.581	39,85	1.788	4,57
2013	21	-	-	788	37,52	110	5,24
Total	**1.091**	**5.810**	**5,33**	**41.810**	**38,32**	**4.696**	**4,30**

Na Scopus é simples identificar os autores que publicaram mais artigos em um periódico e é mais trabalhoso identificar qual par de coautores publicou mais artigos em um determinado periódico. Não é possível, entretanto, mapear os relacionamentos entre os autores de um periódico de acordo com a instituição de sua afiliação. Mas isto é possível ser feito com os dados extraídos utilizando a linguagem ScopusMiner. Na Figura 6.23 é apresentado o mapeamento dos relacionamentos entre os autores dos artigos publicados no JOI de acordo com a instituição de sua afiliação informada na Scopus. Nessa rede, há 326 instituições diferentes de 413 autores distintos (apenas 2 autores não informaram a instituição) e entre eles há 413 relacionamentos. O tamanho dos vértices representa o número de artigos de uma instituição e é colorido de acordo com o número de relacionamentos. Todas as instituições sem relacionamentos foram eliminadas. A cor cinza indica que uma instituição se relaciona apenas com uma única outra instituição. A cor cinza escuro indica que uma instituição se relaciona com duas instituições; a cor laranja com 3 ou 4 instituições, a cor amarelo com 5 instituições; a cor verde com 6 ou 7 instituições; a cor azul com 8 ou 9 instituições; a cor magenta com 10 instituições e a cor vermelho com mais de 10 instituições.

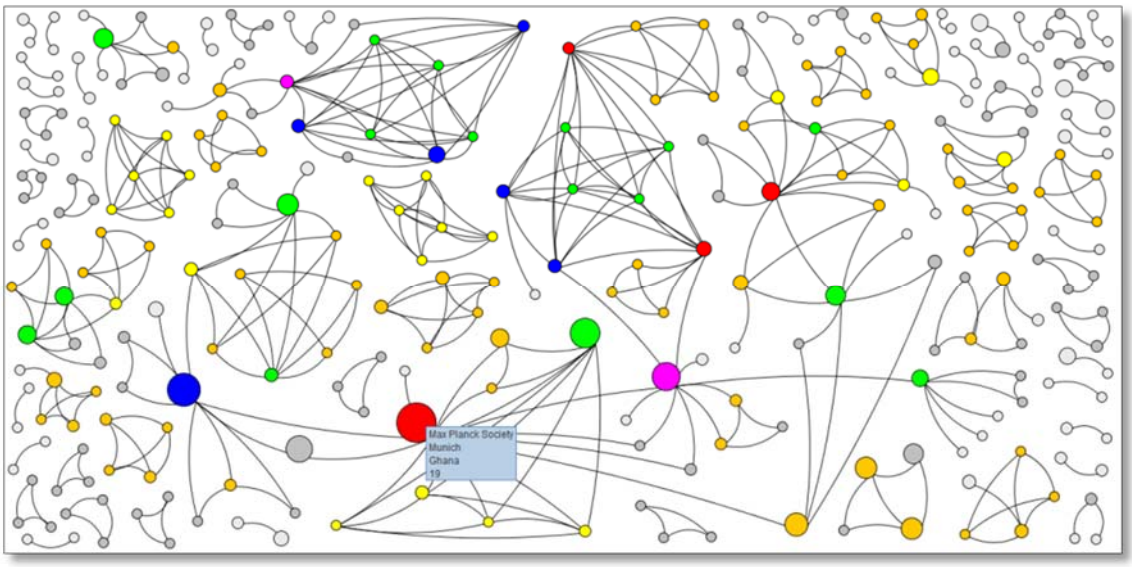


Figura 6.23 - Mapeamento dos relacionamentos dos autores que publicaram artigos no JOI de acordo com a instituição de sua afiliação.

O autor com mais relacionamentos é “Lutz Bornmann” (em destaque na rede). Ele publicou artigos no JOI com autores de outras 19 instituições. Os principais co-autores são “Hans-Dieter Daniel” (Universidade de Zurique, Suíça) com 12 artigos e “Rüdiger Mutz” (*Institut für Forstbenutzung und Forstliche Arbeitswissenschaft*, Alemanha) com 5 artigos publicados juntos no JOI.

Os principais relacionamentos dos autores que publicaram algum artigo no JOI de acordo com instituição de sua afiliação são apresentados na Figura 6.24. Nesta rede (grafo) é possível verificar que foram identificados 5 cliques maximais e dois cliques máximos de tamanho 8.

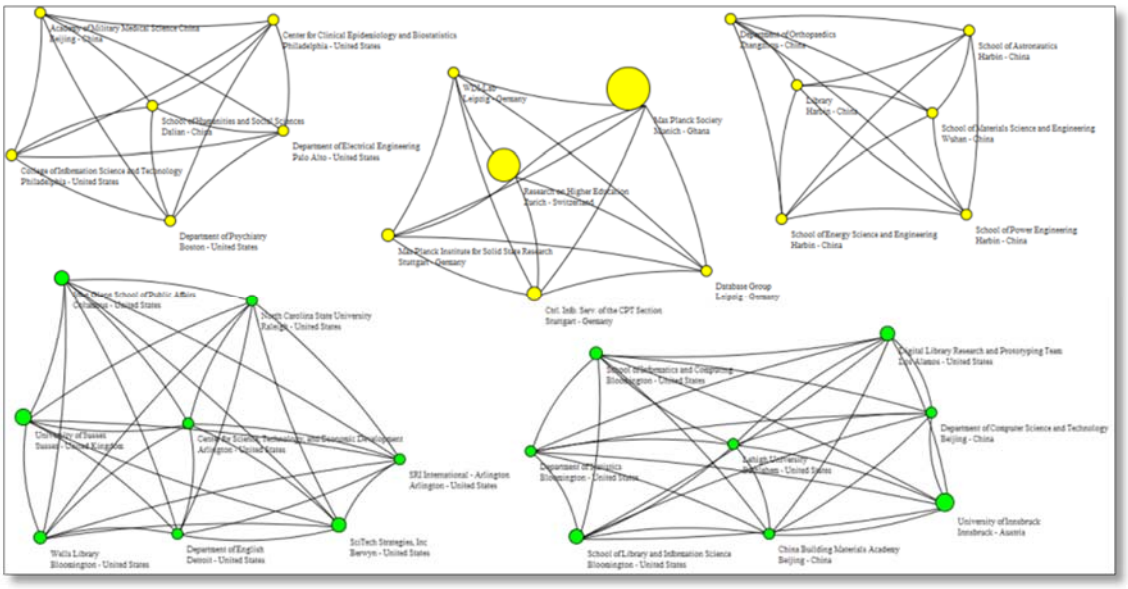


Figura 6.24 - Principais grupos de relacionamentos dos autores que publicaram artigos no JOI de acordo com a instituição de sua afiliação.

Os autores citados anteriormente podem ser facilmente identificados no clique maximal de tamanho 6. As seguintes observações são feitas com relação a afiliação obtida desses autores. “Lutz Bornmann” apareceu como sendo afiliado a uma instituição de “Gana”. Em uma consulta na Scopus realizada em 16/01 já o mostrava como sendo da “Alemanha”. A instituição era “*Max Planck Society*” e depois também como sendo “*Max Planck Society for the Advancement of Science, Division for Science and Innovation Studies*”, na mesma cidade. “Hans -Dieter Daniel” apareceu como sendo afiliado à instituição “*Research on Higher Education*” e depois como sendo afiliado à “*University of Zurich, Evaluation Office*” nas mesmas cidades e países. “Rüdiger Mutz” apareceu como sendo afiliado ao “*Institut für Forstbenutzung und Forstliche Arbeitswissenschaft*” e depois como sendo afiliado ao “*Eidgenossische Technische Hochschule Zurich, Professorship for Social Psychology and Research on Higher Education*”. A cidade era “Freiburg im Breisgau” na Alemanha e, depois, “Zurich”na Suíça.

É importante mencionar que todas essas afiliações citadas aparecem nas afiliações relacionadas de cada autor na Scopus e mudanças podem influenciar os resultados. Portanto, é evidente a importância e conveniência de automatizar a extração de dados, pois eles podem mudar rapidamente. Neste

estudo foi mostrado, por exemplo, que o número de citações de um artigo mudou em um curto período de tempo.

Uma outra contribuição deste estudo foi mostrar que um periódico pode ser analisado utilizando dados da Scopus. Um possível passo seguinte natural a se fazer seria comparar um periódico com outro da mesma área.

6.7. Área de atuação

Neste estudo de caso é apresentado um perfil dos doutores brasileiros cadastrados na PL que atuam na área de Engenharia de Software (ES). O objetivo deste estudo foi tentar obter uma visão geral da área de ES no Brasil nos últimos 25 anos, a partir das informações públicas declaradas pelos pesquisadores brasileiros.

No currículo Lattes um pesquisador pode indicar até seis áreas de atuação de acordo com a classificação das áreas de conhecimento adotadas pelo CNPq. Porém, não é confiável utilizar-se dessa informação uma vez que muitos pesquisadores podem não estar mais ativos na área de ES.

Uma possível proposta para dar maior confiabilidade à informação sobre os pesquisadores que atuam na área de ES no Brasil nos últimos 25 anos foi verificar quantas vezes o termo “ES” (em Português e em Inglês) é citado no currículo Lattes dos doutores. O valor limite “10” foi estabelecido empiricamente considerando o impacto deste valor sobre os membros listados no grupo. Um número menor incluiria muitos pesquisadores que podem não ser reconhecidos por seus pares como sendo parte desta área, um número maior excluiria muitos pesquisadores novos ou jovens atuando nesta área.

Inevitavelmente, erros de classificação podem ocorrer, uma vez que o problema da construção de uma lista de pesquisadores que realmente trabalham na área de ES não é uma tarefa trivial. Para tentar minimizar os inevitáveis erros, foi estabelecido como critério que esses pesquisadores têm de aparecer pelo menos uma vez como autor (ou coautor) de um artigo publicado no Simpósio Brasileiro de Engenharia de Software (SBES) ou em

qualquer outro evento em que o termo ES (em Português ou em Inglês) aparece em seu título. O SBES é o principal congresso brasileiro de ES e sua vigésima quinta edição ocorreu em 2011. Foram considerados também os eventos em que o termo ES aparece porque alguns pesquisadores não informam corretamente o SBES em seus currículos Lattes. Um outro critério adicional é que o pesquisador tenha tido pelo menos um artigo publicado em periódicos calssificados na categoria "*Computer Science, Software Engineering*" (CSSE) do JCR®.

Estes critérios são restritivos, e aparentam ser apropriados para caracterizar um grupo limitado de pesquisadores que atuam na área de ES de modo que os eventuais erros não comprometam o estudo realizado.

Neste estudo de caso do perfil dos doutores brasileiros cadastrados na PL que atuam na área de ES foi feita também uma comparação com a produção científica de outros países e, uma análise do impacto do trabalho de pesquisa da área de ES desenvolvido no Brasil com a de outros países.

6.7.1. Coleta de dados

Inicialmente, foi obtida uma lista de pesquisadores, docentes e profissionais com doutorado registrado na PL. No dia 3 de dezembro de 2012 havia 166.738 doutores brasileiros registrados na PL. O número de identificação (ID) desses doutores foi obtido e com essa identificação, seus currículos Lattes foram obtidos. A linguagem LattesMiner desempenhou papel fundamental em todas essas tarefas.

Um arquivo texto contendo o ID de cada doutor foi inicialmente criado. Em seguida, utilizando a linguagem LattesMiner os números de identificação foram lidos e os currículos Lattes baixados. Os currículos Lattes foram armazenados como arquivos HTML e o nome do arquivo foi salvo junto com o ID de cada doutor.

Todos os 166.738 currículos dos doutores brasileiros foram baixadas em 2 dias ocupando 25,9 GB de espaço em disco para armazená-los. Com os currículos

de todos os doutores a extração de dados foi realizada. A lista de pesquisadores analisados foi construída selecionando um doutor se o termo “ES” (em Português ou em Inglês) aparecia em seu currículo Lattes pelo menos 10 vezes. Com o critério de classificação proposto, 611 dos 166.738 doutores brasileiros foram selecionados. Utilizando como critério pesquisadores que apareceram pelo menos uma vez como autor ou coautor de um artigo publicado no SBES ou em qualquer outro evento em que o termo ES (em Português ou em Inglês) aparece em seu título, e também com pelo menos um artigo publicado em periódicos classificados na categoria CSSE do JCR® foi encontrado um total de 190 doutores. Em 7 dezembro de 2012 o currículo Lattes desses 190 doutores foram baixados utilizando a linguagem LattesMiner. Embora desnecessário, esses currículos foram baixados novamente pois alguns pesquisadores poderiam ter atualizado seu currículo Lattes nos últimos 4 dias e, essa tarefa é muito simples de ser realizada utilizando a linguagem LattesMiner.

