Revista Eletrônica de Sistemas de Informação
ISSN 1677-3071

v. 14, n. 2

mai-ago 2015 - Edição Temática sobre Análise de Redes Sociais e Mineração

doi:10.21529/RESI.2015.1402

Sumário

Editorial

BrASNAM

ISSN: 1677-3071

Esta revista é (e sempre foi) eletrônica para ajudar a proteger o meio ambiente, mas, caso deseje imprimir esse artigo, saiba que ele foi editorado com uma fonte mais ecológica, a *Eco Sans*, que gasta menos tinta.

This journal is (and has always been) electronic in order to be more environmentally friendly. Now, it is desktop edited in a single column to be easier to read on the screen. However, if you wish to print this paper, be aware that it uses Eco Sans, a printing font that reduces the amount of required ink.

EXTRAÇÃO, CARACTERIZAÇÃO E ANÁLISES DE DADOS DE CURRÍCULOS LATTES

EXTRACTION, CHARACTERIZATION AND ANALYSIS OF DATA FROM THE LATTES CURRICULA PLATFORM

(artigo submetido em março de 2013)

Luciano A. Digiampietri
Escola de Artes, Ciências e Humanidades - Universidade de São Paulo (USP)
digiampietri@usp.br

Jesús P. Mena-Chalco
Centro de Matemática, Computação e Cognição – Univ. Federal do ABC (UFABC)
jesus.mena@ufabc.edu.br

José J. Pérez-Alcázar
Escola de Artes, Ciências e Humanidades - Universidade de São Paulo (EACH-USP)
jperez@usp.br

Esteban F. Tuesta
Escola de Artes, Ciências e Humanidades - Universidade de São Paulo (EACH-USP)
tuesta@usp.br

Karina V. Delgado
Escola de Artes, Ciências e Humanidades - Universidade de São Paulo (EACH-USP)
kvd@usp.br

Rogério Mugnaini
Escola de Artes, Ciências e Humanidades - Universidade de São Paulo (EACH-USP)
mugnaini@usp.br

Gabriela S. Silva
Escola de Artes, Ciências e Humanidades - Universidade de São Paulo (EACH-USP)
gabriela.scardine.silva@usp.br

Jamison J. S. Lima
Escola de Artes, Ciências e Humanidades - Universidade de São Paulo (EACH-USP)
jamison.lima@usp.br

ABSTRACT

Curricula from the Lattes Platform are a vast source of information for the creation and analysis of researchers' social networks. However, due to the large amount of data, the manual filling-in, and the use of semi-structured data, there are several challenges in the use of Lattes as a source of data. This paper presents a database produced from the mining of more than one million Brazilian Lattes curricula. Moreover, it highlights some descriptive characteristics and relationships among these curricula and among the knowledge areas, directions and challenges to the production and analyzes of social networks generated from these data.

Keywords: Lattes platform; social network; co-authorship network; data mining.

RESUMO

Os currículos da Plataforma Lattes são uma vasta fonte de informação para a criação e análise de redes sociais de pesquisadores. Contudo, devido à quantidade de dados, ao preenchimento manual e ao uso de dados semiestruturados existem diversos desafios para a utilização desta fonte de dados. Este artigo apresenta um banco de dados produzido a partir da mineração de mais de um milhão de Currículos Lattes, destacando algumas características descritivas e relações entre os currículos e entre as grandes áreas de conhecimento, direções e desafios para a produção e análise de redes sociais a partir destes dados.

Palavras-chave: plataforma Lattes; rede social; rede de coautoria; mineração de dados.

1 INTRODUÇÃO

Atualmente, é possível encontrar na Web uma grande quantidade de dados referentes aos mais diversos assuntos. Dentre esses dados estão informações muito relevantes sobre os pesquisadores, como publicações científicas por eles desenvolvidas, informações sobre projetos de pesquisa e mesmo currículos de pesquisadores.

Ao se tratar de dados referentes a pesquisa, a comunidade acadêmica no Brasil apresenta uma característica peculiar: a existência de um cadastro nacional de currículos de pesquisadores, a Plataforma Lattes, que congrega informações sobre publicações, orientações, projetos de pesquisa, entre outras. O currículo Lattes foi lançado e padronizado em agosto de 1999 pelo CNPq, como sendo o formulário de currículo a ser utilizado no âmbito do Ministério da Ciência e Tecnologia e do CNPq[1]. No ano de 2007 ultrapassou a marca de 1 milhão de currículos.

Currículos da Plataforma Lattes são uma vasta fonte de informação para a criação e análise de redes sociais de pesquisadores (BALANCIERI *et al.,* 2005). Contudo, devido à quantidade de dados, ao preenchimento manual e ao uso de dados semiestruturados, existem diversos desafios computacionais para a utilização desta fonte de dados. Por isso, o grande volume de informações disponível na base tem sido pouco usado, servindo, tipicamente, para avaliar (ou verificar) dados de pesquisadores individualmente ou de pequenos grupos de pesquisadores.

Uma busca sobre "análise de currículo(s)" ("*analysis of curric**") na Web of Science recupera 43 referências a artigos científicos, sendo 50% deles na área de educação. Neste conjunto, evidencia-se o interesse em estudos sobre competência profissional e formação acadêmica. Com a Plataforma Lattes, abre-se a oportunidade de se investigar questões de interesse da área de Política Científica, possibilitando-se estudar não apenas a produtividade dos pesquisadores, mas suas relações de colaboração, vinculando assim dois temas normalmente abordados separadamente: Análise de Redes Sociais e Cientometria.

Porém, o acesso à base Lattes ainda é pouco facilitado pelo CNPq, fazendo com que estudos em nível macro, ou utilizem a interface do Censo do Diretório dos Grupos de Pesquisa no Brasil (GUIMARÃES, 2004) (via que não permite muita liberdade para análise dos dados), ou dependam de solicitação dos dados ao CNPq (LEITE *et al.,* 2011; MUGNAINI *et al.,* 2012) (acesso que nem sempre é garantido).

Este artigo visa a apresentar um banco de dados formado por mais de um milhão de currículos minerados da Plataforma Lattes e que foram processados, organizados e analisados para servirem de base para a produção e análise de redes sociais de pesquisa. O presente artigo revisita e estende o trabalho "Minerando e caracterizando dados de currículos Lattes" (DIGIAMPIETRI *et al.,* 2012), de forma a aprofundar a descrição do

[1] lattes.cnpq.br/conteudo/historico.htm

doi:10.21529/RESI.2015.1402001

conjunto de dados. Além disso, uma análise da rede social formada por todos os currículos do banco de dados foi criada e suas principais características são apresentadas.

O restante deste artigo está organizado da seguinte forma. A Seção 2 apresenta os trabalhos correlatos. A Seção 3 descreve os processos de mineração e produção do banco de dados. A caracterização, análise e cuidados relacionados ao banco de dados são apresentados na Seção 4. Por fim, a Seção 5 contém as considerações finais e sugestões de trabalhos futuros.

2 TRABALHOS CORRELATOS

Há mais de 60 anos existem estudos sobre análise de redes sociais, havendo uma vasta literatura sobre o assunto (NEWMAN, 2001). Esta seção apresenta apenas uma visão geral dos trabalhos correlatos com enfoque na análise de redes sociais de pesquisadores e, especialmente, naqueles relacionados ao uso de Currículos Lattes.

Silva e Smit (2009) avaliaram a organização e qualidade da informação científica disponível nos Currículos Lattes. Eles concluíram que há bastante comprometimento no preenchimento dos currículos, por mais que existam diversos pequenos erros de preenchimento. Por analisar uma amostra não muito grande de currículos, eles não calcularam estatísticas sobre a Plataforma Lattes.

Mena-Chalco e César Júnior (2009) desenvolveram e disponibilizaram uma ferramenta chamada *scriptLattes*[2], que recebe uma lista de identificadores de Currículos Lattes e gera diversas páginas HTML, organizando as informações dos currículos (tanto sumarizando informações quanto separando as informações por categoria). Além disso, a ferramenta gera um mapa da distribuição geográfica dos pesquisadores envolvidos na análise e um grafo de coautorias. Recentemente, os autores analisaram a rede de coautorias de mais de um milhão de Currículos Lattes considerando diferentes métricas e de acordo com as áreas de atuação dos pesquisadores (MENA-CHALCO *et al.,* 2014).

Alves *et al*. (2011) desenvolveram um sistema chamado SUCUPIRA, uma ferramenta *online* que, além de possuir funcionalidades específicas para baixar e organizar currículos Lattes, une algumas ferramentas/APIs para visualizar informações de currículos e suas redes de coautoria. O sistema foi testado analisando-se o conjunto de autores de um programa de pós-graduação da UFMG.

Digiampietri e Silva (2011) desenvolveram uma infraestrutura para encontrar currículos de grupos de pesquisadores. Dada uma lista com o nome dos pesquisadores do grupo de interesse, a infraestrutura procura

[2] scriptlattes.sourceforge.net

pelos currículos utilizando as APIs de busca da Google[3] e da Microsoft[4], baixa esses currículos, gera as redes de coautoria e analisa a produção dos pesquisadores cruzando as informações de cada currículo com as informações dos documentos de área da CAPES.

