

LattesDoctoralDataset: Uma Coleção de Dados Estratificados sobre o Conjunto de Doutores Cadastrados na Plataforma Lattes

Thiago M. R. Dias¹, Alberto H. F. Laender², Gray F. Moita¹

¹Centro Federal de Educação Tecnológica de Minas Gerais

²Departamento de Ciência da Computação - Universidade Federal de Minas Gerais

thiago@div.cefetmg.br, laender@dcc.ufmg.br, gray@dppg.cefetmg.br

Abstract. *Studies on scientific data have attracted the interest of researchers from several areas of knowledge, in view of their potential to better understand how research in a given area has been carried out, or how groups of researchers have collaborated when developing their work. Thus, this work describes the process of extracting, treating and characterizing a stratified dataset containing information about individuals with curricula registered in the Lattes Platform and holding a PhD degree. It also presents a quantitative description of the collected data, as well as an overall description of the datasets made available.*

Resumo. *Os estudos sobre dados científicos têm atraído o interesse de pesquisadores de diversas áreas do conhecimento, tendo em vista seu potencial para melhor compreender como as pesquisas em uma determinada área têm sido realizadas, ou como grupos de pesquisadores têm colaborado no desenvolvimento de seus trabalhos. Assim, este trabalho descreve o processo de extração, tratamento e caracterização de uma coleção de dados estratificados contendo informação sobre os indivíduos com currículos cadastrados na Plataforma Lattes e que possuam doutorado concluído. O trabalho também apresenta uma descrição quantitativa sobre os dados coletados, bem como uma descrição geral dos conjuntos de dados disponibilizados.*

1. Introdução

Uma nova geração de serviços disponíveis principalmente na Web está mudando a forma de divulgar e disponibilizar a produção científica e tecnológica. Existe, atualmente, uma tendência que reforça a troca de informações e a colaboração entre as pessoas. A forte relação entre os domínios científico e socioeconômico tem gerado um interesse crescente pela compreensão dos mecanismos que norteiam as atividades científicas, sendo possível apontar diversos trabalhos que analisam aspectos específicos como as características da linguagem e dos discursos empregados na comunicação científica (HOFFNAGEL, 2009) ou, ainda, a relação de colaboração entre pesquisadores e grupos de pesquisa (DING, 2011; REVOREDO et al., 2012; STROELE; ZIMBRÃO; SOUZA, 2012).

Para Mugnaini et al. (2014), o levantamento da produção científica de um país permite estudar diversos aspectos que podem ser qualificados como resultados mensuráveis de seu respectivo sistema de ciência, tecnologia e inovação. Acompanhar o

fluxo de comunicação científica das diversas áreas facilita o processo de avaliação dos resultados de pesquisa, cujas características são tão diversificadas quanto o é a própria ciência.

No entanto, o grande volume de dados sobre produção científica disponível em diferentes formatos e em diferentes repositórios dificulta a realização de estudos, bem como a consulta por parte de usuários que necessitam de uma visão unificada desses dados para, por exemplo, possibilitar a identificação de grupos de indivíduos que estejam trabalhando com determinado tema em diferentes instituições ou regiões.

Estudos bibliométricos, principalmente em grandes repositórios bibliográficos, não são tarefas triviais tendo em vista a quantidade de dados a serem analisados e as características dos repositórios que, em sua maioria, não possuem um padrão definido. Atualmente, grande parte desses estudos tem utilizado como principais fontes de dados resultados de consultas a repositórios internacionais que apresentam dados sobre trabalhos científicos, geralmente publicados em periódicos indexados. Entretanto, muitos desses repositórios negligenciam trabalhos publicados em periódicos nacionais que muitas vezes não são indexados e grande parte dos artigos publicados em anais de congressos, que constituem importante meio de publicação de algumas áreas do conhecimento como, por exemplo, a Ciência da Computação (LAENDER et al., 2008).

Assim, é evidente a dificuldade existente para se realizar estudos abrangentes que possam apresentar, de forma ampla, análises sobre a produção científica de um grande conjunto de indivíduos que estejam vinculados a diferentes instituições ou que atuem em áreas distintas, como, por exemplo, o conjunto de todos os pesquisadores com um determinado nível de formação ou de uma determinada área de atuação. Diante disso, este trabalho apresenta uma coleção de dados estratificados extraídos dos currículos de todos os indivíduos com doutorado concluído cadastrados na Plataforma Lattes. Essa coleção, denominada LattesDoctoralDataset, inclui dados sobre a formação, orientações concluídas e em andamento, produção científica e colaborações desses indivíduos, possibilitando, desta forma, a realização de diversos estudos sobre esse segmento de pesquisadores.