As informações a seguir foram extraídas automaticamente de cada um dos 190 currículos Lattes dos doutores e armazenadas em um banco de dados utilizando a linguagem LattesMiner: informações pessoais, endereço profissional, formação acadêmica, produção científica em congressos e periódicos, orientações de mestrado e doutorado e contatos (todos os números de identificação dos pesquisadores citados em seu currículo Lattes).

6.7.2. Resultados e discussões

175 (92,11%) dos 190 doutores atualizaram seus currículos Lattes em 2012. Isso pode causar alguma diferença na análise realizada nesse ano em comparação aos anos anteriores. De acordo com o gênero, 133 (70%) são do sexo masculino e 57 (30%) são doutores do sexo feminino. Na Tabela 6.42 é apresentada a distribuição dos doutores da área de ES com bolsa de Produtividade em Pesquisa (PQ) do CNPq por categoria e gênero.

Tabela 6.42 - Distribuição dos doutores da área de ES de acordo com a categoria e gênero.

Categoria	n	Masculino	Feminino
1A	2	1	1
1B	2	2	-
1C	7	6	1
1D	11	8	3
2	40	26	14
Total	**62**	**43**	**19**

O número total de doutores da área de ES com PQ é 62. Desse total, a grande maioria é do sexo masculino (69,35%) e da categoria 2 (64,52%). 57 dos 62 doutores com PQ são da área de Ciência da Computação (CC) no CNPq e representam 15,49% dos 368 doutores da área. Esses valores percentuais são muito semelhantes ao que se observa na área de CC (sexo masculino (73,91%) e da categoria 2 (65,76%)).

183 (96,32%) dos 190 doutores informaram o seu endereço profissional. A predominância absoluta dos endereços está na região Sudeste (50,82%), 41 (22,41%) trabalham no estado do Rio de Janeiro, 34 (18,58%) trabalham no estado de São Paulo, 16 (8,74%) trabalham no estado de Minas Gerais e 2 (1,09%) trabalham no estado de Espírito Santo. A distribuição geográfica dos doutores da área de ES de acordo com a cidade e o estado do endereço profissional é apresentada na Figura 6.25.

É possível observar que há uma concentração de doutores da área de ES na cidade do Rio de Janeiro (36, 19,67%). Outras cidades com número grande de doutores da área de ES são “São Carlos” com 19 (10,38%), “Recife” com 16 (8,74%), “Porto Alegre” com 15 (8,20%) e “Campina Grande” com 10 (5,46%). “São Paulo” e “Campinas” aparecem em seguida com 5 doutores cada. As cidades com um único doutor da área de ES são apresentadas na cor cinza (não indicada na escala apresentada). Os 183 doutores estão distribuídos em 19 estados em todas as cinco regiões do Brasil.

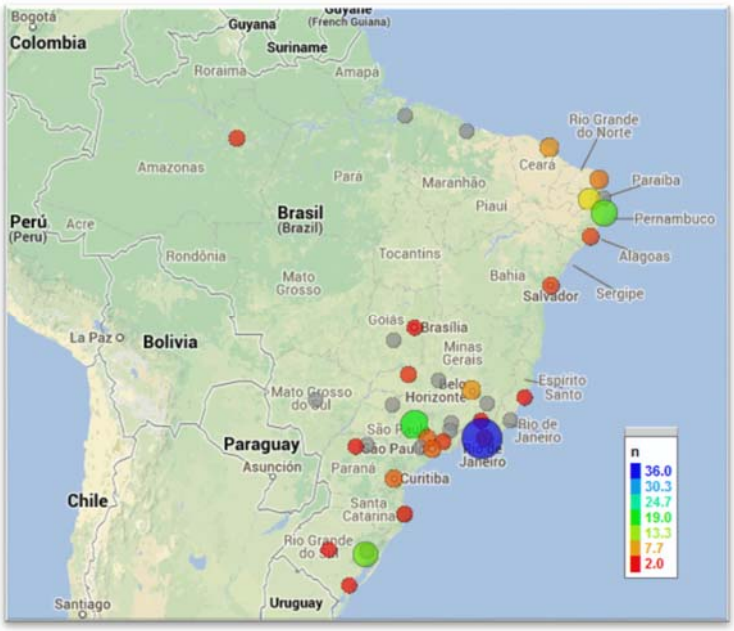


Figura 6.25 - Distribuição geográfica dos doutores da área de ES de acordo com a cidade e o estado do endereço profissional.

Os bolsistas PQ da área de ES concluíram seu doutorado entre 5 e 38 anos, com média de 15,89 anos. Essa média é maior do que a média dos bolsistas PQ da área de CC que é 14,84 anos. Os 190 doutores da área de ES concluíram seu doutorado há 12,08 anos atrás, em média. Na Figura 6.26 é apresentada a distribuição geográfica dos doutores da área de ES com o tempo de conclusão de seu doutorado maior que 25 anos. As cores representam o tempo médio de conclusão do doutorado de acordo com a cidade e o estado do seu endereço profissional e o tamanho dos círculos são proporcionais ao número de doutores em cada cidade.

Os doutores (19) que concluíram seu doutorado há mais de 25 anos atrás estão distribuídos somente em 5 estados e contribuíram para a disseminação da ES no Brasil e, provavelmente, foram responsáveis pelo crescimento da área no Brasil. O “Rio de Janeiro” é a cidade com mais doutores (8) e a cidade com doutores que concluíram seu doutorado há mais tempo (38 anos). Admitindo-se que esses doutores não alteraram os endereços profissionais durante suas carreiras, poder-se-ia afirmar que, desde o início, os doutores da área de ES já se concentravam na região Sudeste.

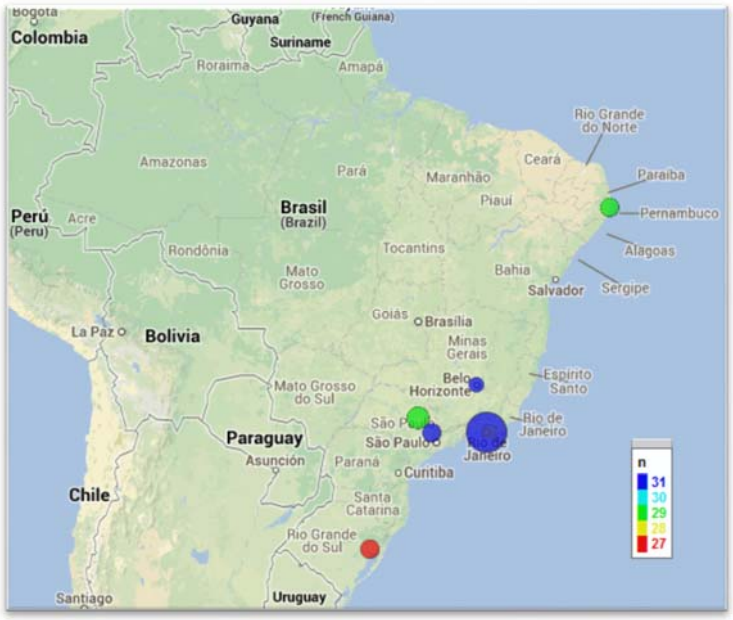


Figura 6.26 - Distribuição geográfica dos doutores da área de ES com tempo de conclusão de doutorado maior que 25 anos.

Atualmente, os doutores recentes (47) da área de ES estão concentrados na região Sudeste, como ilustra a Figura 6.27, mas há doutores em todas as regiões do país em 14 estados. “Rio de Janeiro” é a cidade com mais doutores recentes da área de ES (6).

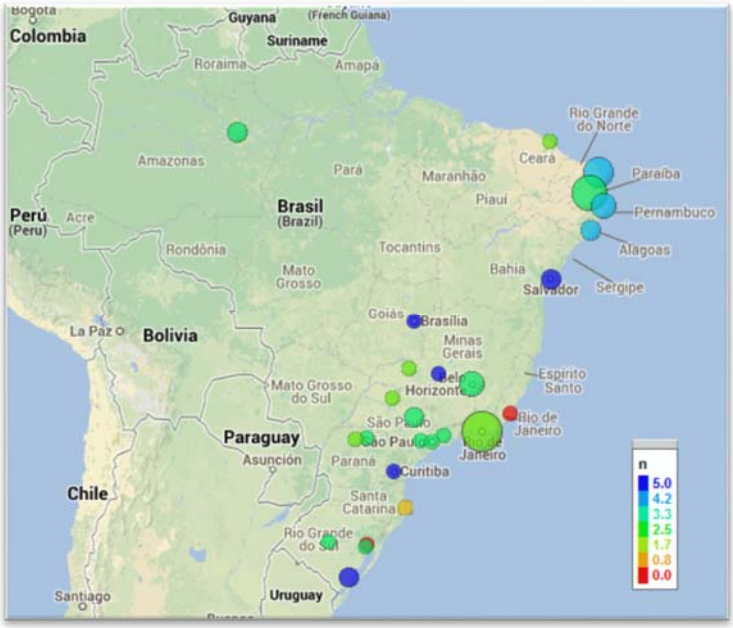


Figura 6.27 - Distribuição geográfica dos doutores da área de ES com tempo de conclusão do doutorado menor ou igual a 5 anos.

De acordo com seus currículos Lattes o número total de artigos publicados pelos 190 doutores da área de ES em periódicos no período de 1987 a 2011 foi 2.312. O número total de artigos publicados em congressos no mesmo período foi 12.237. Na Figura 6.28 é apresentado o número de artigos publicados em periódicos e congressos no período de 1987 a 2011. O número de doutores (entre os 190) envolvidos nestas publicações também é apresentado. O número de artigos publicados em congressos é claramente muito maior do que o número de artigos publicados em periódicos. Nesse período, a razão foi de 5,29 artigos em congressos para cada artigo publicado em periódico, em média. Essa média diminuiu nos últimos anos. Em 2011, foram publicados 3,52 artigos em congressos para cada artigo publicado em periódico.

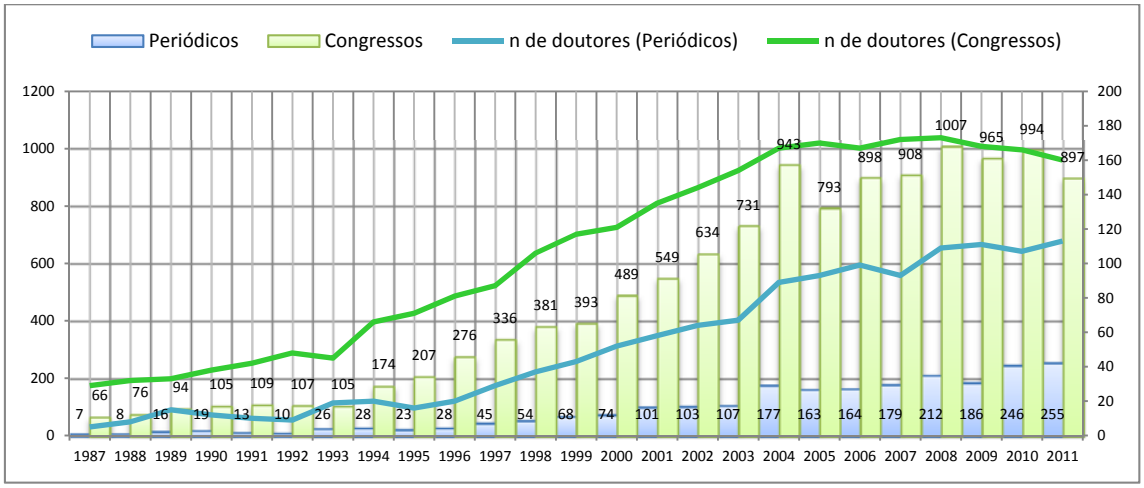


Figura 6.28 - Número de artigos publicados pelos doutores da área de ES em periódicos e congressos no período de 1987 a 2011.

O maior número de artigos e autores que publicaram em periódicos foram de doutores da cidade do “Rio de Janeiro”, como pode ser observado na Figura 6.29. Esses doutores (36) publicaram 28,81% dos artigos. Foram considerados somente 2.239 artigos publicados pelos 183 doutores que informaram o seu endereço profissional. As cidades com doutores que publicaram menos que 25 artigos estão representadas na cor cinza. Cinco cidades (Rio de Janeiro, São Carlos, Porto Alegre, Recife e Campina Grande) contribuíram com 65,52% dos artigos publicados em periódicos.