O banco de dados apresentado neste artigo se diferencia dos demais trabalhos por apresentar um conjunto de currículos bastante representativo (ao nosso entender, muito maior do que os analisados nos demais trabalhos). Outra característica relevante é a granularidade da informação dos currículos, pois neste banco as informações são organizadas campo a campo (e não publicação a publicação, por exemplo). Além disso, os dados foram cuidadosamente estruturados e inseridos em um banco de dados relacional visando a facilitar consultas, análises e enriquecimento dos dados. Por se tratar de conjunto de dados representativo, foi possível analisar algumas características e gerar algumas estatísticas sobre os Currículos Lattes, como será apresentado nas próximas seções.

3 MATERIAIS E MÉTODOS

Esta seção descreve a metodologia utilizada para a criação do banco de dados de currículos Lattes. As três principais atividades realizadas são: busca e obtenção dos currículos; processamento inicial dos arquivos HTML; modelagem e população do banco de dados. Cada uma destas atividades será descrita a seguir.

3.1 BUSCA E OBTENÇÃO DOS CURRÍCULOS

Os currículos da Plataforma Lattes foram obtidos pela Internet, utilizando-se o comando *wget* para baixar cada um dos currículos. Para isto, foi necessário encontrar o identificador numérico de cada currículo, pois este é necessário para compor a URL (*Uniform Resource Locator*) completa do currículo, a qual é parâmetro da ferramenta *wget*.

Para encontrar os identificadores dos currículos, duas estratégias foram usadas. Na primeira, foram feitas consultas na interface de busca provida pela Plataforma Lattes. Mais especificamente, foram feitas 80 consultas, cada uma utilizando como palavras-chave as (sub)áreas de conhecimento da própria plataforma. Foi desenvolvido um *parser* para encontrar os identificadores dos currículos de cada uma das respostas às consultas realizadas. Esta estratégia permitiu a identificação de centenas de milhares de currículos. Estes currículos foram baixados e examinados automaticamente em busca de identificadores de outros currículos (por exemplo, de coautores, orientadores ou orientandos). Este exame constituiu a segunda estratégia de busca. Sempre que um novo currículo era identificado, este era baixado e examinado da mesma forma que os demais.

[3] www.google.com

[4] www.bing.com

doi:10.21529/RESI.2015.1402001

Combinando-se estas duas estratégias, 1.236.548 currículos foram baixados totalizando pouco mais de 16 GB em arquivos HTML.

É importante destacar que as estratégias adotadas não visavam à obtenção de toda a base de currículos da Plataforma Lattes, mas sim um conjunto significativo de currículos para serem processados e servirem de base para a criação e análise de redes sociais de pesquisadores. Levando-se em conta que o CNPq anunciou que no ano de 2007 a base atingiu um milhão de currículos, considera-se que foi obtida uma quantidade representativa de currículos.

Todos os currículos foram baixados em maio de 2011. A partir desta data foram feitos apenas processamentos com base nas informações dos currículos já baixados.

3.2 PROCESSAMENTO INICIAL DOS CURRÍCULOS

Cada um dos currículos baixados foi submetido a duas etapas iniciais de processamento. Na primeira, foram retirados todos os caracteres especiais, espaços em brancos excedentes e fins de linha, de forma a facilitar a identificação de informações executada na etapa subsequente de processamento.

Na segunda etapa de processamento, foram desenvolvidas expressões regulares para dividir cada currículo em suas seções principais (Dados Gerais; Linhas de pesquisa; Projetos; Áreas; Produção em C, T & A; Bancas; Eventos; e Orientações). As informações de cada uma destas seções foram processadas por expressões regulares específicas para a identificação dos itens e campos de interesse. Por exemplo, informações referentes à Produção em C, T & A foram subdivididas em Produção bibliográfica; Produção técnica; Produção artística/cultural; e demais trabalhos. Informações sobre a Produção bibliográfica foram subdivididas em 7 categorias: artigos completos publicados em periódicos; artigos aceitos para publicação; trabalhos completos publicados em anais de congressos; resumos expandidos publicados em anais de congressos; resumos publicados em anais de congressos; livros publicados organizados ou edições; e capítulos de livros publicados. Expressões regulares foram desenvolvidas para identificar os campos de cada uma destas categorias. Apenas para exemplificar, os seguintes campos foram extraídos de cada artigo completo publicado em periódico: título, local, páginas, volume, autores, ano de publicação, nome do periódico, número e ISSN. Para cada currículo processado foi produzido um arquivo XML para estruturar as informações identificadas.

3.3 MODELAGEM E POPULAÇÃO DO BANCO DE DADOS

Para organizar as informações obtidas dos Currículos Lattes, um modelo entidade relacionamento foi especificado e um banco de dados foi criado dentro do SGBD *PostgreSQL*. O esquema deste banco de dados pode ser observado na Figura 1. Cada uma das tabelas é descrita de ma-

neira sucinta no Quadro 1. Neste artigo, cada pessoa que possui um currículo Lattes será chamada de pesquisador.

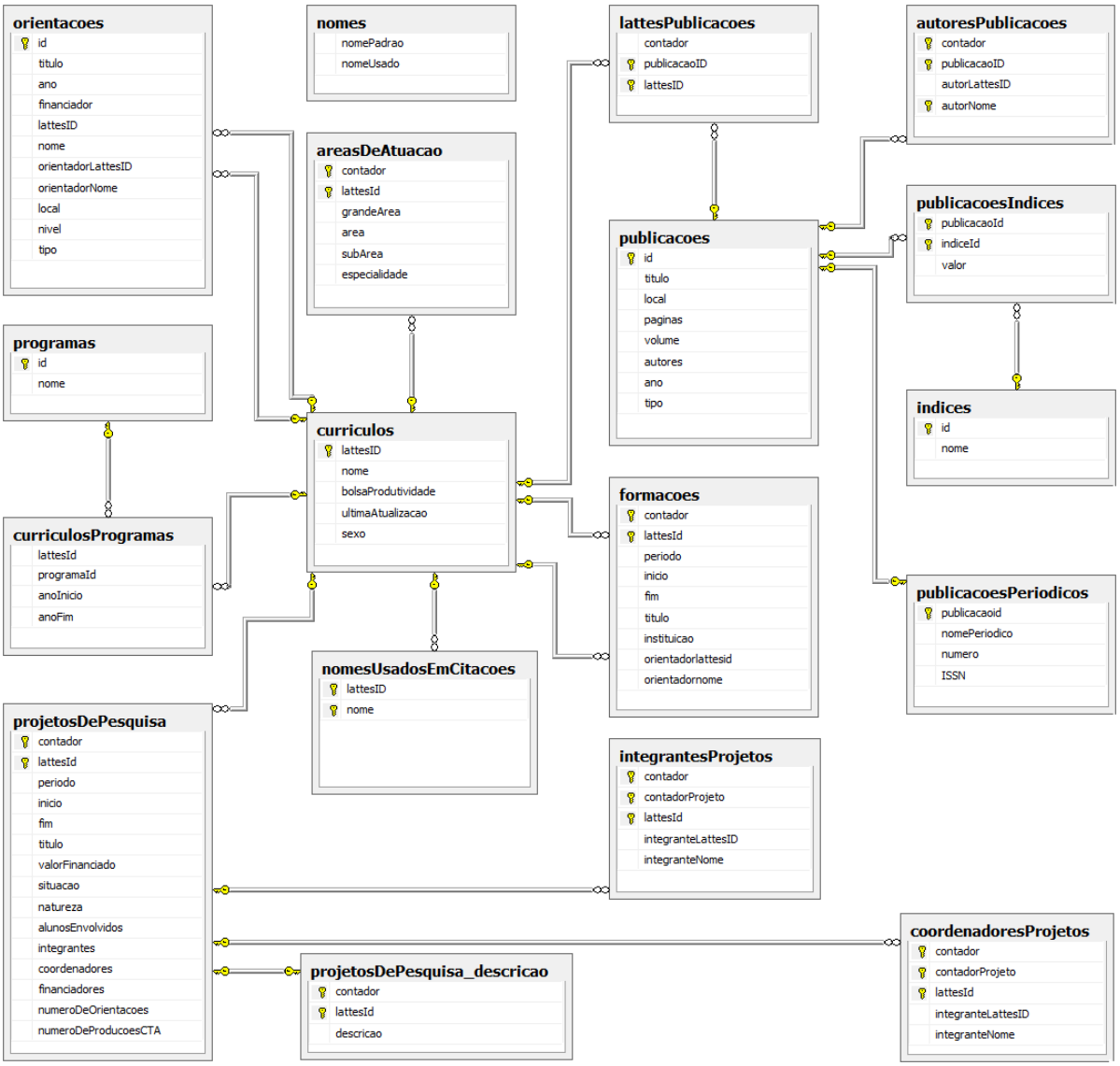

Figura 1. Diagrama entidade relacionamento do banco de dados desenvolvido

Fonte: elaboração própria

Todos os arquivos XML produzidos foram lidos e seus conteúdos inseridos no banco de dados.

Ao todo, 88.402.026 registros foram cadastrados. A Tabela 1 apresenta o total de registros cadastrados em cada uma das tabelas supracitadas.