2. Fonte de Dados

Para a geração da coleção de dados apresentada neste trabalho, foram coletados da Plataforma Lattes os currículos de todos os doutores ali registrados. Segundo Lane, em artigo publicado na revista Nature (LANE, 2010), medir e avaliar o desempenho acadêmico de seus pares é um fator crucial para qualquer comunidade científica. A autora descreve esforços empregados para a construção de repositórios confiáveis de dados científicos que poderiam permitir análises com o objetivo de explorar e compreender como a ciência tem evoluído. Embora todos os esforços sejam importantes, alguns apresentam problemas que comprometem o sucesso de tais iniciativas. Nesse cenário, a Plataforma Lattes é citada como exemplo de boas práticas para o fornecimento de dados de alta qualidade sobre a produção científica de um país, e de como a sua utilização tem sido incentivada por órgãos federais, instituições acadêmicas e agências de fomento à pesquisa. Por fim, a autora destaca que a Plataforma Lattes é uma das fontes de dados sobre pesquisadores mais confiáveis existentes atualmente.

Para Ferraz, Quoniam e Maccari (2014), até o presente momento, não existe no mundo um repositório curricular nacional semelhante à Plataforma Lattes, sendo que somente repositórios de dados referenciais, de onde se podem extrair referências bibliográficas e fontes de informação secundárias, estão disponíveis para livre acesso. Dessa forma, a Plataforma Lattes é um instrumento da maior importância para o estudo da produção científica brasileira.

Mugnaini, Leite e Leta (2011) destacam que muito embora não se apresente como uma base de indexação e catalogação de publicações científicas, a Plataforma Lattes é uma fonte inesgotável de informação sobre a ciência brasileira, sob diversos aspectos e abordagens. Os autores ressaltam que, apesar de todo o volume de informação disponível, o que se observa ainda é uma baixa frequência de estudos cientométricos realizados por especialistas brasileiros que utilizam a Plataforma Lattes, e que isso é reflexo das limitações de seus mecanismos de recuperação e extração de informação. Os autores ainda destacam o fato de o repositório reunir toda a produção científica brasileira, o que viabilizaria estudos que só seriam possíveis se conduzidas em diversas fontes internacionais, mas a um custo considerável para seus autores.

É importante observar ainda que a análise dos dados dos currículos da Plataforma Lattes pode fornecer informações importantes para compreensão do conhecimento científico brasileiro e como ele tem evoluído, tendo em vista a quantidade de trabalhos recentes que têm considerado esses currículos como principal fonte de dados, como os de Oliveira et al. (2012), Perez-Cervantes, Mena-Chalco e Cesar-Junior (2012), Digiampietri, Mugnaini e Alves (2013), Mena-Chalco et al. (2014), Roos et al. (2014), Furtado et al. (2015), Silva et al. (2016), Brito, Quoniam e Mena-Chalco (2016) e Sidone, Haddad e Mena-Chalco (2017)). No entanto, as restrições de acesso impostas pelo CNPq, como, por exemplo, a necessidade de se validar captchas para acesso a cada um dos currículos, tem limitado bastante o potencial de estudos sobre os dados curriculares da Plataforma Lattes. Além disso, particularidades do repositório, como ambiguidade entre nome de indivíduos e a falta de vínculos explícitos entre os coautores dos trabalhos cadastrados, dificultam bastante a análise dos dados, já que a identificação das colaborações passa a ser uma tarefa não trivial, contribuindo para o fato de que a maioria dos atuais trabalhos tem analisado apenas grupos específicos de indivíduos ou pequenos períodos de tempo.

Neste contexto, este trabalho apresenta um importante recurso para disseminação do conteúdo da Plataforma Lattes, abrangendo uma parcela significativa de seus dados já pré-processados, contendo informações relevantes sobre os doutores com currículos ali cadastrados. Logo, tendo em vista a abrangência do conjunto de indivíduos considerado e a diversidade dos dados disponibilizados, inúmeros estudos poderão ser viabilizados, possibilitando assim ampliar o conhecimento sobre a nossa comunidade científica.