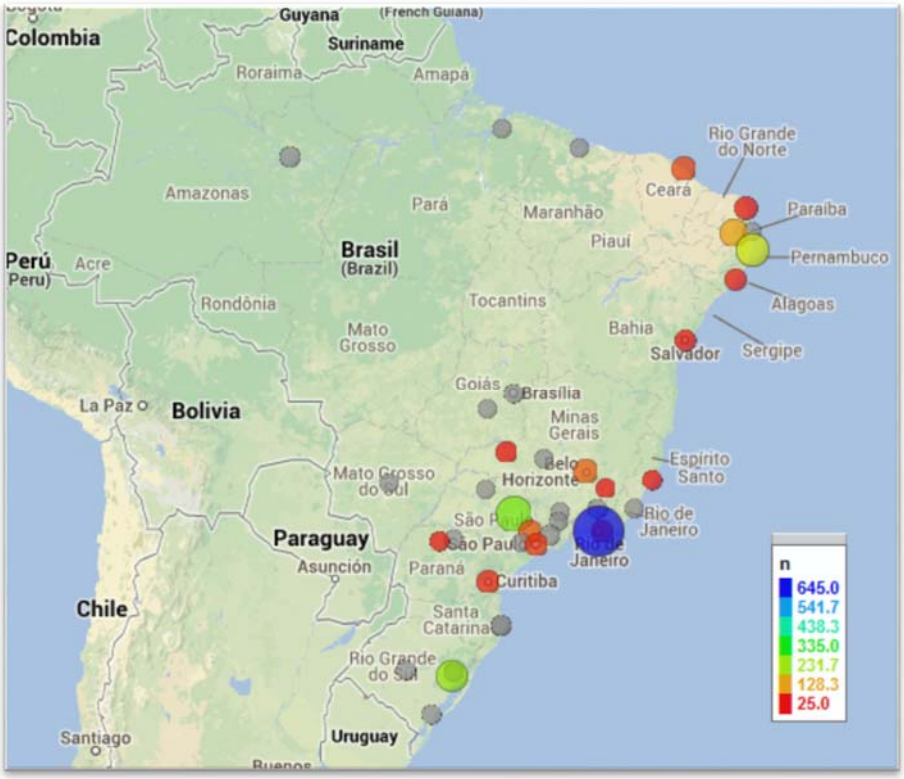


Figura 6.29 - Distribuição geográfica dos doutores da área de ES de acordo com o número de artigos publicados em periódicos no período de 1987 a 2011.

Na Tabela 6.43 é apresentada a lista de doutores que publicaram mais artigos em periódicos classificados na categoria CSSE do JCR® no período de 1987 a 2011. Também é apresentada a média de autores por artigo. Antes do nome de cada pesquisador é indicado se ele/ela se classifica como um expert em ES em seu currículo Lattes (S (Sim) ou N (Não)). Entre parênteses, é apresentado o número de vezes que cada pesquisador citou o termo ES (em Português ou em Inglês) em seu currículo Lattes. Também é apresentado o número de artigos publicados nos últimos 5 anos (2007 a 2011), para dar uma indicação se o doutor continua ativo nessa área. O melhor ano (em termos de número de artigos publicados) de cada doutor também é apresentado.

Tabela 6.43 - Doutores da área de ES que publicaram mais artigos em periódicos classificados na categoria CSSE do JCR® no período de 1987 a 2011.

Nome	Cat.	Artigos	%	Autores	2007 a 2011	%	Melhor Ano	n
S Carlos José Pereira de Lucena (302)	1A	22	73,33	3,45	10	45,45	2008	5
S José Carlos Maldonado (357)	1B	21	91,30	4,52	8	38,10	2006	5
S Maria Emilia Xavier Mendes (134)	-	16	84,21	3,19	9	56,25	2010	3
S Alessandro Fabricio Garcia (262)	2	13	100	4,69	8	61,54	2011	3
S Guilherme Horta Travassos (314)	1D	12	85,71	3,75	7	58,33	2004	3
S Maria Cristina Ferreira de Oliveira (31)	1D	12	60,00	5,17	6	50,00	2007	2
S Cecília Mary Fischer Rubira (54)	1D	9	100	4,00	4	44,44	2009	3
S Augusto Cézar Alves Sampaio (42)	1C	9	69.23	3,11	4	44,44	2010	4
S Julio César Sampaio Prado Leite (140)	1C	9	81,82	4,11	-	-	2005	2
S Ana Lúcia Caneca Cavalcanti (33)	-	8	88,89	4,38	5	62,50	2010	3
S Paulo Henrique Monteiro Borba (136)	2	8	100	2,75	2	25,00	2011	1
S Cláudia Maria Lima Werner (275)	1D	8	72,73	3,00	5	62,50	2007	2
S Daniel Schwabe (19)	1C	8	61,54	3,63	-	-	2002	3
S Eduardo Santana de Almeida (207)	2	7	100	5,29	7	100	2011	4
S Marco Túlio de Oliveira Valente (29)	1D	7	87,50	3,29	6	85,71	2009	2
N Fábio Kon (17)	1D	7	77,78	4,57	2	28,57	2011	1
S Márcio Eduardo Delamaro (158)	-	7	100	3,71	1	14,29	2001	3
S Silvia Regina Vergilio (91)	2	7	70,00	3,29	4	57,14	2010	2
N Cláudia Maria Bauzer Medeiros (10)	1A	7	30,43	2,71	2	28,57	2008	2
S Paulo César Masiero (164)	1D	7	77,78	3,29	3	42,86	1999	2
S Auri Marcelo Rizzo Vincenzi (102)	2	6	100	4,00	1	16,67	2001	2
S André Luís de Medeiros Santos (12)	-	6	100	3,67	3	50,00	2011	2
N Marta Lima de Queiros Mattoso (19)	1C	6	42,86	6,50	4	66,67	2011	2
N Marco Antônio Casanova (13)	1C	6	40,00	3,50	3	50,00	1992	2
S Rafael Prikladnicki (224)	-	5	100	3,80	5	100	2010	4

“Carlos José Pereira de Lucena” foi o doutor que publicou mais artigos em periódicos no período de 1987 a 2011. Ele publicou 22 (73,33%) artigos em periódicos classificados na categoria CSSE entre os 29 artigos publicados em

periódicos indexados no JCR®. O número total de artigos que ele publicou no período foi de 100, sendo 68 deles com ISSN. Observa-se que 4 dos 25 doutores não classificam a si mesmo como um expert em ES em seu currículo Lattes e 3 deles publicaram mais artigos (porcentagem) em outras categorias, em comparação com a categoria CSSE. Todos os outros 22 doutores publicaram pelo menos 60% de suas publicações com JCR® em periódicos classificados na categoria CSSE. Entre esses 25 doutores, 17 (68%), publicaram mais artigos em periódicos classificados na categoria CSSE no período de 2007 a 2011.

Na Tabela 6.44 é apresentada uma lista de 25 periódicos classificados na categoria CSSE do JCR® em que os doutores publicaram mais artigos no período de 1987 a 2011. Essa lista foi obtida utilizando o ISSN dos periódicos informados nos currículos Lattes dos doutores da área de ES. Também é apresentado o número médio de autores dos artigos publicados em cada periódico. O número de artigos publicados nos últimos 5 anos (2007 a 2011) também é apresentado, permitindo verificar se o periódico continua sendo utilizado por essa comunidade. O FI do JCR® de 2011 também é indicado. Entre parênteses, é apresentada a posição relativa em termos do FI do periódico entre os periódicos (104) classificados na categoria CSSE pelo JCR®.

"*Journal of the Brazilian Computer Society*" foi o periódico mais utilizado pelos doutores da área de ES. Esse periódico não aparece na lista pois não é indexado no JCR®. O número de artigos publicados nesse periódico foi 486 e representa 22,86% (de um total de 2.126) das publicações em periódicos com ISSN. O número médio de autores nos artigos publicados em periódicos no período de 1987 a 2011 é 3,80. Os 3 primeiros periódicos da Tabela 6.44 foram mais utilizados no período de 2007 a 2011, contribuindo para suas classificações no topo da lista no período de 1987 a 2011.

Tabela 6.44 - Periódicos classificados na categoria CSSE do JCR® em que os doutores da área de ES publicaram mais artigos no período de 1987 a 2011.

ISSN	Periódico	Artigos	Autores	2007 a 2011	%	FI
0164-1212	Journal of Systems and Software (57/104)	93	3,75	53	56,99	0,836
0948-695X	Journal of Universal Computer Science (85/104)	49	3,92	39	79,59	0,398
0950-5849	Information and Software Technology (29/104)	39	3,90	29	74,36	1,250
0038-0644	Software, Practice & Experience (76/104)	39	4,05	13	33,33	0,519
0167-6423	Science of Computer Programming (70/104)	25	3,28	13	52,00	0,622
0098-5589	IEEE Transactions on Software Engineering (6/104)	24	2,96	8	33,33	1,980
1532-0626	Concurrency and Computation (68/104)	23	4,35	18	78,26	0,636
0740-7459	IEEE Software (18/104)	22	4,36	16	72,73	1,508
1382-3256	Empirical Software Engineering (11/104)	17	6,12	6	35,29	1,854
1751-8806	IET Software (90/104)	17	3,53	16	94,12	0,329
0947-3602	Requirements Engineering (43/104)	16	4,81	8	50,00	0,971
0218-1940	Int. Journal of Soft. Eng. and Knowledge Eng. (99/104)	15	4,73	8	53,33	0,129
0960-0833	Software Testing, Verification & Reliability (46/104)	12	3,17	4	33,33	0,957
1380-7501	Multimedia Tools and Applications (71/104)	12	3,50	7	58,33	0,617
0963-9314	Software Quality Journal (84/104)	11	3,09	3	27,27	0,417
1619-1366	Software and Systems Modeling (37/104)	9	3,22	3	33,33	1,061
0934-5043	Formal Aspects of Computing (80/104)	9	4,00	6	66,67	0,463
0001-0782	Communications of the ACM (9/104)	8	2,63	3	37,50	1,919
0010-4620	Computer Journal (60/104)	8	2,50	3	37,50	0,785
0018-9162	Computer (19/104)	7	3,86	7	100	1,470
0018-9529	IEEE Transactions on Reliability (26/104)	7	3,86	5	71,43	1,285
0163-5808	SIGMOD Record (65/104)	7	7,29	2	28,57	0,667
0929-5585	Design Automation for Embedded Systems (95/104)	7	5,86	6	85,71	0,200
0362-1340	ACM SIGPLAN Notices (101/104)	6	2,67	-	-	0,090
1741-1106	Int. Journal of Web and Grid Services (9/104)	4	4,75	1	25,00	1,919

Dos 2.126 (91,96% do total) dos artigos publicados em periódicos com ISSN no período de 1987 a 2011, 903 (42,47%) foram em periódicos com FI no JCR® e, 546 (60,47%) em periódicos classificados na categoria CSSE. A distribuição dessas publicações nesse período é apresentada na Figura 6.30. Pode-se observar um crescente aumento no número de artigos publicados em periódicos indexados no JCR® nos últimos anos. É interessante observar que a razão entre o número de artigos em periódicos classificados na categoria CSSE em relação aos artigos publicados em periódicos indexados no JCR® permanece em torno de 60% ao longo dos anos.

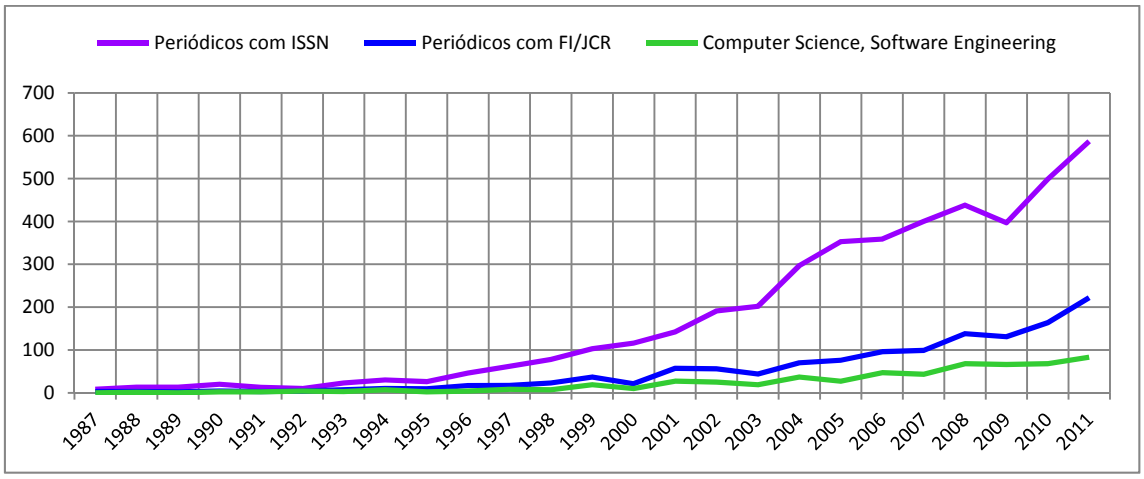


Figura 6.30 - Número de artigos publicados pelos doutores da área de ES em periódicos com ISSN, em periódicos indexados no JCR® e da categoria "*Computer Science, Software Engineering*" no período de 1987 a 2011.

Na Figura 6.31 é apresentada a distribuição geográfica dos países das editoras dos periódicos em que os doutores da área de ES publicaram no período de 1987 a 2011. O país das editoras de cada periódico foi obtido a partir do JCR®. Os periódicos (212) são de 13 países, com predomínio de periódicos dos "Estados Unidos" (358 artigos em 85 periódicos), "Holanda" (222 artigos em 47 periódicos) e "Inglaterra" (197 artigos em 47 periódicos).