Tabela	Descrição
Curriculos	tabela contendo as informações gerais de cada currículo: identificador (*lattesID*), nome, tipo de bolsa produtividade, sexo e última atualização do currículo.
NomesUsadosEmCitacoes	tabela com os nomes/abreviações usadas pelo autor em suas citações.
Formacoes	tabela contendo dados sobre a formação de uma dada pessoa, incluindo o período de formação, o título, a instituição, o identificador do orientador e o nome do orientador.
AreasDeAtuacao	contém a lista das áreas de atuação informadas pelo pesquisador (contendo grande área, área, subárea e especialidade).
ProjetosDePesquisa	lista dos projetos de pesquisa nos quais cada pesquisador está envolvido.
ProjetosDePesquisa_descricao	descrição dos projetos de pesquisa, na forma de texto livre.
CoordenadoresProjetos	tabela para listar os coordenadores de cada projeto de pesquisa.
IntegrantesProjetos	tabela para listar os integrantes de cada projeto de pesquisa.
Publicacoes	tabela para armazenar os registros de todas as publicações de todos os currículos. Há sete tipos de publicações que estão sendo consideradas: artigos completos publicados em periódicos; artigos aceitos para publicação; trabalhos completos publicados em anais de congressos; resumos expandidos publicados em anais de congressos; resumos publicados em anais de congressos; livros publicados organizados ou edições; e capítulos de livros publicados.
LattesPublicacoes	tabela para vincular cada currículo ao conjunto de publicações cadastradas pelo pesquisador.
AutoresPublicacoes	tabela com o conjunto de (co)autores de cada publicação.
Orientacoes	lista das orientações feitas por cada pesquisador. Há sete tipos de orientação que estão sendo consideradas: orientações de pós-doutorado; teses de doutorado; orientações de outra natureza; dissertações de mestrado; monografias de conclusão de curso de aperfeiçoamento; iniciações científicas; e trabalhos de conclusão de graduação.

Quadro 1. Descrição das tabelas apresentadas na Figura 1

Fonte: elaboração própria

Tabela 1. Número de registros em cada uma das tabelas do banco de dados

Tabela	Número de Registros
projetosdepesquisa_descricao	1.069.884
curriculos	1.236.548
nomesusadosemcitacoes	1.353.467
projetosdepesquisa	1.378.885
coordenadoresprojetos	1.380.087
formacoes	3.250.846
areasdeatuacao	3.256.019
orientacoes	4.329.993
integrantesprojetos	4.915.223
publicacoes	11.529.218
lattespublicacoes	11.529.218
autorespublicacoes	43.172.638
Total	**88.402.026**

Fonte: elaboração própria

Além dessas tabelas, algumas tabelas adicionais foram criadas para permitir ou facilitar análises mais sofisticadas sobre o conjunto de dados. Por exemplo, a tabela *Nomes* contém os nomes na forma que foram usados pelo pesquisador e uma versão canônica (sem acentos e caracteres especiais) para facilitar a comparação entre nomes registrados pelo pesquisador e por seus coautores, orientandos ou orientadores. As tabelas *Indices* e *PublicacoesIndices* foram criadas para armazenar informações adicionais sobre cada publicação, por exemplo, JCR e número de citações (informações que não necessariamente estão presentes nos currículos Lattes, mas que serão obtidas numa etapa futura de enriquecimento da base de dados). Além disso, um campo adicional foi inserido na tabela *Publicacoes* chamado *idUnico* que está sendo usando para relacionar diferentes registros de publicações que se referem à mesma entidade (por exemplo, três coautores registraram o mesmo artigo, então há três registros deste artigo na base e o campo *idUnico* conterá o mesmo valor para esses registros). Esta informação também não pode ser obtida diretamente/explicitamente dos currículos, mas, como será visto na Subseção 4.2, ela pode ser inferida das informações presentes nos currículos.

4 CARACTERIZAÇÃO DO BANCO DE DADOS DE CURRÍCULOS

Esta seção contém a caracterização da distribuição dos dados nos currículos, uma breve análise sobre dados e um conjunto de dicas e cuidados que devem ser considerados na análise de Currículos Lattes.

4.1 DESCRIÇÃO DO BANCO DE DADOS

Nesta seção serão apresentadas algumas características do banco de dados formado com as informações dos currículos analisados.

Na média, cada pesquisador informou que atua em 2,61 grandes áreas, áreas ou subáreas diferentes. Em mais detalhes, 263.775 pesquisadores informaram que atuam em mais de uma grande área (das oito grandes áreas disponíveis para seleção na Plataforma Lattes). A Figura 2 apresenta a distribuição dos pesquisadores nas grandes áreas.

As grandes áreas são divididas em dezenas de áreas, porém as 13 áreas mais informadas pelos pesquisadores correspondem, juntas, a mais de 50% do total de declarações de áreas de atuação. A Figura 3 apresenta a distribuição dos pesquisadores nas áreas de atuação mais frequentemente informadas.

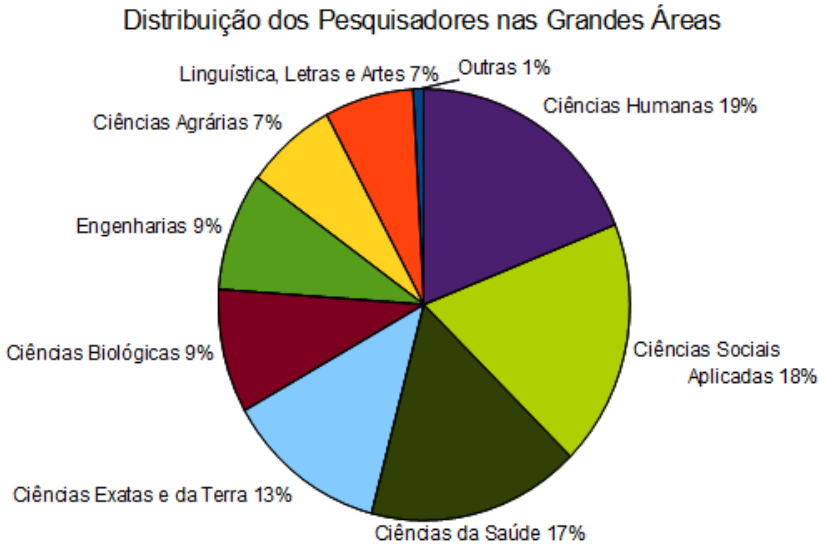

Figura 2. Distribuição nas grandes áreas de atuação

Fonte: elaboração própria a partir de dados da pesquisa

Conforme apresentado, todos os currículos foram baixados em maio de 2011. Enquanto uma grande parcela dos currículos foi atualizada nos últimos doze meses (contados de junho de 2010 a maio de 2011), há diversos currículos que estão desatualizados há diversos anos. Na média, cada currículo foi atualizado pela última vez há 27 meses, mas a mediana é de apenas 10 meses. As Figuras 4 e 5 apresentam, respectivamente, o ano da última atualização dos currículos e o período (em intervalos de doze meses contados a partir de maio de 2011) da última atualização. Pode-se observar que mais de 75% dos currículos foram atualizados nos últimos 36 meses.