3. Coleta dos Dados

Para a criação dos conjuntos de dados que compõem o LattesDoctoralDataset, utilizou-se um arcabouço denominado LattesDataXplorer (Figura 1) desenvolvido especificamente para a coleta, extração e tratamento de dados da Plataforma Lattes (DIAS, 2016). Esse arcabouço adota técnicas usualmente empregadas na coleta e extração de dados de

documentos disponíveis na Web para realizar essas tarefas sobre os currículos da Plataforma Lattes.

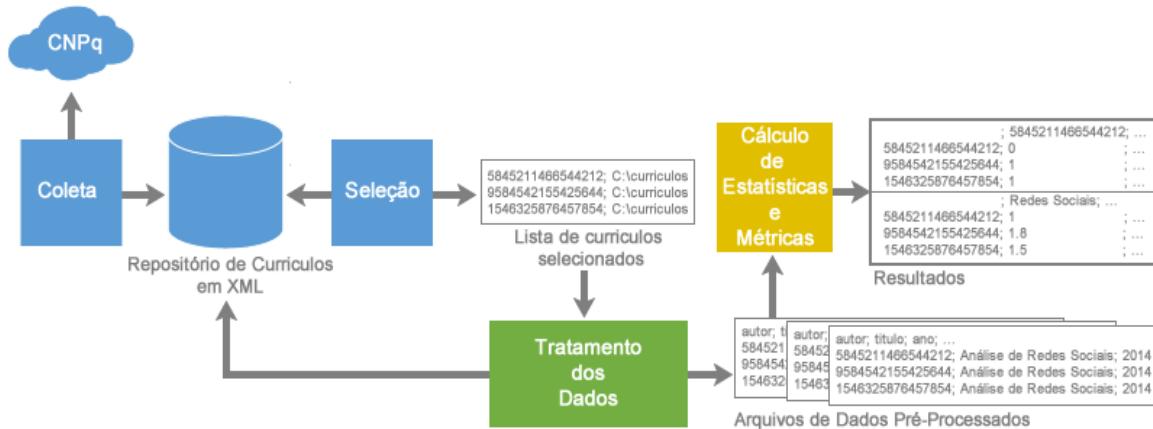


Figura 1: Visão geral do LattesDataXplorer.

A Figura 2 apresenta os componentes do módulo de coleta e extração de dados do LattesDataXplorer. O processo de coleta e extração dos dados curriculares da Plataforma Lattes envolve três etapas que são realizadas por meio de três componentes específicos que, para minimizar o custo computacional envolvido, executam respectivamente as seguintes funções: 1) extração de URLs, que visa obter os códigos de identificação de todos os currículos cadastrados na plataforma, possibilitando assim acessar individualmente cada um deles; 2) extração de Ids e Datas de Atualização, que visa extrair de cada currículo o seu identificador individual e a data de sua última atualização; e 3) coleta dos currículos, que visa coletar e armazenar em um repositório local os currículos cuja data de atualização na Plataforma Lattes seja divergente da data de atualização armazenada localmente ou que ainda não tenham sido coletados.

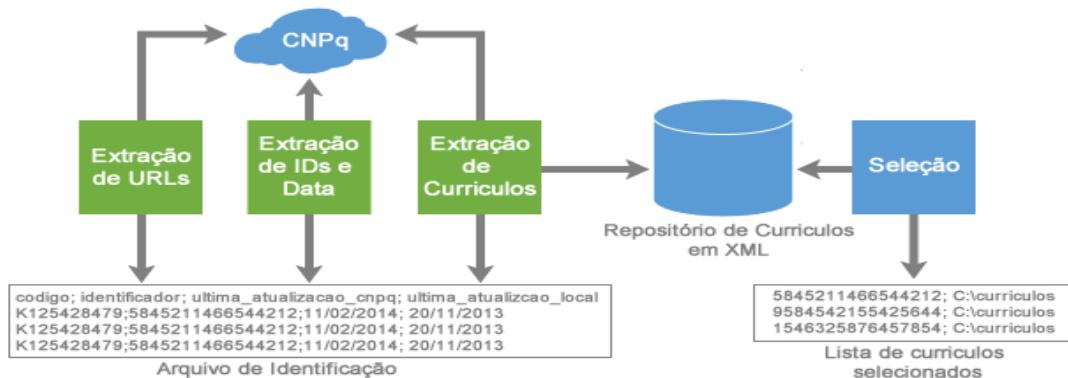


Figura 2: Processo de extração e seleção de dados.