Apenas em editoras de 6 (46,15%) países os doutores publicaram mais que 10 artigos no período de 1987 a 2011. O número de artigos publicados em periódicos dos "Estados Unidos" representa 39,65% do total de artigos com

JCR®, e 143 (75,26%) doutores da área de ES publicaram pelo menos 1 artigo em algum periódico desse país. O número de artigos publicados (6) em periódicos brasileiros (5) é muito pequeno, provavelmente porque a maioria dos periódicos brasileiros não está indexado no JCR® e não há nenhum periódico brasileiro classificado na categoria CSSE. As cores representam o número de artigos publicados de acordo com o país da editora e o tamanho dos círculos é proporcional ao número de periódicos em cada país.

Figura 6.31 - Distribuição geográfica dos países das editoras dos periódicos em que os doutores da área de ES publicaram no período de 1987 a 2011.

Considerando o número de artigos publicados em congressos (11.937) no período de 1987 a 2011 pelos 183 doutores que informaram o seu endereço profissional, “Rio de Janeiro” também é a cidade que publicou mais artigos (3.375; 28,27%) e também, é a cidade com mais doutores (36; 19,67%) que publicaram nesse período, como pode ser observado na Figura 6.32. É interessante observar que as mesmas 5 cidades que publicaram mais artigos em periódicos também publicaram mais artigos em congressos. Essas cidades publicaram juntas 65,64% dos artigos. Essa porcentagem é muito semelhante a porcentagem em periódicos.

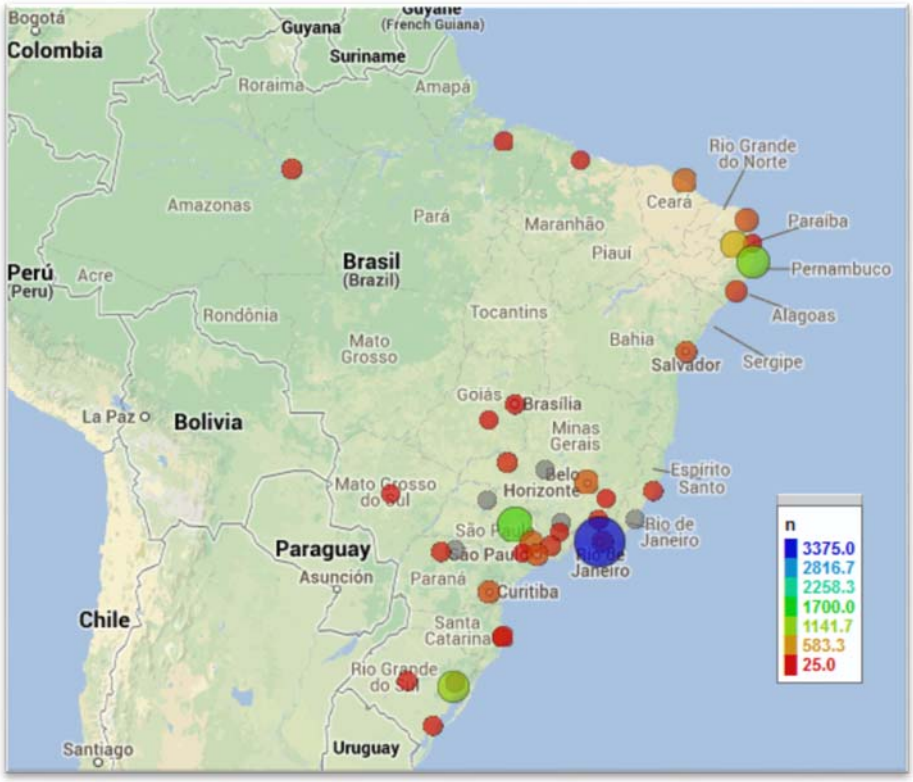


Figura 6.32 - Distribuição geográfica dos doutores da área de ES de acordo com o número de artigos publicados em congressos no período de 1987 a 2011.

Na Tabela 6.45 é apresentada uma lista dos doutores que publicaram mais artigos no SBES ou em congressos relacionados com ES no período de 1987 a 2011. No caso de empate foi considerado o ano de conclusão do doutorado. É importante mencionar que apenas a quantidade foi considerada.

O pesquisador com o maior número de publicações no SBES ou em congressos relacionados com ES no período de 1987 a 2011 também é “Carlos José Pereira de Lucena”. Ele publicou 400 artigos em congressos nesse período e 91 dessas publicações foram no SBES ou em congressos relacionados com ES. Entre os 25 doutores, 18 (72%) publicaram mais artigos no SBES ou em congressos relacionados com ES no período de 1987 a 2011.

Tabela 6.45 - Doutores da área de ES que publicaram mais artigos no SBES ou em congressos relacionados com ES no período de 1987 a 2011.

Nome	Cat.	Artigos	%	Autores	2007 a 2011	%	Melhor Ano	n
Y Carlos José Pereira de Lucena (302)	1A	91	22,75	3,90	30	32,97	2006	16
Y Guilherme Horta Travassos (314)	1D	79	38,92	3,04	29	36,71	2009	10
Y Cláudia Maria Lima Werner (275)	1D	75	37,69	3,45	20	26,67	2004	9
Y José Carlos Maldonado (357)	1B	75	40,11	3,44	17	22,67	2008	8
Y Silvio Romero de Lemos Meira (120)	-	61	29,05	3,61	19	31,15	2008	6
Y Alessandro Fabricio Garcia (262)	2	48	32,88	4,65	24	50,00	2006	10
Y Paulo César Masiero (164)	1D	41	30,37	4,15	13	31,71	2001	4
Y Eduardo Santana de Almeida (207)	2	36	31,30	4,11	25	69,44	2011	6
Y Uira Kulesza (132)	2	35	38,89	6,17	19	54,29	2009	6
Y Maria Emilia Xavier Mendes (134)	-	34	36,17	2,79	20	58,82	2010	10
Y Antônio Francisco do Prado (214)	-	34	20,99	3,44	5	14,71	2002	7
Y Ana Regina Cavalcanti da Rocha (62)	-	34	18,38	3,53	4	11,76	1999	9
Y Rafael Prikladnicki (224)	-	33	42,86	3,18	24	72,73	2007	8
Y Márcio de Oliveira Barros (117)	2	32	50,00	2,91	12	37,50	2010	5
Y Roberto Tom Price (57)	-	29	37,66	2,69	3	10,34	1990	5
Y Júlio César Sampaio Prado Leite (140)	1C	27	18,88	3,04	5	18,52	1998	4
Y Manoel Gomes Mendonça Neto (205)	-	26	40,00	3,85	12	46,15	2009	3
Y Leonardo Gresta Paulino Murta (100)	2	24	26,67	4,25	9	37,50	2007	4
Y Jorge César Abrantes Figueiredo (60)	-	24	24,24	3,29	4	16,67	2003	4
Y Mário Jino (47)	-	24	32,00	3,38	-	-	1999	5
Y Arndt von Staa (87)	2	24	44,44	4,42	5	20,83	2006	3
Y Angelo Perkusich (68)	-	23	11,11	3,74	3	13,04	2006	4
Y Jorge Luis Nicolas Audy (66)	-	22	18,64	2,73	12	54,55	2007	6
Y Tayana Uchôa Conte (104)	-	20	47,62	4,20	16	80,00	2011	8
Y Silvia Regina Vergilio (91)	2	20	21,98	3,20	8	40,00	2007	3

O número médio de autores por artigo publicado em congressos no período de 1987 a 2011 é 3,66. Essa média é muito semelhante ao número médio de autores por artigo publicado em periódicos (3,80) no mesmo período. No período de 1987 a 1991, o número médio de autores por artigo publicado era bem menor, 2,26 em congressos e 2,57 em periódicos. No período de 2007 a 2011, essas médias aumentaram consideravelmente, para 4,13 em congressos e para 4,16 em periódicos. Esse aumento pode ser observado na Figura 6.33.

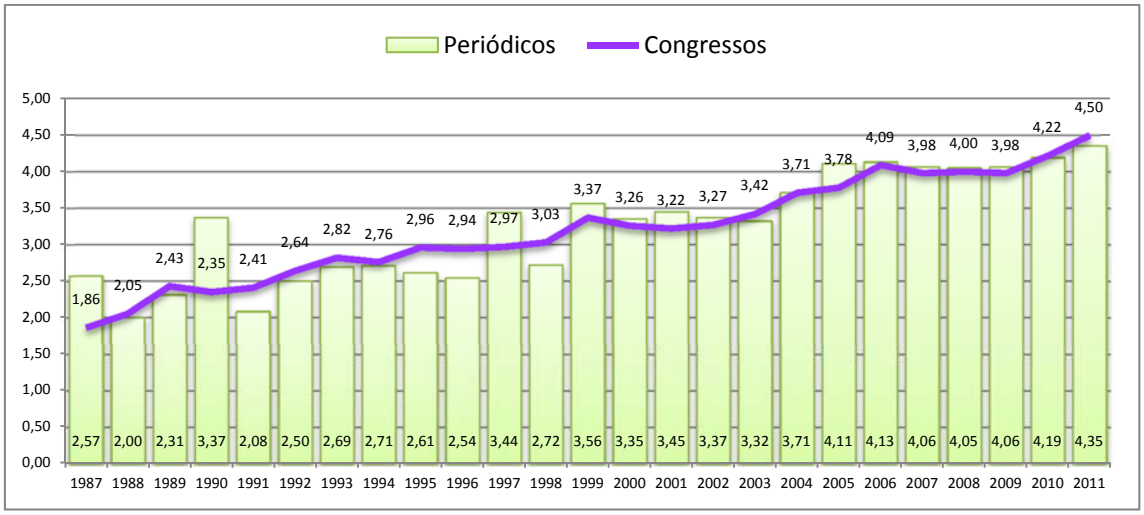


Figura 6.33 - Número médio de autores por artigo publicado pelos doutores da área de ES em periódicos e congressos no período de 1987 a 2011.

Utilizando a linguagem LattesMiner também é possível identificar todos os contatos de cada um dos doutores da área de ES. Na Figura 6.34 é apresentada a página inicial do sistema SUCUPIRA com a distribuição geográfica dos 25 doutores da área de ES com mais contatos no grupo, de acordo com o endereço profissional deles informado em seus currículos Lattes. Pode-se observar que a maioria deles está localizado na região Sudeste (14), estão distribuídos em 7 estados e “São Paulo” é o estado com mais doutores (7).

Figura 6.34 - Distribuição geográfica dos 25 doutores da área de ES com mais contatos distintos entre todos os doutores dessa área.

Na Figura 6.35 é apresentado o grafo de contatos desses 25 doutores. Nessa rede social acadêmica cada vértice tem um rótulo com o nome do doutor e é colorido de acordo com a sua categoria PQ: a cor azul indica a categoria 1A, a cor verde a categoria 1B, amarelo a categoria 1C, cor laranja a categoria 1D e a cor vermelha a categoria 2. A cor cinza é utilizada para representar os doutores sem bolsa PQ. O tamanho dos vértices indica o número de contatos de cada doutor. As cores das arestas representam o número de relacionamentos entre os doutores, sendo que a intensidade da cor indica o número de relacionamentos. Nessa rede apenas os relacionamentos que apareceram pelo menos 10 vezes são apresentados e é possível visualizar relacionamentos entre os doutores com um grau de separação igual a 1. Apenas um desses doutores não tem nenhum contato com qualquer um dos outros doutores desse grupo.

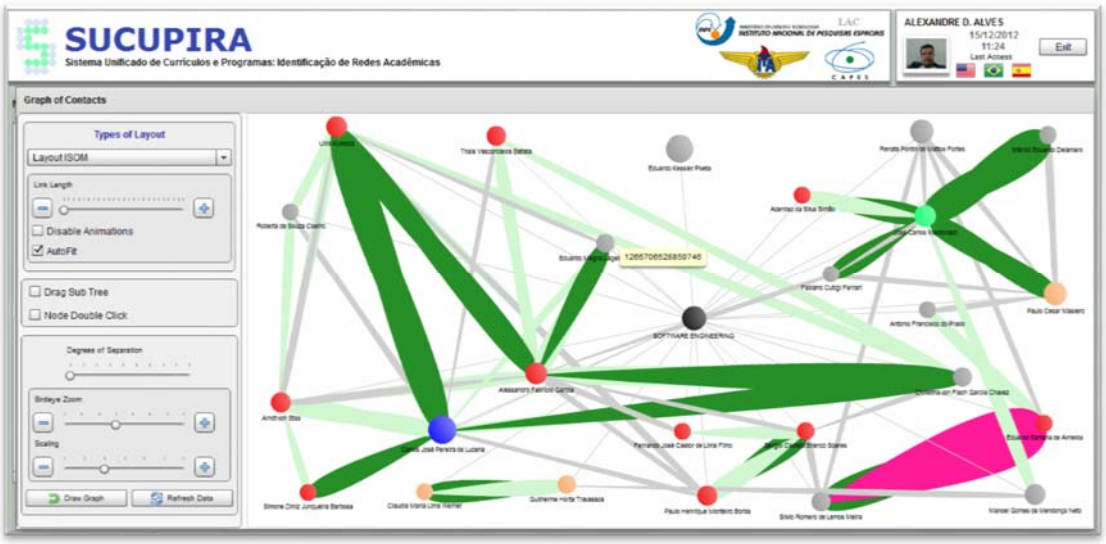


Figura 6.35 - Grafo de contatos dos 25 doutores da área de ES com mais contatos distintos entre todos os doutores dessa área.