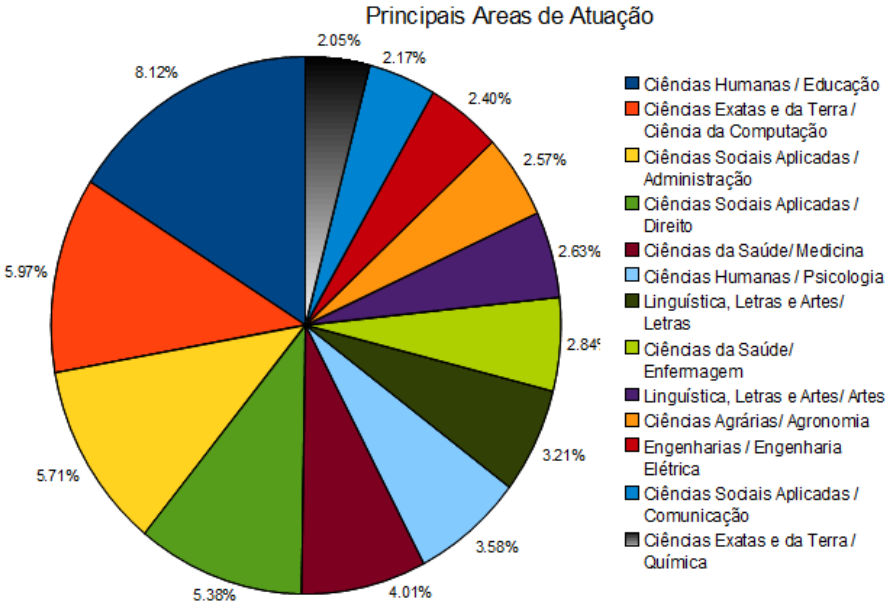

Figura 3. Áreas de atuação mais frequentes

Fonte: elaboração própria a partir de dados da pesquisa

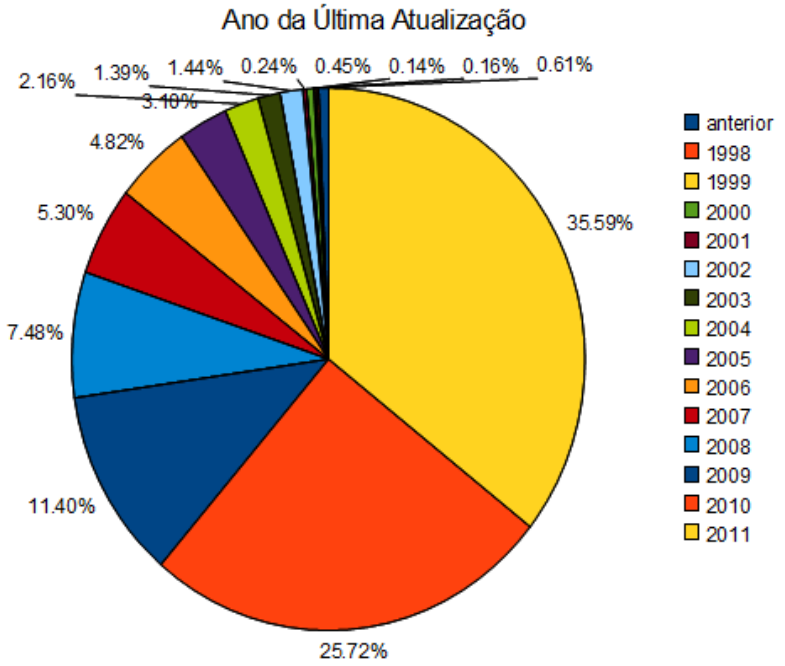

Figura 4. Ano da última atualização do currículo

Fonte: elaboração própria a partir de dados da pesquisa

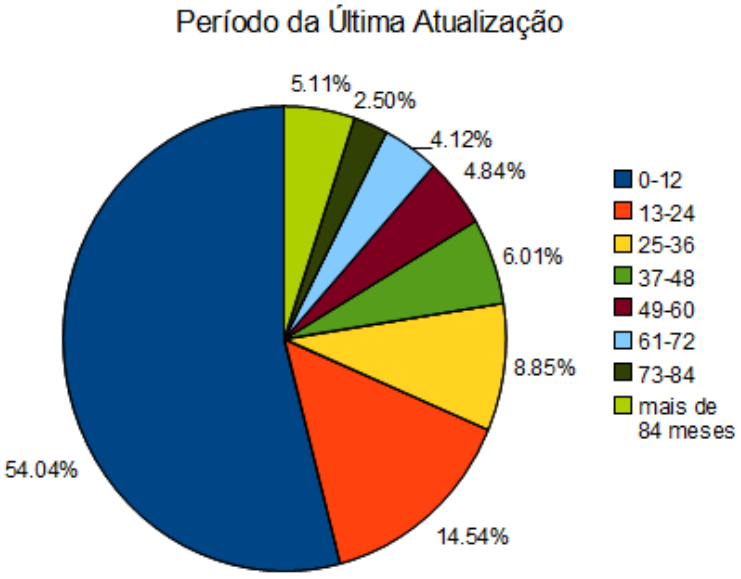

Figura 5. Período da última atualização do currículo

Fonte: elaboração própria a partir de dados da pesquisa

Nos currículos analisados foram encontrados 11.529.218 de registros de publicações, ou seja, uma média de 9,32 publicações por currículo. É importante lembrar que esses mais de 11,5 milhões de registros de publicações possuem redundâncias, pois diferentes coautores inserem uma mesma publicação em seus currículos.

As publicações estão organizadas em sete tipos, conforme pode ser observado na Figura 6. Na média, cada publicação tem 3,74 autores. A Figura 7 apresenta as médias de autores para cada um dos tipos de publicação.

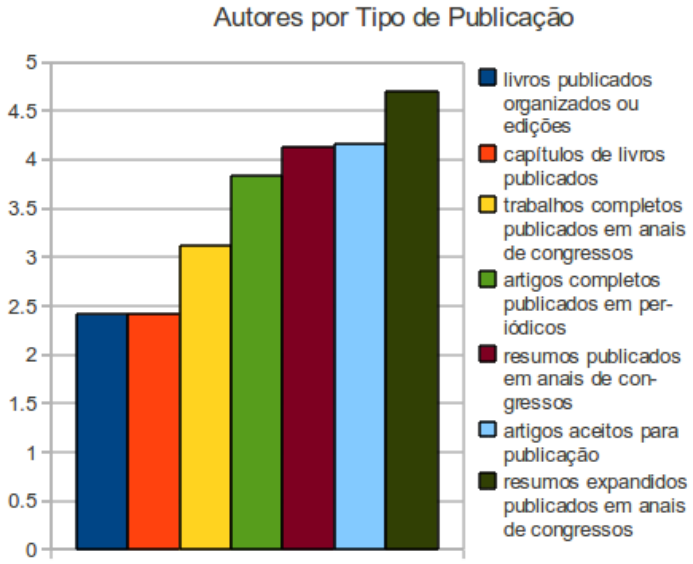

Figura 6. Distribuição das publicações por tipo

Fonte: elaboração própria a partir de dados da pesquisa

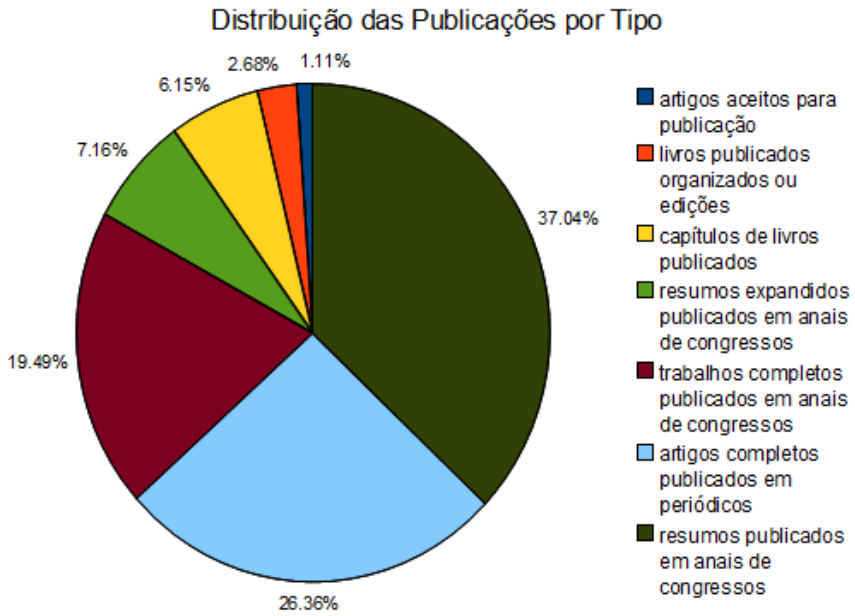

Figura 7. Distribuição das publicações por tipo

Fonte: elaboração própria a partir de dados da pesquisa

A Figura 8 apresenta a evolução do número de registros de publicações cadastrados nos currículos ao longo dos anos. Nesta figura podemos observar um crescimento no número de publicações ano a ano. Nos últimos três anos esse crescimento não pode ser observado, mas isto pode ocorrer devido à não atualização dos currículos (que na média foram atualizados pela última vez há 27 meses).

Uma visualização melhor no número de registros de publicações pode ser feita considerando-se apenas os trinta anos de 1979 até 2008, período no qual a grande maioria dos currículos está atualizada. A Figura 9 apresenta estas informações. Ao observarmos os registros de artigos publicados em periódicos é possível constatar um crescimento exponencial no número desses registros. De fato, nessa janela de trinta anos, o número desses registros cresceu cerca de 12% ao ano.

Dos 1.236.548 currículos cadastrados, 13.797 possuem Bolsa Produtividade do CNPq (1,12% dos pesquisadores). A Figura 10 apresenta a distribuição das bolsas em seus sete níveis.

No banco de dados produzido, há 4.329.993 orientações cadastradas (cerca de 3,5 orientações por currículo). Estas orientações estão divididas em sete categorias, conforme ilustrado na Figura 11. As orientações de mestrado e doutorado correspondem a pouco menos de 16,5% do total de orientações.