Essas etapas são necessárias para manter o repositório local sempre atualizado sem que se faça necessário coletar novamente todos os currículos a cada nova extração, possibilitando assim a realização de análises com dados sempre atualizados. Além disso, é importante ressaltar que os currículos atualizados podem não só conter novos dados, como também ter dados já registrados alterados ou excluídos, o que torna o processo de atualização de campos específicos uma tarefa complexa. Com a estratégia adotada, todo currículo atualizado é substituído pela sua versão mais recente, o que ameniza consideravelmente o processo de atualização do repositório local de currículos.

O arquivo de identificação é a base para a extração de dados dos currículos. Toda a vez que por algum motivo seja necessário atualizar o repositório local de currículos, a primeira etapa do processo de extração é executada, resultando na extração de todos os códigos cadastrados na Plataforma Lattes. Códigos já registrados no arquivo de identificação são ignorados por corresponderem a currículos já incluídos no repositório local, enquanto que novos códigos são inseridos ao final do arquivo, já que representam novos currículos que ainda não foram coletados.

Posteriormente, com o uso dos códigos de identificação, são acessados os cabeçalhos de cada um dos currículos e extraídos os respectivos códigos identificadores e as datas de atualização junto à Plataforma Lattes, tanto para currículos já extraídos como para os novos currículos, atualizando o arquivo de identificação a cada nova extração. O acesso ao cabeçalho possibilita maior rapidez a todo o processo, agilizando de forma significativa a extração de dados, já que não se faz necessário esperar que todo o currículo seja gerado.

Finalmente, ocorre a extração de dados dos currículos. Inicialmente, o extrator verifica no arquivo de identificação se existem currículos cuja data de atualização na Plataforma Lattes é divergente da data de atualização local. Em caso afirmativo, esses currículos são coletados, substituindo as atuais versões armazenadas no repositório local e tendo as suas respectivas datas de atualização alteradas no arquivo local de identificação. Em seguida, são coletados os novos currículos cadastrados na Plataforma Lattes, cujos códigos de identificação foram inseridos ao final do arquivo local de identificação. Diante disso, os dados desses currículos são extraídos pela primeira vez e sua data de atualização local é registrada no arquivo de identificação. Todo esse processo possibilita manter um repositório atualizado com baixo custo computacional, já que um percentual considerável de currículos não é atualizado frequentemente.

Com todos os currículos armazenados localmente em formato XML, é possível manipular os dados com mais flexibilidade, permitindo explorar todo o potencial que os dados curriculares da Plataforma Lattes oferecem. Para isso é utilizado o módulo de seleção de currículos, que permite a criação de subgrupos de currículos por meio de consultas expressas em XPath.

Para a geração dos conjuntos de dados que compõem a coleção LattesDoctoralDataset disponibilizada neste trabalho, foram selecionados todos os indivíduos com doutorado concluído informado em seus currículos. Foi possível identificar que alguns indivíduos, mesmo com pós-doutorado concluído, não haviam inserido qualquer informação sobre seus cursos de doutorado, mas mesmo assim foram incluídos na coleção considerada.

4. Tratamento dos Dados

Na etapa de Tratamento dos Dados, o componente é responsável por analisar os currículos dos grupos selecionados e gerar um conjunto de dados para análises. O componente utiliza a lista de currículos selecionados, gerada para um determinado grupo, o que possibilita identificar quais são os currículos que vão ser analisados. A lista possui o identificador e o local de armazenamento do referido currículo e, diante desta indicação, apenas os currículos presentes na lista são considerados. O componente é responsável por analisar cada um dos currículos, de tal forma que sejam produzidos arquivos de dados pré-processados que vão possibilitar a realização de análises

bibliométricas da produção científica, análises do processo de formação e orientação do conjunto, como também da colaboração científica dos indivíduos que compõem os grupos em análise (Figura 3).