O principal relacionamento é destacado por uma aresta na cor rosa e representa relacionamentos que ocorreram entre 100 e 200 vezes. O relacionamento entre os doutores “Eduardo Santana de Almeida” e “Silvio Romero de Lemos Meira” é representado nessa cor. É interessante observar que o relacionamento entre esses doutores não é recíproco em relação à intensidade. “Eduardo Santana de Almeida” informou o relacionamento no seu

currículo Lattes 146 vezes e “Silvio Romero de Lemos Meira” informou 85 vezes. “Silvio Romero de Lemos Meira” foi orientador de doutorado de “Eduardo Santana de Almeida” e ele teve relacionamentos com outros 129 pesquisadores, sendo 18 deles doutores considerados neste estudo. “Eduardo Santana de Almeida” teve relacionamentos com 45 pesquisadores e 12 deles são doutores da área de ES.

Os relacionamentos importantes são destacados por uma aresta na cor verde escuro e representam relacionamentos que ocorreram entre 50 e 100 vezes. Foram identificados os seguintes relacionamentos: “Alessandro Fabricio Garcia” e “Uira Kulesza” (recíproco), “Christina von Flach Garcia Chavez” e “Alessandro Fabricio Garcia”, “Eduardo Magno Lages Figueiredo” e “Alessandro Fabricio Garcia”, “José Carlos Maldonado” e “Márcio Eduardo Delamaro” (recíproco), “José Carlos Maldonado” e “Paulo César Masiero” (recíproco), “José Carlos Maldonado” e “Adenilso da Silva Simão”, “Fabiano Cutigi Ferrari” e “José Carlos Maldonado”, “Carlos José Pereira de Lucena” e “Uira Kulesza” (recíproco), “Christina von Flach Garcia Chavez” e “Carlos José Pereira de Lucena”, “Cláudia Maria Lima Werner” e “Guilherme Horta Travassos”, “Silvio Romero de Lemos Meira” e “Eduardo Santana de Almeida”, “Simone Diniz Junqueira Barbosa” e “Carlos José Pereira de Lucena” e dos doutores “Sérgio Castelo Branco Soares” e “Paulo Henrique Monteiro Borba”.

Nessa rede também é possível observar os relacionamentos entre os doutores “Antônio Francisco do Prado”, “Fabiano Cutigi Ferrai”, “Adenilso da Silva Simão”, “Renata Pontim de Mattos Fortes”, “Márcio Eduardo Delamaro”, “Paulo César Masiero” e “José Carlos Maldonado”. Além de serem todos da mesma área, todos eles moram na cidade de “São Carlos” e, todos trabalham na mesma instituição, exceto os dois primeiros. Outro grupo identificado nessa rede é “Alessandro Fabricio Garcia”, “Carlos José Pereira de Lucena”, “Arndt von Staa”, “Simone Diniz Junqueira Barbosa”, “Guilherme Horta Travassos” e “Cláudia Maria Lima Werner”. Todos eles moram na cidade do “Rio de Janeiro”, os quatro primeiros trabalham na mesma instituição e os outros dois em outra instituição. Outro grupo identificado é “Paulo Henrique Monteiro Borba”, “Sérgio

Castelo Branco Soares", "Fernando José Castor de Lima Filho" e "Silvio Romero de Lemos Meira". Eles moram na cidade de "Recife" e trabalham na mesma instituição.

"Carlos José Pereira de Lucena" é o doutor com mais contatos no grupo. De 1987 a 2011, ele teve relacionamentos com 66 (10,80%) doutores da área de ES. O número total de relacionamentos com esses doutores foi 1.285. Seu principal relacionamento foi com "Hugo Fuks". Ambos moram na cidade do "Rio de Janeiro" e trabalham na mesma instituição. Esse relacionamento ocorreu 134 vezes e não é apresentado na rede porque "Hugo Fuks" não está entre os 25 doutores com mais contatos.

Para obter a orientação acadêmica dos doutores foi considerado o número de identificação (ID) do orientador informado nos currículos Lattes. O grafo de orientações dos 25 doutores da área de ES com mais contatos é apresentado na Figura 6.36. O tamanho dos vértices é proporcional ao número de alunos orientados por cada doutor. A intensidade da cor das arestas indica se é uma orientação de mestrado, doutorado ou ambas.

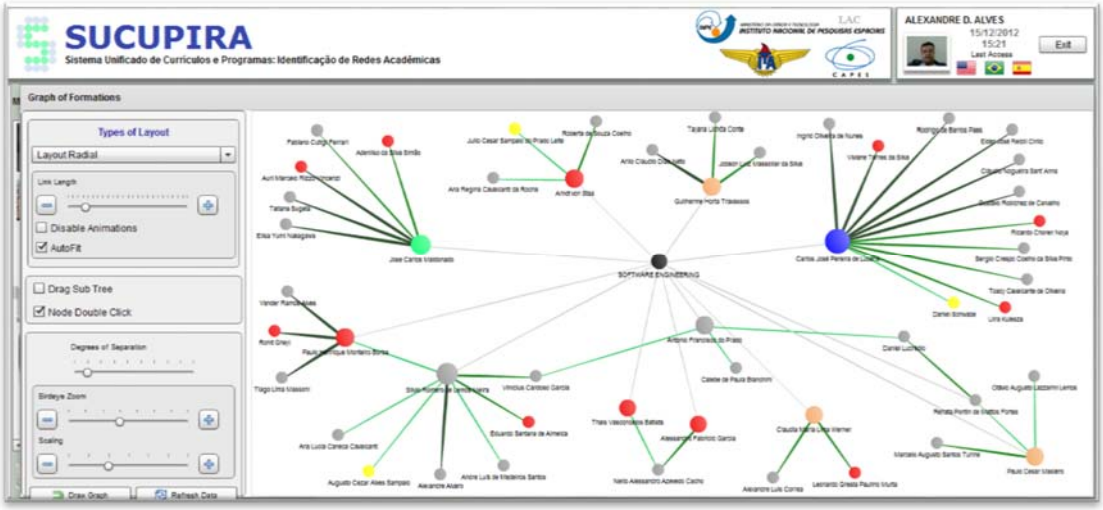


Figura 6.36 - Grafo de orientações dos 25 doutores da área de ES com mais contatos distintos entre todos os doutores dessa área.

"Carlos José Pereira de Lucena" também é o doutor que orientou mais alunos no período de 1987 a 2011. Ele orientou 11 alunos (5 mestrados e 10 doutorados) que atualmente estão na lista dos doutores da área de ES. Os 5

doutores que orientaram mais alunos formaram 29 (15,26%) doutores da área de ES. 75 (39,47%) dos 190 doutores foram orientados por algum doutor da área de ES. Esse número pode ser muito maior uma vez que apenas orientações com o ID do orientador informado no currículo Lattes foram considerados.

Na Figura 6.37 é apresentada a distribuição geográfica desses 75 doutores, destacando sua migração entre os estados brasileiros. De um total de 75 doutores, 44 (58,67%) doutores continuam no mesmo estado de sua orientação. No mapa, os estados marcados com quadrados, são aqueles que tiveram migração para outros estados. O número nos quadrados indica a quantidade de doutores que continuaram no estado. As migrações (orientações nos arcos) são representadas por três cores: a preta representa apenas 1 migração, a azul representa 2 migrações e a verde representa 3 migrações.

Figura 6.37 - Distribuição geográfica dos doutores orientados por algum doutor da área de ES.

“Rio de Janeiro” é o estado em que mais doutores foram formados. De um total de 30 orientações, 18 (60%) doutores continuaram no estado. Os outros 12 doutores estão em 7 estados em quatro regiões do Brasil. No estado de “São Paulo” foram formados 19 doutores. 11 (57,89%) continuaram no estado e 8 estão em outros 7 estados em quatro regiões do Brasil. No estado de “Pernambuco” foram formados 11 doutores. 4 (36,35%) continuaram no estado e 7 estão em outros 6 estados em quatro regiões do Brasil. Pode-se observar que os doutores que se formaram em ES migraram para outros estados e também para outras regiões do Brasil.

A seguir é comparada a produção científica brasileira na categoria CSSE definida na WoS com a produção de outros países. É analisado o impacto do trabalho de pesquisa da área de ES desenvolvido no Brasil comparando com o número de artigos e citações de outros países. O Brasil ocupa a vigésima primeira (21ª) posição no ranking mundial dos países em termos de artigos publicados em periódicos classificados na categoria CSSE no período de 1987 a 2011, conforme apresentado na Tabela 6.46. Em termos do número médio de citações por artigo, o Brasil passa para a décima oitava posição (18ª).

Tabela 6.46 - Países que publicaram mais artigos em periódicos classificados na categoria CSSE do JCR® no período de 1987 a 2011.

Posição	País	Artigos (A)	Citações (C)	C / A	Posição
1ª	Estados Unidos	43.127	550.811	12,77	2ª ↓1
2ª	China	8.063	42.432	5,26	22ª ↓20
3ª	Inglaterra	7.751	71.784	9,26	8ª ↓5
4ª	Alemanha	7.602	66.452	8,74	12ª ↓8
5ª	Japão	6.281	32.410	5,16	23ª ↓18
6ª	França	5.751	54.169	9,42	7ª ↓1
7ª	Canadá	5.726	58.261	10,17	5ª ↑2
8ª	Itália	4.999	43.575	8,72	13ª ↓5
9ª	Coréia do Sul	4.017	20.422	5,08	24ª ↓15
10ª	Espanha	3.606	19.043	5,28	21ª ↓11
11ª	Austrália	3.391	26.679	7,87	14ª ↓3
12ª	Taiwan	3.313	21.320	6,44	17ª ↓5
13ª	Holanda	2.827	27.214	9,63	6ª ↑7
14ª	Israel	2.068	35.992	17,40	1ª ↑13
15ª	Suíça	1.753	18.797	10,72	3ª ↑12
16ª	Índia	1.744	9.993	5,73	20ª ↓4
17ª	Áustria	1.533	13.618	8,88	11ª ↑6
18ª	Singapura	1.528	11.225	7,35	15ª ↑3
19ª	Suécia	1.468	13.395	9,12	9ª ↑10
20ª	Grécia	1.369	9.368	6,84	16ª ↑4
21ª	**Brasil**	**1.355**	**8.534**	**6,30**	**18ª ↑3**
22ª	Bélgica	1.348	13.954	10,35	4ª ↑18
23ª	Rússia	1.162	4.155	3,58	25ª ↓2
24ª	Escócia	1.055	9.497	8,78	10ª ↑14
25ª	Polônia	1.023	6.349	6,21	19ª ↑6

Os 1.355 artigos publicados no período de 1987 a 2011 de autoria de brasileiros tiveram a participação de autores de 46 outros países. A distribuição geográfica desses coautores de acordo com o seu país de afiliação na WoS é apresentada na Figura 6.38. O país que o Brasil mais colaborou foi o “Estados Unidos”, com 187 artigos nesse período.

Figura 6.38 - Distribuição geográfica dos coautores que publicaram junto com autores brasileiros na categoria CSSE de acordo com o país de sua afiliação na WoS.

Na Figura 6.39 é apresentada as 25 palavras-chave mais utilizadas nos artigos publicados por autores brasileiros na categoria CSSE. “*Grid Computing*” foi a palavra-chave mais utilizada e foi utilizada em 23 desses artigos. Essa palavra-chave foi utilizada de 2004 a 2010. “*Algorithms*” (14) e “*Aspect-Oriented Programming*” (14) também foram citadas por um número significativo de artigos. Foi observado que apenas 950 (70,11%) dos 1.355 artigos publicados por autores brasileiros na categoria CSSE informaram pelo menos uma palavra-chave. Nesses artigos um total de 3.192 palavras-chave distintas foram informadas.