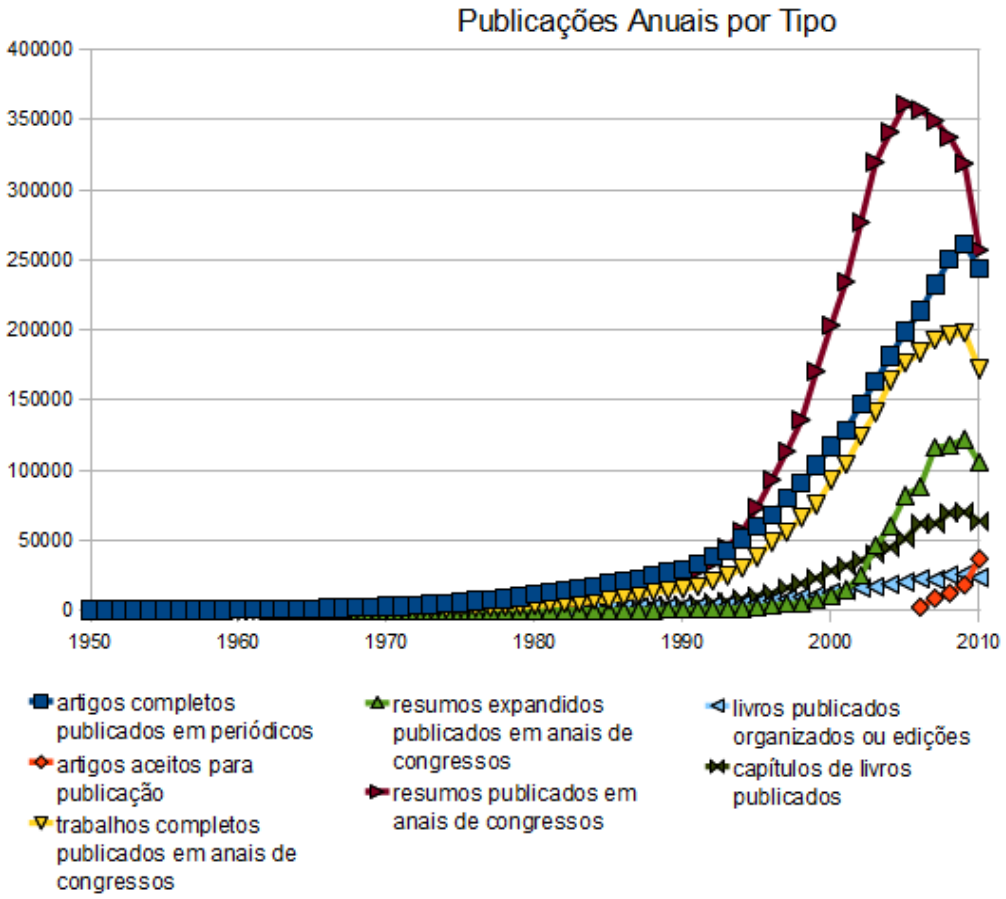

Figura 8. Evolução no número de publicações no tempo

Fonte: elaboração própria a partir de dados da pesquisa

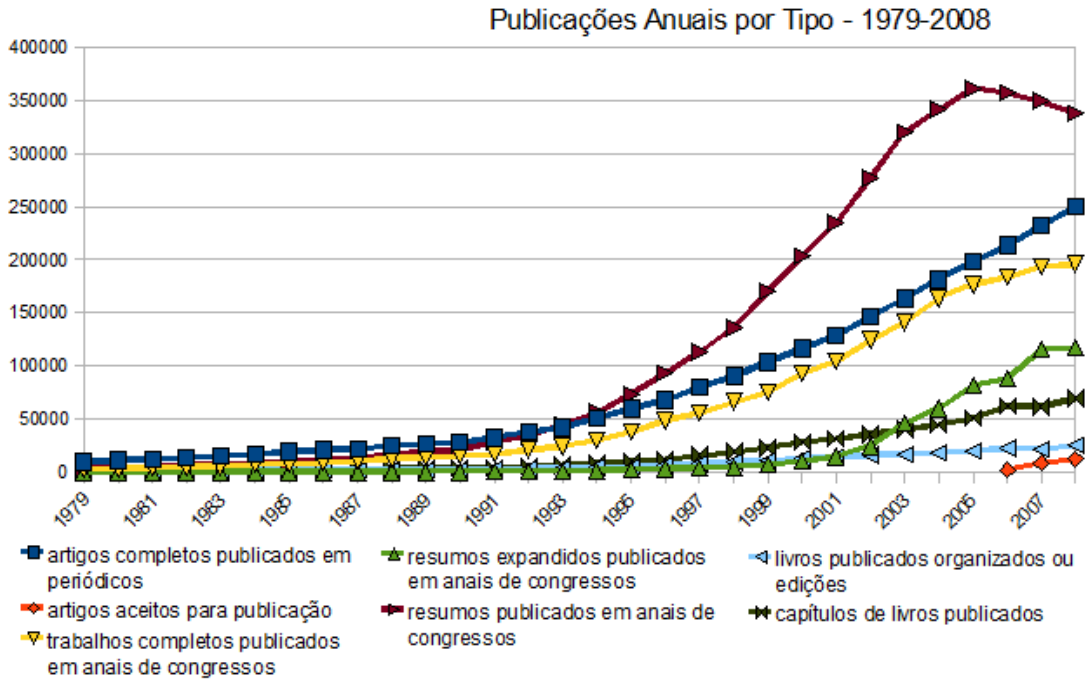

Figura 9. Evolução das Publicações no Período de 1979 a 2008

Fonte: elaboração própria a partir de dados da pesquisa

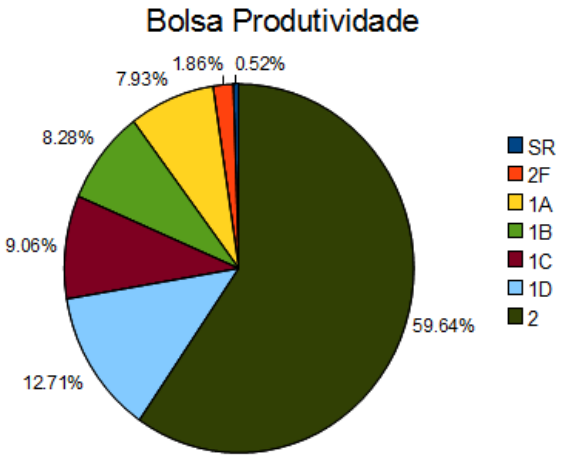

Figura 10. Distribuição das bolsas produtividade

Fonte: elaboração própria a partir de dados da pesquisa

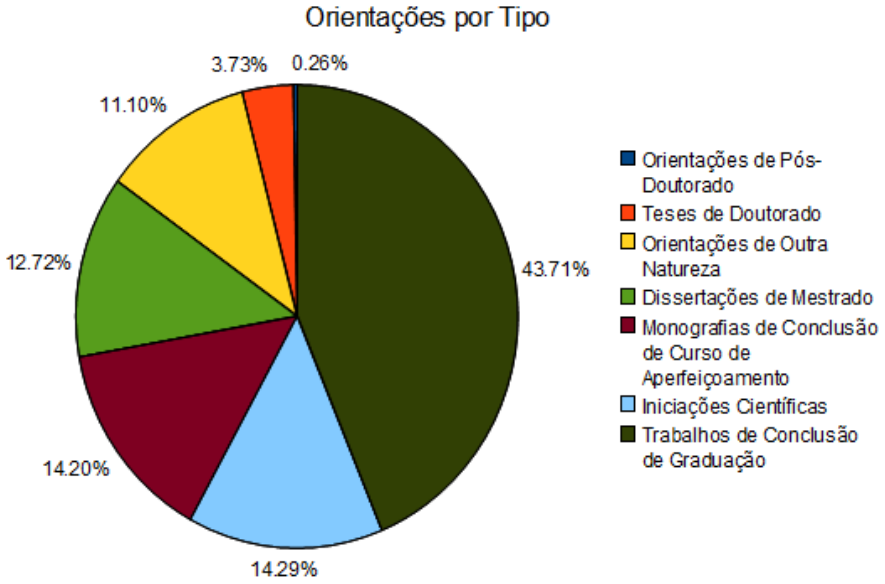

Figura 11. Distribuição das orientações por categoria

Fonte: elaboração própria a partir de dados da pesquisa

A Figura 12 apresenta a evolução da quantidade de orientações de cada tipo no período de trinta anos, de 1979 a 2008. Duas informações merecem destaque nesta evolução: a primeira é a grande quantidade de trabalhos de conclusão de curso que vêm sendo cadastrados nos últimos anos. Conforme apresentado na Figura 12, este tipo de orientação corresponde a mais de 43% do total de orientações. A segunda informação é a relação entre o número de orientações de mestrado e doutorado. Considerando esta janela de trinta anos, nos dez primeiros anos de análise, para cada 5,4 orientações de mestrado havia uma orientação de doutorado. Ao considerar os dez anos seguintes essa relação caiu para 3,9 para 1. Observando os dez últimos anos desse período (de 1999 a 2008) a relação caiu para menos de 3,3 para 1, indicando um crescimento proporcional bem maior no número de orientações de doutorado em relação às orientações de mestrado.

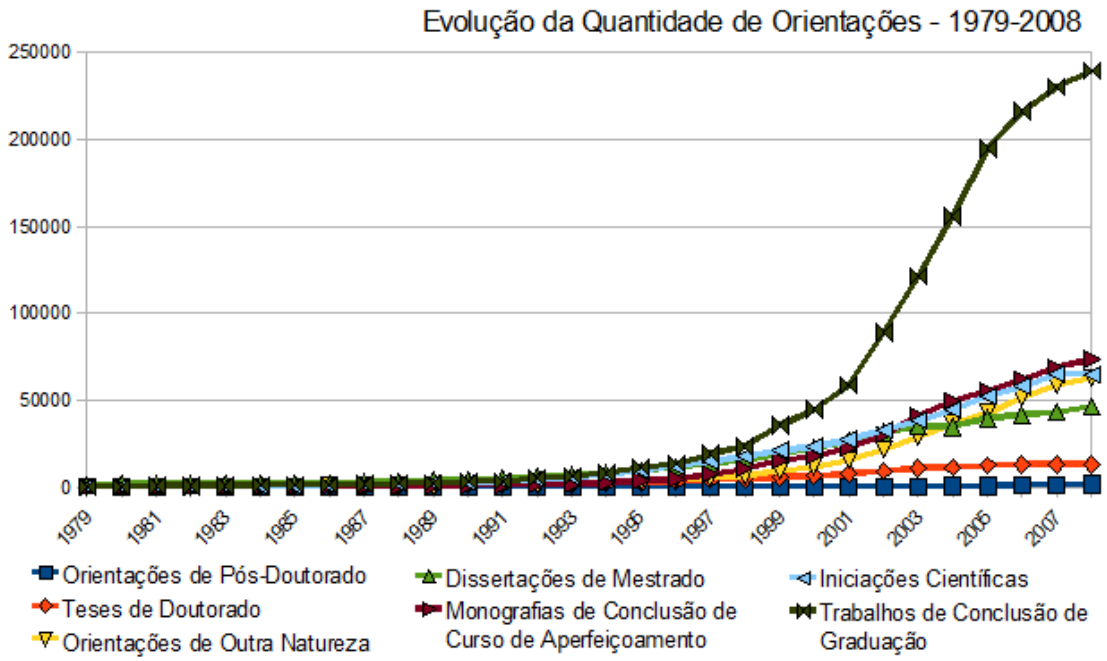

Figura 12. Quantidade de orientações por tipo ao longo dos anos

Fonte: elaboração própria a partir de dados da pesquisa

4.2 RELAÇÕES ENTRE CURRÍCULOS

Uma das grandes motivações da criação do banco de dados de Currículos Lattes é a geração e análise das redes sociais de pesquisadores. Para a montagem dessas redes é necessário estabelecer quais características serão consideradas para relacionar diferentes currículos.