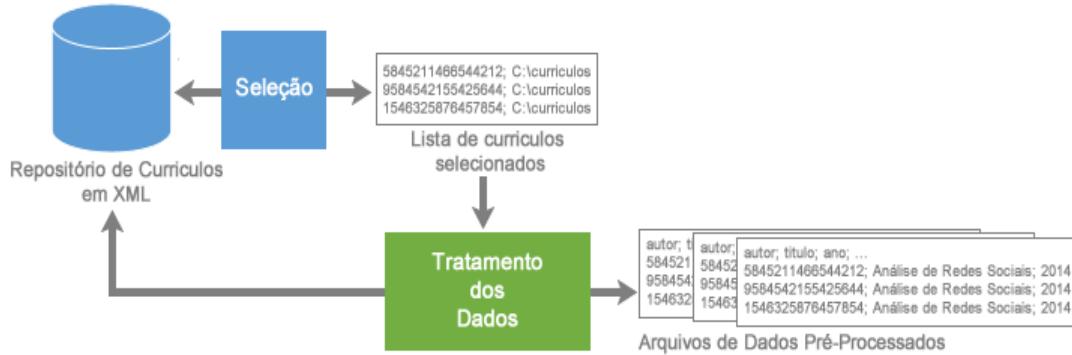


Figura 3: Processo de tratamento dos dados.

O maior desafio no tratamento dos dados coletados dos currículos da Plataforma Lattes está relacionado com a maneira com que cada indivíduo preenche os seus dados no currículo: entradas são realizadas manualmente e, portanto, não é uma situação incomum dois indivíduos cadastrarem o mesmo trabalho com informações divergentes, como o título do trabalho ou, até mesmo, nomes de colaboradores incompletos (BOAVENTURA et al., 2014).

É importante destacar que este trabalho não possui ênfase na desambiguação de autores, sendo esta tarefa foco de diversos outros trabalhos que tentam tratar a melhor forma de desambiguar nome de autores (FERREIRA; GONÇALVES; LAENDER, 2012). Neste trabalho, são considerados os identificadores de cada autor e não o nome dos autores para o processo de identificação de colaborações. Ou seja, cada indivíduo é referenciado pelo seu identificador único na Plataforma Lattes. O método de identificação utilizado neste trabalho para a caracterização das colaborações científicas dos doutores pode ser visualizado em Dias e Moita (2015).

Além da identificação das colaborações, como resultados da etapa de tratamento dos dados, são produzidos arquivos de dados pré-processados que contêm todas as informações necessárias para o cálculo de diversas métricas bibliométricas e baseadas em análise de redes sociais. O cálculo das estatísticas e métricas utilizando esses arquivos é facilitado, já que eles sumarizam todos os dados contidos nos currículos.

5. Descrição dos Conjuntos de Dados

Os dados utilizados para a seleção do conjunto de doutores foram coletados em junho de 2017 e correspondem a 5.251.540 currículos da Plataforma Lattes. Esses dados foram coletados conforme descrito na seção anterior. Após a coleta de todos os currículos e objetivando mostrar a facilidade do componente de seleção, foram selecionados todos os indivíduos com doutorado ou pós-doutorado concluído, resultando em um conjunto de 265.187 currículos, possibilitando realizar uma caracterização geral dos dados contidos nos currículos dos indivíduos com o maior nível de formação acadêmica. Além disso, em geral esse conjunto de currículos possui data de atualização recente e praticamente em sua totalidade possui algum tipo de publicação registrada. Este grupo de indivíduos que, em sua maioria, tem atuado em pesquisas, seja

em instituições de ensino seja em institutos de ciência e tecnologia, ainda é responsável pela formação dos alunos nos principais programas de pós-graduação stricto sensu no Brasil, e vários deles são reconhecidos por sua elevada produção científica. Com isso, ressalta-se que o conjunto de indivíduos considerados neste trabalho para a geração dos conjuntos de dados, compreende grande parte dos docentes de pós-graduação e os bolsistas de produtividade em pesquisa do CNPq.

Em Dias (2016) é apresentado que apesar do conjunto de indivíduos com doutorado concluído representar apenas 5,38% de todos os currículos cadastrados na Plataforma Lattes, estes são detentores de 64,67% dos artigos em anais de congressos e 74,51% dos artigos em periódicos registrados em todo o conjunto de currículos. Corroborando com a importância do conjunto considerado neste trabalho.

É importante destacar a diversidade dos dados que estão registrados neste conjunto de currículos, como artigos publicados em anais de congresso e periódico, apresentação de trabalhos, participação em eventos, nível de formação, orientações, dentre outros. Ressalta-se que um determinado trabalho pode estar inserido em currículos distintos, já que este pode ter sido realizado em colaboração. Logo, no repositório da Plataforma Lattes, um trabalho pode aparecer várias vezes, tendo em vista que ele pode ser registrado por cada um de seus autores. A Tabela 1 apresenta o quantitativo geral de todos os trabalhos registrados nos currículos dos doutores.