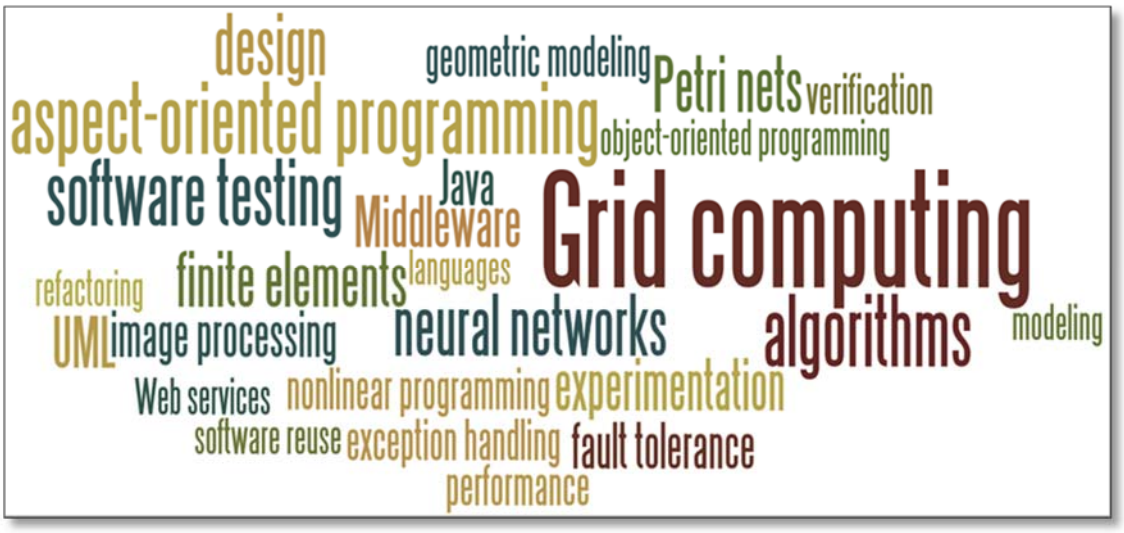


Figura 6.39 - Palavras-chave mais utilizadas nos artigos publicados por autores brasileiros na categoria CSSE.

De 1987 a 2011, um total de 7.425 artigos distintos (sem autocitações) citou algum artigo (1.019 de um total de 1.355) publicado por autores brasileiros na categoria CSSE. Os autores desses artigos estão distribuídos em 97 países diferentes. Na Figura 6.40 são apresentados os 25 países dos autores que citaram mais artigos publicados por autores brasileiros na categoria CSSE. “Brasil” (1.394), “Estados Unidos” (1.373) e “China” (1.037) foram os países com mais autores nessa condição. No caso do Brasil, não foram consideradas as autocitações. Do total de 97 países, 16 (16,49%) países citaram apenas um artigo publicado por autores brasileiros na categoria CSSE.

Figura 6.40 - Distribuição geográfica dos autores que citaram algum artigo publicado por autores brasileiros na categoria CSSE no período de 1987 a 2011 de acordo com o país de sua afiliação na WoS.

Uma das contribuições deste estudo foi o estabelecimento dos requisitos para definir a comunidade de pesquisadores para ser incluída neste conjunto representativo de estudo da área de ES. Essa definição de comunidade pode ser estendida e utilizada em outras áreas do conhecimento com simples ajustes. Agências do governo como a CAPES e o CNPq poderiam mapear o estado atual do conhecimento e das competências de qualquer região do Brasil para obter informações relevantes para dar suporte à definição de políticas públicas.

6.8. Considerações finais

Os estudos de caso apresentados neste Capítulo ilustram algumas análises que podem ser feitas e também mostram a evolução deste trabalho ao longo de sua execução. É possível perceber que as análises se tornaram mais complexas, envolvendo um número maior de bases de dados e mais tópicos sendo considerados. Percebe-se também que se trata de uma evolução natural, pois algumas análises exigem outras bases de dados. Inicialmente,

conforme já mencionado, a ideia era utilizar apenas dados da PL. Porém, isso limitaria o trabalho a estudos envolvendo apenas pesquisadores brasileiros e cadastrados na PL. O fato de considerar bases de dados internacionais permite que análises mais abrangentes sejam realizadas e a comparação da produção científica brasileira com a de outros países. Além disso, essas bases são reconhecidas mundialmente e isso contribui para que este trabalho possa ter uma maior visibilidade. No próximo Capítulo são apresentadas as conclusões deste trabalho e sugestões para trabalhos futuros.

7 CONCLUSÕES

Com este trabalho procurou-se preencher uma dificuldade encontrada por diversos pesquisadores interessados em realizar análise de dados científicos de grupos de pesquisadores com informações disponíveis em bases de dados científicas cuja coleta e extração era muito complicada e trabalhosa. O fato das informações estarem públicas não garante que sua utilização ocorra de forma simples.

É conveniente ter ferramentas computacionais que permitam automatizar e agilizar o processo de obtenção das informações necessárias. Como se sabe, as bases de dados científicas estão em constante atualização e, se o tempo de obtenção das informações for muito demorado, o estudo poderá considerar informações desatualizadas.

No desenvolvimento deste trabalho foram encontrados vários desafios. No caso do currículo Lattes, ele não está disponível em um formato estruturado. Recentemente, o CNPq disponibilizou o currículo Lattes em XML mas restringiu o acesso por meio de um *captcha* que evita que buscadores automáticos obtenham os currículos dos pesquisadores, o que inviabiliza análises de grandes grupos de pesquisadores.

Um dos problemas em extrair dados de um único currículo Lattes disponível na Web é a falta de padronização dos dados registrados. Muitos currículos são parcialmente preenchidos e muitos pesquisadores não atualizam seus currículos periodicamente. O conteúdo das páginas Web dos currículos Lattes não é estruturado e uma das maiores dificuldades é derivada das inconsistências que podem afetar a identificação de relacionamentos. É importante ressaltar que muitos desses problemas somente foram identificados devido ao fato desses dados poderem ser armazenados em um banco de dados. Isso permite rapidamente que diversas consultas podem ser realizadas e assim poder identificar esses problemas.

As bases como a Scopus e a WoS também têm problemas e também impõem restrições quanto ao acesso das informações. A Scopus limita o resultado da consulta em 2.000 registros e a WoS em 100.000 registros.

Todos esses desafios foram superados. Do melhor do nosso conhecimento, não foram encontrados outros trabalhos que tratam desses problemas. Com as ferramentas computacionais desenvolvidas é possível realizar estudos mais abrangentes envolvendo grandes quantidades de dados e que eram impraticáveis.

Uma das contribuições deste trabalho então foi o conjunto de ferramentas computacionais desenvolvidas que permitem extrair automaticamente informações de bases públicas de dados científicas. Foram desenvolvidas duas linguagens de domínio específico que permitem extrair informações da PL e da base Scopus.

O processo de aprendizagem de uma linguagem de programação não é uma tarefa simples. A cada dia que passa mais bibliotecas são adicionadas as linguagens de programação com o objetivo de atender as necessidades crescentes. Numa tentativa de recuperar a usabilidade das linguagens de programação tem sido disseminada a ideia da criação de linguagens de domínio específico. Dessa forma, esperamos que as linguagens desenvolvidas contribuam para que um número maior de usuários possam utilizá-las.

A Extração de Informação na PL de forma automática pode ajudar grupos e o próprio Governo a levantar informações importantes de determinadas áreas e com isso, tendo uma visão ampla, e conhecendo melhor o cenário real atualizado, estabelecer planos estratégicos melhor fundamentados. Estudos envolvendo outras bases de dados científicas também podem contribuir para que políticas públicas sejam melhor definidas. Além disso, esses estudos podem contextualizar a produção científica brasileira no cenário internacional.

As ferramentas desenvolvidas neste trabalho possibilitam também que muitos outros trabalhos futuros possam ser realizados. Além dos exemplos de estudos

apresentados neste trabalho, diversos outros estudos podem ser realizados utilizando as informações de bases de dados científicas. Um dos estudos inovadores realizados, por exemplo, fez uso de uma técnica utilizada em detecção de fraudes. Com isso foi possível identificar inconsistências em bases como a Scopus e WoS. Em outro estudo realizado foi definida uma metodologia para identificar pesquisadores que realmente atuam em uma determinada área do conhecimento, que não é trivial. Essa metodologia pode ser aplicada para qualquer área do conhecimento.

Outra contribuição deste trabalho foi o índice de colaboração proposto, que permite medir a colaboração entre os autores de um artigo. Espera-se que esse índice (e/ou suas variações) possa efetivamente contribuir para entendimento da colaboração entre os pesquisadores.

Uma limitação deste trabalho é que algumas ferramentas desenvolvidas utilizam dados de bases que apenas podem ser acessadas de instituições liberadas pelo Portal de Periódicos da CAPES. Outra limitação é que este trabalho contempla apenas dados de algumas bases. Isso impede que parte da produção científica mundial seja considerada, embora as bases de dados científicas consideradas neste trabalho foram escolhidas principalmente devido a sua abrangência e alcance em nível nacional e mundial.

Pretende-se que as ferramentas desenvolvidas sejam disponibilizadas na Web, permitindo que sejam utilizadas por qualquer pesquisador com conhecimentos básicos em banco de dados e possibilitando que novos estudos sejam realizados, contribuindo para um maior entendimento da produção científica de pesquisadores tanto em nível nacional bem como em nível internacional.

As ferramentas computacionais desenvolvidas, principalmente as linguagens, poderiam ser estendidas adicionando novos recursos e funcionalidades. Além disso, essas ferramentas poderiam ser acessadas através de serviços Web contornando o problema de acesso a algumas bases de dados.

REFERÊNCIAS BIBLIOGRÁFICAS

ABRIZAH, A.; ZAINAB, A. N.; KIRAN, K.; RAJ, R. G. LIS journals scientific impact and subject categorization: a comparison between Web of Science and Scopus. **Scientometrics**, v. 94, n. 2, p. 721-740, 2013.

AGUILLO, I. F. Is Google Scholar useful for bibliometrics? A webometric analysis. **Scientometrics**, v. 91, n. 2, p. 343-351, 2012.

AJIFERUKE, I.; BURELL, Q.; TAGUE, J. Collaborative coefficient: a single measure of the degree of collaboration in research. **Scientometrics**, v. 14, n. 5-6, p. 421-433, 1988.

ALMEIDA, A. M. de. **Proposição de indicadores para avaliação técnica de projetos de Data Warehouse:** um estudo de caso no Data Warehouse da Plataforma Lattes. 83 p. Dissertação (Mestrado em Engenharia de Produção) - Universidade Federal de Santa Catarina (UFSC), Florianópolis, 2006.

ALMEIDA, E. C. E. **A evolução da produção científica nacional, os artigos de revisão e o papel do Portal de Periódicos da CAPES**. 139 p. Tese (Doutorado em Educação em Ciências) - Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, 2013.

ALMEIDA, E. C. E.; GUIMARÃES, J. A. Brazil's growing production of scientific articles—how are we doing with review articles and other qualitative indicators? **Scientometrics**, v. 97, n. 2, p. 287-315, 2013.

ALMEIDA, E. C. E.; GUIMARÃES, J. A.; ALVES, I. T. G. Dez anos do Portal de Periódicos da CAPES: histórico, evolução e utilização. **Revista Brasileira de Pós-Graduação (RBPG)**, v. 7, n. 13, p. 218-246, 2010.

ALONSO, S.; CABRERIZO, F. J.; HERRERA-VIEDMA, E.; HERRERA, F. h-index: a review focused in its variants, computation and standardization for different scientific fields. **Journal of Informetrics**, v. 3, n. 4, p. 273-289, 2009.

ARNOLD, D. N. Integrity under attack: the state of scholarly publishing. **News Journal of the Society for Industrial and Applied Mathematics (SIAM News)**, v. 42, n. 10, p. 1-3, 2009.

ARRUDA, D.; BEZERRA, F.; NERIS, V. A.; TORO, P. R. de.; WAINER, J. Brazilian computer science research: gender and regional distributions. **Scientometrics**, v. 79, n. 3, p. 651-665, 2009.

BALANCIERI, R. **Análise de redes de pesquisa em uma plataforma de gestão em ciência e tecnologia:** uma aplicação à Plataforma Lattes. 117 p. Dissertação (Mestrado em Engenharia de Produção) - Universidade Federal de Santa Catarina (UFSC), Florianópolis, 2004.

BARATA, R. B.; GOLDBAUM, M. Perfil dos pesquisadores com bolsa de produtividade em pesquisa do CNPq da área de Saúde Coletiva. **Cadernos de Saúde Pública**, v. 19, n. 6, p. 1863-1876, 2003.

BATISTA, P. D. Qual seu índice H? **Ciência Hoje**, v. 46, n. 273, p. 28-33, 2010.

BEBER, B.; SCACCO, A. What the numbers say: a digit-based test for election fraud. **Political Analysis**, v. 20, n. 2, p. 211-234, 2012.

BENFORD, F. The law of anomalous numbers. **Proceedings of the American Philosophical Society**, v. 78, n. 4, p. 551-572, 1938.

BITRAN, G. R.; NOVAES, A. G. Linear Programming with a fractional objective function. **Operations Research**, v. 21, n. 1, p. 22-29, 1973.

BORGES, T.; RIBEIRO-JUNIOR, L. C.; LOH, S.; LICHTNOW, D.; KICKHÖFEL, R. B.; GOUVEA, C.; SALDAÑA, R. Identificação automática de expertise analisando currículos no formato Lattes. In: SIMPÓSIO BRASILEIRO DE SISTEMAS DE INFORMAÇÃO (SBSI), 1., 2004, Porto Alegre, RS. **Anais…** Porto Alegre: PUC, 2004. p. 127-134.