Nesta seção serão apresentadas algumas das características mais óbvias para a criação dessas redes, bem como a quantidade de relações que já foram obtidas considerando cada uma delas.

A primeira relação obtida é a de coautorias, ou seja, dois ou mais pesquisadores que são coautores de uma mesma publicação. Na base de dados há 11.529.218 registros de publicações, porém vários desses registros representam a mesma publicação referenciada em diferentes currículos. Uma consulta simples à base de dados, procurando apenas por publicações com o mesmo título e do mesmo tipo permite identificar 1.843.464 relações de coautoria. Este critério de publicações com o mêsmo título e do mesmo tipo não é robusto o suficiente para garantir que dois registros diferentes se refiram à mesma publicação, bem como haverá registros referentes à mesma publicação com títulos diferentes (devido a erros no preenchimento, excesso ou falta de espaços ou pontuações etc.). Assim, é necessário o desenvolvimento de métodos eficientes e eficazes de resolução de entidades. De qualquer forma, esse valor de mais de 1,8 milhão de relações identificadas é bastante relevante. Uma análise mais detalhada e com tratamento mais sofisticado dos dados de coautoria é realizada em Digiampietri *et al.* (2015).

Uma segunda maneira de utilizar a relação de coautorias é através da verificação dos nomes e identificadores dos autores de cada uma das publicações. Essa relação, da maneira que está disponível nos Currículos Lattes não liga dois registros de publicações, mas relaciona a publicação de um currículo com outro currículo, servindo de base para a ligação entre currículos. Os 11.529.218 registros de publicações possuem, ao todo, 43.172.638 registros de (co)autores. Destes, há identificadores de currículos de 21.102.745 (co)autores, sendo que 11.529.218 são dos donos dos currículos nos quais a publicação foi informada. Assim, restam 9.573.527 relações de coautoria identificadas.

Outra relação relevante quando analisados os currículos de pesquisadores é a relação orientador/orientando. Ao se cruzar as informações de formação de um pesquisador com as informações de orientação de outro foi possível identificar um total de 210.112 relações de orientação de doutorado. Aqui, só foram consideradas as relações onde foi possível fazer a verificação dupla: um pesquisador informou que foi orientado por um segundo pesquisador e este segundo informou que foi orientador do primeiro.

É possível também relacionar os pesquisadores por áreas (grandes áreas, áreas, subáreas ou especialidades) de interesse. Devido ao fato de, na média, cada pesquisador informar cerca de 2,6 áreas de interesse, só esta característica originaria milhões de relações entre pesquisadores.

Para este artigo, as relações explícitas existentes nos currículos Lattes foram utilizadas para a montagem de uma rede social. Estas relações são formadas principalmente por relações de coautoria (quando no nome de um dos autores de um artigo há uma ligação [um *link* html] para o currículo deste autor); mas há também relações de orientação; co-participação em projetos de pesquisa e participação em bancas. Ao todo foram encontrados 14.411.364 relações entre currículos de pesquisadores envolvendo 424.558 currículos (incluindo múltiplas relações entre o mesmo par de currículos). As relações explícitas correspondem às ligações de um currículo referenciando outro. Desta forma, as relações são direcionadas. Do total de 424.558 currículos envolvidos nestas relações apenas 280.834 possuem *links* (arestas) para outros currículos, ou seja, 143.724 currículos são referenciados por outros currículos, porém sem referenciá-los.

A Tabela 2 apresenta um resumo da quantidade total de relações entre currículos, considerando-se apenas os 280.834 currículos que possuem alguma relação explícita com outros currículos. Destaca-se que a média de relações de cada um destes currículos é acima de 41, porém a mediana é de apenas 3 relações. Outra informação interessante é que o currículo com maior quantidade de relacionamentos possui 14.302 *links* que referenciam outros currículos. Sendo que destas relações, 1.920 são múltiplas relações entre o mesmo par de currículos. Este par de currículos foi investigado e corresponde a um casal de pesquisadores, livres-docen-

tes, ambos doutores em Cardiologia, sendo que um deles foi orientador de doutorado do outro.

Tabela 2. Resumo das relações explícitas entre currículos

Característica	Número de relações
Média	41,48
Mediana	3
Desvio padrão	130,49
Total	14.411.364
Mínimo	1
Máximo	14.302

Fonte: elaboração própria a partir de dados da pesquisa

Ao se contar uma única vez o total de relações entre cada par de pesquisadores, ou seja, ao se contar as relações entre pesquisadores diferentes, observou-se um total de 1.923.147 relações direcionadas e 1.125.193 relações não direcionadas (ou seja, do total de relações entre pares de pesquisadores, 797.954 aparecem em uma única direção entre os currículos). A Tabela 3 contém um resumo destas relações. Na média, cada currículo possui referências explícitas para cerca de 5,54 currículos. Porém, a mediana é apenas 1. Há um pesquisador cujo currículo aponta para outros 598 pesquisadores. Este pesquisador é um professor livre-docente, doutor em Psicofarmacologia e bolsista de produtividade 1A.

Tabela 3. Resumo das relações entre currículos diferentes

Característica	Número de relações
Média	5,54
Mediana	1
Desvio padrão	11,50
Total	1.923.147
Mínimo	1
Máximo	428

Fonte: elaboração própria a partir de dados da pesquisa

Uma rede social foi montada onde os nós correspondem aos 424.558 pesquisadores envolvidos em ao menos uma relação com outro pesquisador e as arestas correspondem às 1.125.193 relações não direcionadas. O grafo correspondente à rede possui 2.878 componentes conexos, sendo a grande maioria (1.727) de tamanho 2 (ou seja, dois pesquisadores que estão ligados entre si, mas não têm nenhuma outra relação com os outros pesquisadores do conjunto de dados considerado). O maior componente conexo é formado por 415.483 pesquisadores (ou seja, cerca de 97,86% do total de pesquisadores que possuem ao menos uma relação). A Figura 13 apresenta a distribuição da quantidade de componentes conexos de acordo com seus tamanhos.

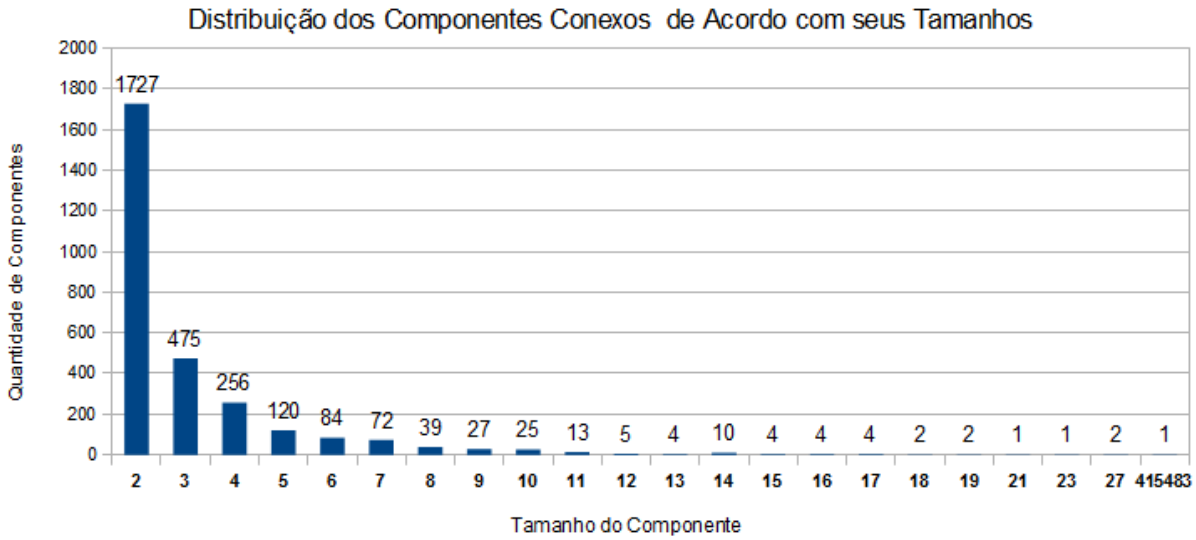

Figura 13. Distribuição dos componentes conexos

Fonte: elaboração própria a partir de dados da pesquisa

Para uma análise mais completa da rede social formada, foi utilizada a informação das grandes áreas de conhecimento. De todos os currículos no banco de dados formado, 880.332 declararam atuar em apenas uma das oito grandes áreas de conhecimento do CNPq. Os demais currículos ou não tinham nenhuma grande área associada, ou estavam associados a mais de uma grande área, ou estavam associados à grande área 'Outras'. A Tabela 4 apresenta a distribuição nestas nove categorias dos 424.558 envolvidos em relações.