Tabela 1: Quantitativo dos dados dos currículos dos doutores em junho de 2017.

Tipo de Trabalho	Geral
Artigos em Anais de Congresso	9.051.680
Artigos em Periódico	4.660.430
Capítulos de Livro	1.054.844
Demais Trabalhos	445.894
Livros	397.934
Textos em Jornais e Revistas	858.789
Trabalhos Técnicos	1.342.416
Outras Produções Bibliográficas	625.621

A quantidade de dados registrada corrobora a importância da Plataforma Lattes, confirmando ser um dos maiores repositórios de dados científicos mundiais da atualidade (LANE, 2010), caracterizando-se como uma fonte extremamente rica para análise da produção científica brasileira. É possível observar a tendência de publicação de artigos em anais de congresso, seguida em menor número pela publicação de artigos em periódico.

Os estratos de dados estão disponíveis em formato *.csv*, em oito arquivos *.rar*, em que as colaborações científicas, que identificam pares de colaboradores estão divididas em dois arquivos. Além disso, um descritor contendo um detalhamento de todos os estratos com seus respectivos dados pode ser visualizado no arquivo *descritor.pdf*. Informações adicionais como por exemplo, datas de atualização dos estratos podem ser encontradas no arquivo *info.txt*.

Já a Tabela 2 apresenta a descrição dos estratos de dados disponibilizados. Como pode ser observado, os dados dos doutores estão organizados em conjuntos de dados que agrupam informações extraídas dos currículos e que possibilitam a realização de análises específicas.

Tabela 2: Descrição dos estratos de dados disponibilizados.

Estrato	Descrição
Formação Acadêmica	Dados sobre a formação acadêmica de cada doutor da graduação até o pós-doutorado.
Proficiência	Dados sobre nível de conhecimento (compreensão, fala, leitura e escrita) de cada doutor nos idiomas que domina.
Orientações	Dados sobre as orientações em andamento e concluídas de cada doutor nos diversos níveis de capacitação.
Produção Científica	Dados sobre a produção científica de cada doutor nos diversos tipos de veículo de publicação (periódicos, conferências, livros, etc.).
Atuação Profissional	Dados sumarizados sobre a atuação profissional de cada doutor.
Colaborações	Dados sobre as colaborações científicas de cada doutor com os demais indivíduos do mesmo grupo, tendo como base as publicações produzidas no quinquênio 2012 a 2016.

O primeiro estrato de dados (*Formação Acadêmica*) refere-se à sumarização de todos os cursos realizados por cada um dos doutores. No conjunto são apresentados os dados com o identificador de cada doutor e informações sobre cursos de graduação, especialização, mestrado, doutorado e pós-doutorado, incluindo ano de início, ano de conclusão e local de realização. Caso um doutor tenha realizado mais de um curso em um determinado nível de formação, o mais recente é considerado.

As análises deste conjunto de dados possibilitam compreender o processo de formação nos diversos níveis de capacitação dos doutores com currículos cadastrados na Plataforma Lattes. Estudos que visam compreender a duração média de cada nível de capacitação ou que levem em consideração as instituições em que os doutores se capacitaram poderiam propiciar informações inéditas que visam descrever o processo de formação destes indivíduos.

Já o estrato de dados *Proficiência* apresenta para cada identificador que corresponde a um determinado doutor, o nível de proficiência em outros idiomas. É apresentado o grau de conhecimento na compreensão, fala, leitura e escrita. Em cada uma destas habilidades podem ser registrados os seguintes níveis de conhecimento: pouco, razoavelmente e bem. Caso não seja feito nenhum registro de conhecimento sobre um determinado idioma, nenhuma informação sobre este idioma é apresentada.

Este conjunto de dados pode ser de extrema relevância para estudos que visam compreender o grau de conhecimento dos doutores de uma determinada área e realizar a correlação deste conhecimento com a produção científica dela. Tendo em vista que quanto maior o conhecimento em outros idiomas, maior o potencial de leitura de trabalhos escritos em revistas internacionais, e ainda, maior a qualidade de escrita de trabalhos a serem enviados a eventos e periódicos que exigem um determinado idioma. Além disso, uma análise sobre as áreas que possuem maior conhecimento em um determinado idioma pode proporcionar mecanismos para políticas de incentivo à capacitação visando melhorar a visibilidade de seus trabalhos realizados.