BORNMANN, L.; MUTZ, R.; DANIEL, H. Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from Biomedicine. **Journal of the American Society for Information Science and Technology**, v. 59, n. 5, p. 830-837, 2008.

BOVO, A. B. **Um método de tradução de fontes de informação em um formato padrão que viabilize a extração de conhecimento por meio de link analysis e teoria dos grafos**. 102 p. Dissertação (Mestrado em Engenharia de

Produção) - Universidade Federal de Santa Catarina (UFSC), Florianópolis, 2004.
CAMPANARIO, J. M.; COSLADO, M. A. Benford's Law and citations, articles and impact factors of scientific journals. **Scientometrics**, v. 88, n. 2, p. 421-432, 2011.

COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR (CAPES). **Cursos recomendados e reconhecidos**. 2013. Disponível em: <http://www.capes.gov.br/cursos-recomendados>. Acesso em: 08 out. 2013.

CARDOSO, O. N. P. **Gestão do conhecimento usando Data Mining: estudo de caso na UFLA**. 135 p. Dissertação (Mestrado em Administração) - Universidade Federal de Lavras (UFLA), Lavras, 2005.

CARDOSO, O. N. P.; MACHADO, R. T. M. Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras. **Revista de Administração Pública**, v. 42, n. 3, p. 495-528, 2008.

CASTAÑO, A. C. **Populando ontologias através de informações em HTML - o caso do currículo Lattes**. 100 p. Dissertação (Mestrado em Ciência da Computação) - Universidade de São Paulo (USP), São Paulo, 2008.

CAVALCANTE, R. A.; BARBOSA, D. R.; BONAN, P. R. F.; PIRES, M. B. de O.; MARTELLI JÚNIOR, H. Perfil dos pesquisadores da área de Odontologia no Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). **Revista Brasileira de Epidemiologia**, v. 11, n. 1, p. 106-113, 2008.

CAVALCANTI, A. L.; PEREIRA, D. S. de A. Perfil do bolsista de produtividade em pesquisa do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) na área de Odontologia. **Revista Brasileira de Pós-Graduação (RBPG)**, v. 5, n. 9, p. 67-88, 2008.

CHANG, C.-H,; HSU, C.-N; LUI, S.-C. Automatic information extraction from semi-structured web pages by pattern discovery. **Decision Support Systems**, v. 35, n. 1, p. 129-147, 2003.

CIVIDANES, F. de S. **CollectLattes:** sistema para extração de conhecimento sobre a Plataforma Lattes. 167 p. Dissertação (Mestrado em Engenharia

Eletrônica e Computação) - Instituto Tecnológico de Aeronáutica (ITA), São José dos Campos, 2010.

CLIPPE, P.; AUSLOOS, M. Benford's Law and Theil transform of financial data. **Physica A: Statistical Mechanics and its Applications**, v. 391, n. 24, p. 6556-6567, 2012.

CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO (CNPq). **Plataforma Lattes**. 2013. Disponível em: <http://lattes.cnpq.br/>. Acesso em: 02 out. 2013.

COSTAS, R.; BORDONS, M. The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. **Journal of Informetrics**, v. 1, n. 3, p. 193-203, 2007.

COURY, H. J. C. G.; VILELLA, I. Perfil do pesquisador fisioterapeuta brasileiro. **Revista Brasileira de Fisioterapia**, v. 13, n. 4, p. 356-363, 2009.

DEURSEN, A. van; KLINT, P. Little languages: Little maintenance? **Journal of Software Maintenance**, v. 10, n. 2, p. 75-92, 1998.

DEURSEN, A. van; KLINT, P.; VISSER, J. Domain-specific languages: an annotated bibliography. **ACM SIGPLAN Notices**, v. 35, n. 6, p. 26-36, 2000.

THE SAN FRANCISCO DECLARATION ON RESEARCH ASSESSMENT (DORA). **San Francisco Declaration on Research Assessment (DORA)**. 2013. Disponível em: <http://am.ascb.org/dora/>. Acesso em: 12 dez. 2013.

EGGHE, L. Five years "Journal of Informetrics". **Journal of Informetrics**, v. 6, n. 3, p. 422-426, 2012.

EGGHE, L. Theory and practise of the g-index. **Scientometrics**, v. 69, n. 1, p. 131-152, 2006.

ELSEVIER. **Elsevier is a world-leading provider of scientific, technical and medical information products and services | Elsevier**. 2013. Disponível em: < http://www.elsevier.com/>. Acesso em: 08 out. 2013.

FOWLER, M. A pedagogical framework for domain-specific languages. **IEEE Software**, v. 26, n. 4, p. 13-14, 2009.

FREIRE, R. S.; OLIVEIRA, E. A.; SILVEIRA, M. F.; MARTELLI, D. R. B.; OLIVEIRA, M. C. L.; MARTELLI JÚNIOR, H. Perfil dos pesquisadores na área de Fisioterapia e Terapia Ocupacional no Conselho Nacional de Desenvolvimento Cientifico e Tecnológico. **Revista Brasileira de Pós-Graduação (RBPG)**, v. 10, n. 19, p. 11-24, 2013.

FREEMAN, S.; PRYCE, N. Evolving an embedded domain-specific language in java. In: **OOPSLA '06:** Companion to the 21st ACM SIGPLAN symposium on Object-oriented programming systems, languages, and applications. New York, NY, USA: ACM, 2006. p. 855-865.

FREITAS, C. M. D. S.; NEDEL, L. P.; GALANTE, R. LAMB, L. C.; SPRITZER, A. S.; FUJII, S.; OLIVEIRA, J. P. M. de; ARAÚJO, R. M.; MORO, M. M. Extração de conhecimento e análise visual de redes sociais. In: SEMINÁRIO INTEGRADO DE SOFTWARE E HARDWARE (SEMISH), 28., 2008, Belém do Pará, PA. **Anais...** Porto Alegre: SBC, 2008. p. 106-120.

GARFIELD, E. Citation indexes for Science: a new dimension in documentation through association of ideas. **Science**, v. 122, n. 3159, p. 108-111, 1955.

GLÄNZEL, W.; LANGE, C. de. A distributional approach to multinationality measures of international scientific collaboration. **Scientometrics**, v. 54, n. 1, p. 75-89, 2002.

GONZÁLEZ-PEREIRA, B.; GUERRERO-BOTE, V. P.; MOYA-ANEGÓN, F. A new approach to the metric of journals' scientific prestige: the SJR indicator. **Journal of Informetrics**, v. 4, n. 3, p. 379-391, 2010.

HEIN, J.; ZOBRIST, R.; KONRAD, C.; SHUEPFER, G. Scientific fraud in 20 falsified anesthesia papers. **Der Anaesthesist**, v. 61, n. 6, p. 543-549, 2012.

HIRSCH, J. E. An index to quantify an individual's scientific research output. **Proceedings of the National Academy of Sciences of the United States of America (PNAS)**, v. 102, n. 46, p. 16569-16572, 2005.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Censo de 2010**. 2013. Disponível em: <http://www.ibge.gov.br/estadosat/>. Acesso em: 08 out. 2013.

JCR. **Journal Citation Reports - Thomson Reuters**. 2013. Disponível em: <http://wokinfo.com/products_tools/analytical/jcr/>. Acesso em: 08 out. 2013.

KALIL, F. **Uma ferramenta de suporte à análise do comportamento científico dos pesquisadores**. 82 p. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) - Universidade de Passo Fundo (UPF), Passo Fundo, 2008.

KARAHALIOS, K. G.; VIÉGAS, F. B. Social visualization: exploring text, audio, and video interaction. In: CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (CHI '06), 2006, Montreal, Canada. **Extended Abstracts…** New York, NY, USA: ACM, 2006. p. 1667-1670.

KARAMOURZOV, R. The development trends in science in the CIS countries on the basis of some scientometrics indicator. **Scientometrics**, v. 91, n. 1, p. 1-14, 2012.

KATZ, J. S.; MARTIN, B. R. What is research collaboration? **Research Policy**, v. 26, n. 1, p. 1-18, 1997.

KAYED, M.; SHAALAN, K. F. A survey of web information extraction systems. **IEEE Transactions on Knowledge and Data Engineering**, v. 18, n. 10, p. 1411-1428, 2006.

KOOPMAN, B. O. New mathematical methods in Operations Research. **Journal of the Operations Research Society of America**, v. 1., n. 1, p. 3-9, 1952.

KOSAR, T.; LÓPEZ, P. E. M.; BARRIENTOS, P. A.; MERNIK, M. A preliminary study on various implementation approaches of domain-specific language. **Information and Software Technology**, v. 50, n. 5, p. 390-405, 2008.

KOSMULSKI, M. A new Hirsch-type index saves time and works equally well as the original h-index. **ISSI Newsletter**, v. 2, n. 3, p. 4-6, 2006.

KOSMULSKI, M. Hirsch-type index of international recognition. **Journal of Informetrics**, v. 4, n. 3, p. 351-357, 2010.

LANE, J. Let´s make science metrics more scientific. **Nature**, v. 464, n. 7288, p. 488-489, 2010.

LAWANI, S. M. Some bibliometric correlates of quality in scientific research. **Scientometrics**, v. 9, n. 1-2, p. 13-25, 1986.

LEITE, B. D.; OLIVEIRA, E. A.; QUEIROZ, I. N.; MARTELLI, D. R.; OLIVEIRA, M. C.; MARTELLI JÚNIOR, H. Profile of the researchers with productivity grants in the Brazilian National Research Council (CNPq) of the Physical Education Area, **Motricidade**, v. 8., n. 3, p. 90-98, 2012.

LIM, E.-P; MAUREEN; M.; IBRAHIM, N. L.; SUN, A.; DATTA, A.; CHANG, K. SSnetViz: a visualization engine for heterogeneous semantic social networks. In: INTERNATIONAL CONFERENCE ON ELECTRONIC COMMERCE, 11., 2009, New Iork. **Proceedings...** New York, NY, USA: ACM, 2009. p. 213-221.

MARINHO, I. **A comunicação científica e o modelo de comunicação organizacional: análise quantitativa de produtividade dos programas de pós-graduação em Ciência da Informação por meio do currículo Lattes**. 107 p. Dissertação (Mestrado em Ciência da Informação) - Universidade de Brasília (UnB), Brasília, 2007.

MARTELLI JÚNIOR, H.; MARTELLI, D. R. B.; QUIRINO, I. G.; OLIVEIRA, M. C. L.; LIMA, L. S.; OLIVEIRA, E. A. de. Pesquisadores do CNPQ na área de Medicina: comparação das áreas de atuação. **Revista da Associação Médica Brasileira**, v. 56, n. 1, p. 478-483, 2010.

MELO, P. L. da C. e. **Produtividade, internacionalização e visibilidade da comunidade científica brasileira na virada do milênio**. 177 p. Tese (Doutorado em Química Biológica) - Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, 2011.

MENA-CHALCO, J. P.; CESAR-JUNIOR, R. M. scriptLattes: an open-source knowledge extraction system from the Lattes platform. **Journal of the Brazilian Computer Society**, v. 15, n. 4, p. 31-39, 2009.

MENDES, P. H. C.; MARTELLI, D. R. B.; SOUZA, W. P. de; FILHO, S. Q.; MARTELLI JÚNIOR, H. Perfil dos pesquisadores bolsistas de produtividade científica em Medicina no CNPQ, Brasil. **Revista Brasileira de Educação Médica**, v. 34, n. 4, p. 535-541, 2010.

MICHELS, C.; SCHMOCH, U. The growth of science and database coverage. **Scientometrics**, v. 93, n. 3, p. 831-846, 2012.

MILGRAM, S. The small word problem. **Psychology Today**, v. 1, n. 1, p. 61-67, 1967.

MIR, T. A. The law of the leading digits and the world religions. **Physica A: Statistical Mechanics and its Applications**, v. 391, n. 3, p. 792-798, 2012.

MOED, H. F. Measuring contextual citation impact of scientific journals. **Journal of Informetrics**, v. 4, n. 3, p. 265-277, 2010.

MOHAMMADHASSANZADEH, H.; SAMADIKUCKAKSARAEI, A.; SAEMI, N.; MOHAMMAD S. Two new scientometric indices for measurement of collaboration activities of departments and their researchers in academic institutions. **Malaysian Journal of Library & Information Science**, v. 16, n. 3, p. 1-7, 2011.

MOREIRA, L. M. **Formação de competências em ciência e tecnologias espaciais:** uma análise da trajetória da pós-graduação no Instituto Nacional de Pesquisas Espaciais. 209 p. Tese (Doutorado em Política Científica e Tecnológica) - Universidade Estadual de Campinas (UNICAMP), Campinas, 2009.

MOREIRA, L. M.; VELHO, L. Pós-Graduação no INPE: a aliança pesquisa-desenvolvimento e ensino. **Cadernos de Pesquisa**, v. 39, n. 136, p. 243-268, 2009.