Tabela 4. Distribuição dos Pesquisadores nas Grandes Áreas

Grande área do conhecimento	Número de pesquisadores
Ciências Agrárias	19.762
Ciências Biológicas	21.824
Ciências da Saúde	67.392
Ciências Exatas e da Terra	33.445
Ciências Humanas	47.978
Ciências Sociais Aplicadas	53.413
Engenharias	23.258
Linguística, Letras e Artes	17.260
Outras/Mais de uma	140.226

Fonte: elaboração própria a partir de dados da pesquisa

A Tabela 5 apresenta o número de relações existentes entre currículos pertencentes a cada uma das grandes áreas. Destaca-se a pequena proporção de relações entre currículos de uma mesma grande área, por exemplo, há 131.858 relações onde ao menos um dos pesquisadores disse atuar apenas na grande área de *Engenharia*. Porém, deste total de relações, apenas 4.174 são entre dois pesquisadores exclusivamente desta grande área.

Tabela 5. Número de relações considerando as diferentes grandes áreas

	Ciências Agrárias	Ciências Biológicas	Ciências da Saúde	Ciências Exatas e da Terra	Ciências Humanas	Ciências Sociais Aplicadas	Engenharias	Linguística, Letras e Artes	Outras/Mais de uma
Ciências Agrárias	**2912**	6547	19732	9975	14293	16116	7021	5048	30365
Ciências Biológicas	6547	**3527**	22084	10952	15823	17733	7629	5506	32969
Ciências da Saúde	19732	22084	**33655**	33842	47630	54271	23293	17463	102599
Ciências Exatas e da	9975	10952	33842	**8783**	24781	27366	12243	8817	52232
Ciências Humanas	14293	15823	47630	24781	**17237**	39251	16673	12419	73226
Ciências Sociais	16116	17733	54271	27366	39251	**22103**	19039	14097	83158
Engenharias	7021	7629	23293	12243	16673	19039	**4174**	6035	35751
Linguística, Letras e Artes	5048	5506	17463	8817	12419	14097	6035	**2189**	26594
Outras/Mais de uma	30365	32969	102599	52232	73226	83158	35751	26594	**78040**

Fonte: elaboração própria a partir de dados da pesquisa

A Figura 14 apresenta a porcentagem das relações dos currículos de cada uma das grandes áreas. É possível observar que, por exemplo, dos currículos dos pesquisadores que disseram atuar apenas na área de *Ciências Agrárias*, a maioria de suas relações envolve pesquisadores de *Outras/Mais de uma grande área* (cerca de 27% das relações); *Ciências da Saúde* (cerca de 17,5%); e *Ciências Sociais Aplicadas* (cerca de 14,5%).

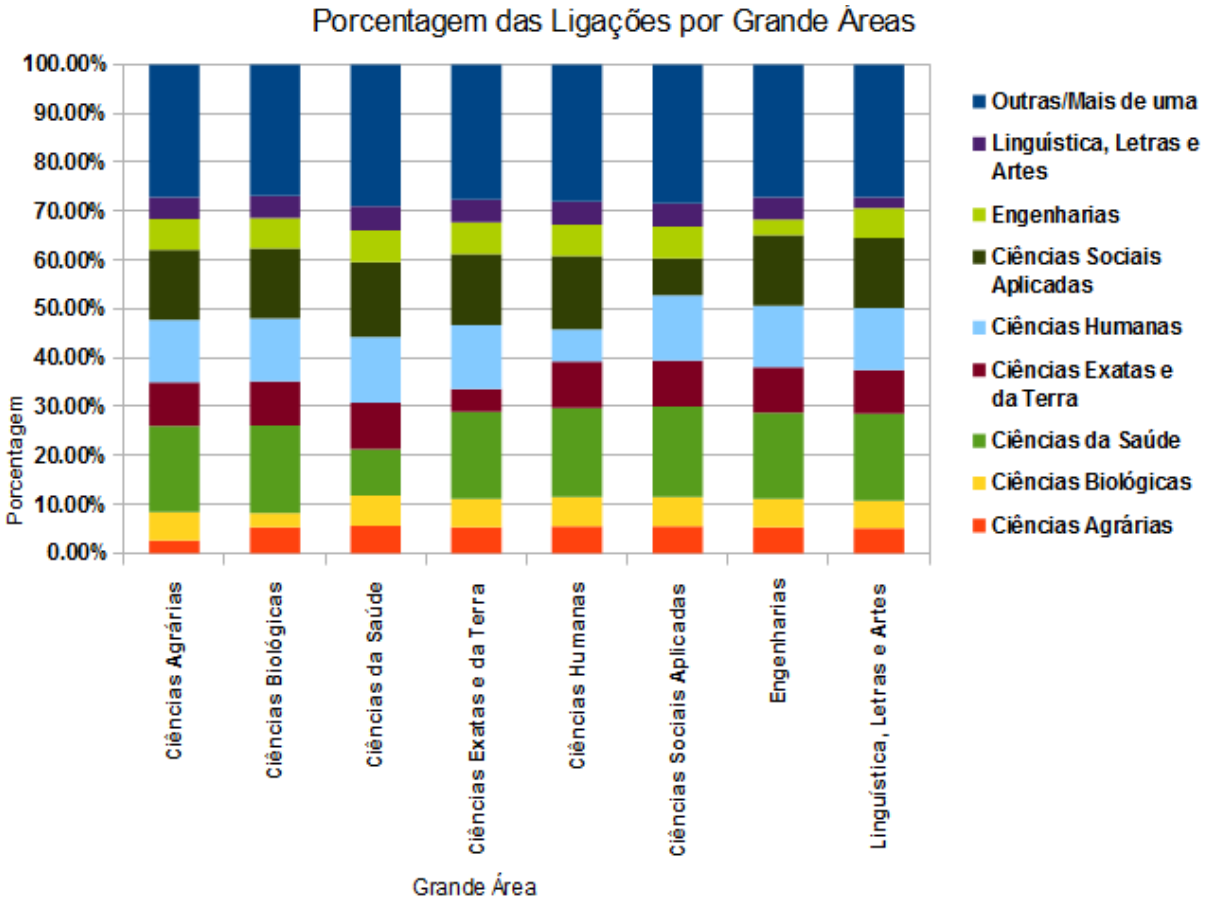

Figura 14. Porcentagem das relações entre pesquisadores

Fonte: elaboração própria a partir de dados da pesquisa

A Figura 15 apresenta a rede social formada pelas relações explícitas existentes nos currículos Lattes. É possível observar que a grande maioria dos currículos pertence a um único componente conexo. Há algumas concentrações de nós de uma mesma cor (isto é, pessoas que atuam em uma mesma grande área) em certas regiões do grafo, mas, de um modo geral, as cores estão bastante misturadas indicando a presença de relacionamentos interdisciplinares.

A análise da rede social acadêmica brasileira produzida a partir de dados dos currículos Lattes é uma atividade importante e desafiadora (MENA-CHALCO *et al.,* 2014), porém foge ao escopo do presente trabalho. Desta forma, a Figura 15 visa a apenas ilustrar essa rede.

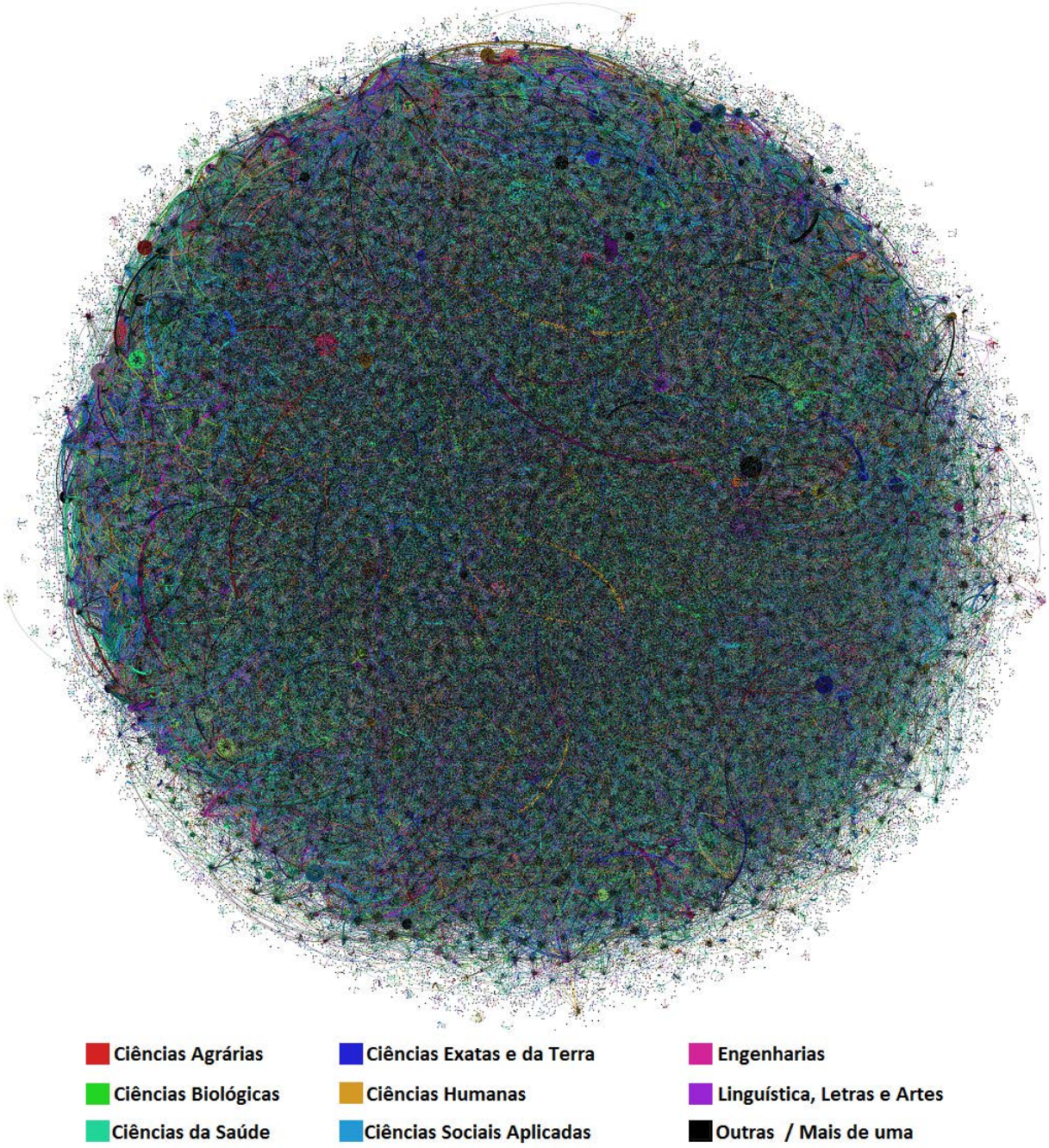

Figura 15. Rede de relações explícitas

Fonte: elaboração própria a partir de dados da pesquisa

4.3 CUIDADOS E PROBLEMAS RELACIONADOS AOS DADOS

Os currículos preenchidos na Plataforma Lattes possuem certa padronização imposta pelos formulários da plataforma e algumas verificações de valores (por exemplo, o ano de conclusão de um projeto não pode ser anterior ao ano de início).

Porém, algumas das verificações só foram introduzidas nos últimos anos, o que permite que informações mais antigas estejam inconsistentes.

Além disso, grande parte da informação é preenchida manualmente, o que possibilita a ocorrência de diversos tipos de problemas que precisam ser tratados durante o processamento e análise dos currículos.

Um primeiro cuidado que deve ser considerado é a existência de currículos de homônimos na Plataforma Lattes, isto por si só não seria um problema, mas precisa ser considerado sempre que o nome do pesquisador for utilizado para o estabelecimento de qualquer tipo de relação entre currículos. No banco de dados produzido neste artigo foram encontrados 22.169 currículos de homônimos (1,79% dos currículos do banco).

Um problema que ocorria principalmente nas versões mais antigas da Plataforma Lattes era o preenchimento incompleto das informações. Em suas primeiras versões havia uma menor quantidade de campos obrigatórios e de mecanismos de verificação. Assim, não é incomum encontrar um registro de publicação sem o nome de pelo menos um autor; publicações em revista sem indicação do nome da revista e assim por diante.

Um problema comum e cuja solução é difícil de ser totalmente automatizada é o preenchimento incorreto de informações. Por exemplo, é comum que coautores preencham os campos referentes a uma mesma publicação de maneiras diferentes, variando a grafia do título do artigo, veículo de publicação, nome dos coautores e assim por diante. Além disso, diferentes autores classificam a mesma publicação de diferentes maneiras (por exemplo, um autor diz que uma publicação é um resumo e outro diz que é um resumo expandido). Desta forma, o processamento das informações dos currículos deve incluir mecanismos de casamento aproximado de *strings* e/ou análise de padrões.

Outra característica dos dados de Currículos Lattes é a diferença entre as atualizações dos currículos (DIGIAMPIETRI *et al.,* 2014). Alguns autores atualizam seus currículos mensalmente enquanto outros atualizam menos de uma vez por ano. Mesmo um currículo atualizado recentemente pode conter dados desatualizados. Por exemplo, o banco de dados contém mais de 10.000 artigos classificados como *aceitos para publicação* entre 2006 e 2007, sendo que a maioria dos currículos possuidores destes artigos foi atualizada depois de maio de 2010.

5 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Os currículos da Plataforma Lattes contêm uma quantidade muito grande e diversificada de informações que podem ser utilizadas como base para a construção e análise de redes sociais de pesquisa, sendo uma das bases de pesquisadores mais completas do mundo.

Enquanto algumas relações podem ser extraídas diretamente desta plataforma, há diversas outras que podem ser obtidas por meio de algoritmos de resolução de entidades. O banco de dados também pode ser enriquecido com informações relacionadas a publicações (número de citações de cada artigo, por exemplo); dos veículos de publicação (obtenção

de índices como JCR e SJR) e dados derivados dessas informações, como índices G e H.

Além disso, cada pesquisador pode ser caracterizado por diferentes índices como *Page Rank* ou *Author Rank*, ou novos índices podem ser criados exclusivamente para analisar pesquisadores com base em sua produção científica e/ou orientações.

Neste artigo foram apresentados os primeiros passos na direção de uma análise ampla dos dados de Currículos Lattes, das relações entre esses dados e das redes formadas pelos seus pesquisadores. Uma primeira rede social acadêmica foi construída a partir das relações explícitas existentes nos currículos Lattes e algumas de suas características foram analisadas.

Como trabalhos futuros, pretende-se desenvolver algoritmos robustos para a resolução de entidades de forma a determinar uma maior quantidade de relações de orientação e coautoria. Pretende-se também criar e analisar redes de pesquisadores considerando diferentes relações e diferentes métricas de redes.

AGRADECIMENTOS

O trabalho apresentado neste artigo foi parcialmente financiado pela FAPESP (Projeto Jovem Pesquisador processo 2009/10413-5 e Bolsa de Iniciação Científica processo 2013/06084-1), pelo CNPq (Bolsa de Iniciação Científica e Bolsa Produtividade em Pesquisa processos 304937/2010-0 e 306046/2013-0) e pelo Programa de Educação Tutorial (MEC/SESu).

REFERÊNCIAS

ALVES, A.; YANASSE, H.; SOMA, N. Sucupira: a system for information extraction of the Lattes platform to identify academic social networks. *Proceedings of the 6th Iberian Conference on Information Systems and Technologies (CISTI),* p. 1-6, 2011.

BALANCIERI, R.; BOVO, A. B.; KERN, V. M.; PACHECO, R. C. D. S.; BARCIA, R. M. A análise de redes de colaboração científica sob as novas tecnologias de informação e comunicação: um estudo na Plataforma Lattes. *Ciência da Informação*, v. 34, p. 64-77, 2005. http://dx.doi.org/10.1590/S0100-19652005000100008

DIGIAMPIETRI, L. A.; DA SILVA, E. E. A framework for social network of researchers analysis. *Iberoamerican Journal of Applied Computing*, v. 1, p. 1-24, 2011.

DIGIAMPIETRI, L. A.; MENA-CHALCO, J. P.; PÉREZ-ALCÁZAR, J. J.; TUESTA, E. F.; DELGADO, K.; MUGNAINI, R. Minerando e caracterizando dados de currículos Lattes. *I Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2012)*, 2012.

DIGIAMPIETRI, L. A.; MUGNAINI, R.; MENA-CHALCO, J. P.; DELGADO, K. V.; ALCAZAR, J. J. P. Análise da atualização dos currículos Lattes. *Anais do IV Encontro Brasileiro de Bibliometria e Cientometria (EBBC),* 2014.

DIGIAMPIETRI, L. A.; MENA-CHALCO, J. P.; SILVA, G. S.; OLIVEIRA, L. B.; LIMA, J. J. S.; MALHEIRO. A. P., MEIRA, D. Análise da evolução das relações de coautoria nos programas de pós-graduação em computação no Brasil. *Revista Eletrônica de Sistemas de Informação,* v. 14, n. 1, 2015.

GUIMARÃES, J. A. A pesquisa médica e biomédica no Brasil: comparações com o desempenho científico brasileiro e mundial. *Ciência & Saúde Coletiva*, v. 9, p. 303-327, 2004. http://dx.doi.org/10.1590/S1413-81232004000200009

LEITE, P.; MUGNAINI, R.; LETA, J. A new indicator for international visibility: exploring Brazilian scientific community. *Scientometrics*, v. 88, p. 311-319, 2011. http://dx.doi.org/10.1007/s11192-011-0379-9

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, p. 31-39, 2009. http://dx.doi.org/10.1007/BF03194511

MENA-CHALCO, J. P.; DIGIAMPIETRI, L.; CESAR JUNIOR, R. M.; LOPES, F. M. Brazilian bibliometric coauthorship networks. *Journal of the Association for Information Science and Technology*, v. 65, p. 1424-1445, 2014. http://dx.doi.org/10.1002/asi.23010

MUGNAINI, R.; LEITE, P.; LETA, J. Fontes de informação para análise de internacionalização da produção científica brasileira. *Ponto de Acesso*, v. 5, n. 3, 2012.

NEWMAN, M. E. J. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, v. 98, n. 2, p. 404–409, 2001. http://dx.doi.org/10.1073/pnas.98.2.404

SILVA, F. M.; SMIT, J. W. Organização da informação em sistemas eletrônicos abertos de Informação Científica & Tecnológica: análise da Plataforma Lattes. *Perspectivas em Ciência da Informação*, v. 14, p. 77-98, 2009. http://dx.doi.org/10.1590/S1413-99362009000100007