MOREIRA, L. M.; VELHO, L. Pós-Graduação do Instituto Nacional de Pesquisas Espaciais numa perspectiva de gênero. **Cadernos Pagu**, n. 35, p. 279-308, 2010.

MOREIRA, L. M.; VELHO, L. Trajetória de egressos da Pós-Graduação do Instituto Nacional de Pesquisas Espaciais: uma ferramenta para avaliação. **Avaliação**, v. 17, n. 1, p. 257-288, 2012.

MORENO, J. L. **Who shall survive?:** a new approach to the problem of human interrelations. Washington, D.C.: Nervous and Mental Disease Publishing Co., 1934. 440 p.

NANNO, T.; SAITO, S.; OKUMURA, M. Structuring web pages based on repetition of elements. In: In: INTERNATIONAL WORKSHOP ON WEB DOCUMENT ANALYSIS (WDA), 2., 2003, Edinburgh - UK. **Proceedings...** Edinburgh, 2003. p. 7-10.

NASCIMENTO-JÚNIOR, E. **Representação visual de rede social de pesquisa**. 50 p.Trabalho de Graduação (Divisão de Ciência da Computação) - Instituto Tecnológico de Aeronáutica (ITA), São José dos Campos, 2008.

NEWCOMB, S. Note on the frequency of use of the different digits in natural numbers. **American Journal of Mathematics**, v. 4, n. 1, p. 39-40, 1881.

NEWMAN, M.; BARABÁSI, A-L.; WATTS, D. J. **The Structure and Dynamics of Networks**. Princeton: Princeton University Press, 2006. 582 p.

NIGRINI, M. **Benford's Law:** applications for forensic accounting, auditing, and fraud detection. Wiley Corporate F&A, 2012. 330 p.

NORUZI, A. Google Scholar: the new generation of citation indexes. **Libri: International Journal of Libraries and Information Services**, v. 55, n. 4, p. 170-180, 2005.

OLIVEIRA, E. A.; COLOSIMO, E. A.; MARTELLI, D. R.; QUIRINO, I. G.; OLIVEIRA, M. C. L.; LIMA, L. S.; SILVA, A. C. S. e; MARTELLI JÚNIOR, H. Comparison of Brazilian researchers in Clinical Medicine: are criteria for ranking well-adjusted? **Scientometrics**, v. 90, n. 2, p. 429-443, 2012.

OLIVEIRA, E. A.; PÉCOITS-FILHO, R.; QUIRINO, I. G.; OLIVEIRA, M. C.; MARTELLI, D. R.; LIMA, L. S.; MARTELLI JÚNIOR, H. Perfil e produção científica dos pesquisadores do CNPq nas áreas de Nefrologia e Urologia. **Jornal Brasileiro de Nefrologia**, v. 33, n. 1, p. 31-37, 2011.

OLIVEIRA, E. A.; RIBEIRO, A. L. P.; QUIRINO, I. G.; OLIVEIRA, M. C. L.; MARTELLI, D. R.; LIMA, L. S.; COLOSIMO, E. A.; LOPES, T. J.; SILVA, A. C. S. e; MARTELLI JÚNIOR, H. Profile and scientific production of CNPq

researchers in Cardiology. **Arquivos Brasileiros de Cardiologia**, v. 97, n. 3, p. 186-193, 2011.

OLIVEIRA, M. C. de; BERNUSSOU, J.; GEROMEL, J. C. A new discrete-time robust stability condition. **Systems & Control Letters**, v. 37, n. 4, p. 261-265, 1999.

OLIVEIRA, M. C. L.; MARTELLI, D. R. B.; PINHEIRO, S. V.; MIRANDA, D. M.; QUIRINO, I. G.; LEITE, B. G. L.; COLOSIMO, E. A.; SILVA, A. C. S. e; MARTELLI-JÚNIOR, H.; OLIVEIRA, E. A. Profile and scientific production of Brazilian National Council of Technological and Scientific Development researchers in Pediatrics. **Revista Paulista de Pediatria**, v. 31, n. 3, p. 278-284, 2013.

PACHECO, R. C. S.; FORCELLINI, F. A.; KERN, V. M.; GONÇALVES, A. L.; IGARASHI, W. Uma análise da pesquisa em Engenharia e Ciências Mecânicas no Brasil a partir dos dados da Plataforma Lattes. **Associação Brasileira de Engenharia e Ciências Mecânicas (ABCM)**, v. 12, n. 1, p. 18-24, 2007.

PACHECO, R. C. S.; KERN, V. M. Uma ontologia comum para a integração de bases de informações e conhecimento sobre ciência e tecnologia. **Ciência da Informação**, v. 30, n. 3, p. 56-63, 2001.

PAULA, M. V. de. **Explorando o potencial da Plataforma Lattes como fonte de conhecimento organizacional em ciência e tecnologia**. 148 p. Dissertação (Mestrado em Gestão do Conhecimento e da Tecnologia da Informação) - Universidade Católica de Brasília (UCB), Brasília, 2004.

POPOFF, D. A. V.; FERREIRA, R. C.; MARTELLI, D. R. B.; OLIVEIRA, E. A. de; VIEIRA JÚNIOR, J. R.; MARTELLI JÚNIOR, H. Profile and scientific production of Brazilian researchers in Dental Materials. **Brazilian Journal of Oral Sciences**, v. 11, n. 1, p. 56-61, 2012.

QUALIS. **Qualis Periódicos (CAPES)**. 2013. Disponível em: <http://www.capes.gov.br/avaliacao/qualis>. Acesso em: 08 out. 2013.

ROMANO-SILVA, M. A.; CORREA, H.; OLIVEIRA, M. C. L.; QUIRINO, I. G.; COLOSIMO, E. A.; MARTELLI, D. R.; DUARTE, M. G.; LIMA, L. S.; SIMÕES E SILVA, A. C.; MARTELLI JÚNIOR, H.; OLIVEIRA, E. A. Perfil e análise da

produção científica dos pesquisadores brasileiros em Neurociência Clínica, **Revista de Psiquiatria Clínica**, v. 40, n. 2, p. 53-58, 2013.

ROSA, S. P. **O campo de conhecimento da Educação Física: uma abordagem cientométrica**. 197 p. Tese (Doutorado em Química Biológica) – Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, 2010.

SANTOS, N. C. F.; CÂNDIDO, L. F. O.; KUPPENS, C. L. Produtividade em pesquisa do CNPq: análise do perfil dos pesquisadores da Química, **Química Nova**, v. 33, n. 2, p. 489-495, 2010.

SANTOS, S. M. C.; LIMA, L. S.; MARTELLI, D. R. B.; MARTELLI JÚNIOR, H. Perfil dos pesquisadores da Saúde Coletiva no Conselho Nacional de Desenvolvimento Científico e Tecnológico. **Physis: Revista de Saúde Coletiva**, v. 19, n. 3, p. 761-775, 2009.

SCARPELLI, A. C.; SARDENBERG, F.; GOURSAND, D.; PAIVA, S. M.; PORDEUS, I. A. Academic Trajectories of Dental researchers receiving CNPq´s productivity grants. **Brazilian Dental Journal**, v. 19, n. 3, p. 252-256, 2008.

SCIENTI. **ScienTI Network -** international network of information and knowledge sources for science, technology and innovation management. 2013. Disponível em: <http://www.scienti.net/php/index.php?lang=en>. Acesso em: 15 nov. 2013.

SCHUBERT, A.; GLÄNZEL, W. A systematic analysis of Hirsch-type indices for journals. **Journal of Informetrics**, v. 1, n. 3, p. 179-184, 2007.

SCIMAGO. **SCImago Journal & Country Rank**. 2013. Disponível em: <http://www.scimagojr.com/>. Acesso em: 08 out. 2013.

SCOPUS. **Scopus**. 2013. Disponível em: <http://www.scopus.com/>. Acesso em: 08 out. 2013.

SEGLEN, P. O. Why the impact factor of journals should not be used for evaluating research. **British Medical Journal**, v. 314, n. 7079, p. 497-502, 1997.

SILVA, E. F. A.; BARROS, F. A.; PRUDÊNCIO, R. B. C. Uma abordagem de aprendizagem híbrida para extração de informação em textos semi-estruturados. In: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL (ENIA), 5., 2005, São Leopoldo - RS. **Anais...** 2005. p. 504-513.

SILVA, F. M. **Organização da informação em sistemas eletrônicos abertos de informação científica e tecnológica:** análise da Plataforma Lattes. 163 p. Tese (Doutorado em Cultura e Informação) - Universidade de São Paulo (USP), São Paulo, 2007.

SILVA, L. L. Estudo do perfil científico dos pesquisadores com bolsa de produtividade do CNPq que atuam no ensino de Ciências e Matemática. **Revista Brasileira de Pesquisa em Educação em Ciências**, v. 11, n. 3, p. 75-99, 2011.

SILVA, S. R. P.; PINHEIRO, J. M. Um framework para criação de linguagens de domínio específico. In: SIMPÓSIO BRASILEIRO DE LINGUAGENS DE PROGRAMAÇÃO (SBLP), 8., 2004, Niterói – RJ. **Anais…** Niterói, 2004.

SINGH, L.; BEARD, M.; GETOOR, L.; BLAKE, M. B. Visual mining of multi-modal social networks at different abstraction levels. In: **11th International Conference Information Visualization (IV)**. Zurich, Switzerland: IEEE Computer Society Press, 2007. p. 672-679.

SOUTO, R. Q.; LACERDA, G. da S.; COSTA, G. M. C.; CAVALCANTI, A. L.; FRANÇA, I. S. X. de; SOUSA, F. S. de. Characterization of the productivity of scholar researchers of CNPq of Nursing: a cross-sectional study. **Online Brazilian Journal of Nursing**, v. 11, n. 2, p. 261-272, 2012.

SPILKI, F. R. Perfil dos bolsistas de produtividade do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) na área de Medicina Veterinária. **Pesquisa Veterinária Brasileira**, v. 33, n. 2, p. 205-213, 2013.

STALLINGS, J.; VANCE, E.; YANG, J.; VANNIER, M. W.; LIANG, J.; PANG, L.; DAI, L. Determining scientific impact using a collaboration index. **Proceedings of the National Academy of Sciences of the United States of America (PNAS)**, v. 110, n. 24, p. 9680-9685, 2013.

STORN, R.; PRICE, K. Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. **Journal of Global Optimization**, v. 11, n. 4, p. 341-359, 1997.

SUBRAMANYAM, K. Bibliometric studies of research collaboration: a review. **Journal of Information Science**, v. 6, n. 1, p. 33-38, 1983.

TAHA, W. Plenary talk III Domain-specific languages. In: INTERNATIONAL CONFERENCE ON COMPUTER ENGINEERING & SYSTEMS (ICCES), 2008, Cairo. **Proceedings…** Cairo: IEEE, 2008. p. xxiii-xxvii.

VADREVU, S.; GELCI, F.; DAVULCU, H. Information extraction from web pages using presentation regularities and domain knowledge. **World Wide Web**, v. 10, n. 2, p. 157-179, 2007.

VANZ, S. A. de S. **As redes de colaboração científica no Brasil (2004-2006)**. 204 p. Tese (Doutorado em Comunicação e Informação) - Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, 2009.

VANZ, S. A. de S.; STUMPF, I. R. C. Colaboração científica: revisão teórico-conceitual. **Perspectivas em Ciência da Informação**, v. 15, n. 2, p. 42-55, 2010.

VASCONCELOS, S. M. R. de. **Ciência no Brasil:** uma abordagem cientométrica e linguística. 206 p. Tese (Doutorado em Química Biológica) - Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, 2008.

VASCONCELOS, S. M. R. de.; SORENSON, M. M.; LETA, J. A new input indicator for the assessment of science & technology research? **Scientometrics**, v. 80, n. 1, 2009.

WASSERMAN, S.; FAUST, K. **Social network analysis:** methods and applications (structural analysis in the social sciences). Cambridge: Cambridge University Press, 1994. 857 p.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world' networks. **Nature**, v. 393, n. 6684, p. 440-442, 1998.

WENDT, G. W.; LISBOA, C. S. de M.; DESOUSA, D. A.; KOLLER, S. H. Perfil dos bolsistas de produtividade em pesquisa do CNPQ em Psicologia. **Psicologia: Ciência e Profissão**, v. 33, n. 3, p. 536-547, 2013.

WOS. Web of Science - Thomson Reuters. 2013. Disponível em: <http://wokinfo.com/wok/products_tools/multidisciplinary/webofscience/>. Acesso em: 12 nov. 2013.

XIAO, L.; WISSMANN, D.; BROWN, M.; JABLONSKI, S. Information extraction from the web: systems and techniques. **Applied Intelligence**, v. 21, n. 2, p. 195-224, 2004.

YU, H.; DAVIS, M.; WILSON, C. S.; COLE, F. T. H. Object-relational data modelling for informetric databases. **Journal of Informetrics**, v. 2, n. 3, p. 240-251, 2008.