O estrato de dados *Orientações* apresenta para cada um dos doutores a quantidade de orientações em andamento e concluídas nos seguintes níveis de capacitação: Outra Natureza, Iniciação Científica, Graduação, Especialização, Mestrado, Doutorado e Pós-Doutorado. Tais valores que consideram o quantitativo de orientações em cada um destes níveis não consideram os cursos ou instituições onde tais orientações foram realizadas, sendo basicamente o somatório de orientações em cada um dos níveis de capacitação.

Logo, com este conjunto de dados, estudos que visam analisar a quantidade de orientações, dado o tempo de carreira de cada doutor (podendo utilizar para isso seus dados de formação acadêmica) podem ser realizados. Além disso, estudos que visam realizar análises comparativas sobre o processo de orientação nas diversas áreas do conhecimento também podem ser realizados, apresentando informações inéditas sobre o processo de orientação pelos principais orientadores, tendo em vista que a maioria dos docentes de pós-graduação no Brasil estão presentes no conjunto analisado.

O estrato de dados *Produção Científica* caracteriza-se como um conjunto importante para análises da produção científica dos principais pesquisadores brasileiros. Este conjunto de dados apresenta o quantitativo de artigos em anais de congressos e periódicos, textos em jornais e revistas, apresentação de trabalho, produção de trabalhos técnicos, demais produção técnica, livros e capítulos de livros. O conjunto de doutores é responsável por grande parte da produção científica registrada nos currículos cadastrados na Plataforma Lattes. Logo, a análise deste conjunto, representa de forma consistente a produção geral registrada nos currículos de todos os indivíduos.

Análises que consideram a correlação entre tempo de carreira e produção científica, podendo utilizar o conjunto de dados de *Formação Acadêmica*, ou ainda, análises que consideram a correlação entre o número de orientações e a produção científica, podendo utilizar o conjunto de dados *Orientações*, podem apresentar informações importantes para a comunidade científica. Além disso, estudos que consideram as grandes áreas e áreas de pesquisa para análises comparativas da produção científica de seus doutores podem propiciar informações inéditas sobre como as pesquisas vêm sendo realizadas e divulgadas em cada área do conhecimento.

O estrato de dados *Informações Profissionais* apresenta um conjunto de informações sobre cada um dos doutores. Nele, além do identificador do doutor é também descrito sua grande área e área de atuação, informações sobre seu vínculo profissional e sua localização, tendo como base para isso seu endereço profissional, além da data de atualização de seu currículo. Para indivíduos com mais de uma grande área e área de atuação registrada em seu currículo, a primeira delas foi considerada para a caracterização do conjunto de dados.

Logo, com os dados profissionais dos doutores, diversas análises que consideram as instituições em que estes doutores estão vinculados podem ser realizadas, bem como, utilizar do endereço profissional para análises por regiões geográficas. Além disso, ele possibilita o agrupamento dos doutores por áreas e grandes áreas de atuação, bem como permite visualizar o período da última atualização dos currículos. Este conjunto de dados é de extrema importância para ser utilizado com os outros conjuntos, potencializando ainda mais as análises a serem realizadas.

Por fim, o último estrato de dados *Colaborações* descreve o conjunto de colaborações em trabalhos publicados pelos doutores no quinquênio de 2012 a 2016. Ele contém o identificador de um determinado doutor e de seus colaboradores, e ainda, a quantidade de trabalhos, que pode ser de natureza distinta, realizados em colaboração.

Com estes dados é possível caracterizar e analisar a rede de colaboração científica dos doutores brasileiros. Logo, diversos estudos baseados em análises de redes podem ser desenvolvidos. Aliado a outros conjuntos de dados, tais estudos podem ser potencializados.

Os estratos de dados podem ser obtidos aqui:
<https://github.com/thiagomagela/LattesDoctoralDataset>.

6. Considerações Finais

Considerando o grande interesse de diversos trabalhos recentes que visam analisar dados de publicações científicas, os conjuntos de dados disponibilizados neste trabalho, caracterizam-se como importante fonte de informações para diversos novos estudos em diferentes áreas. Por sumarizar dados específicos, como produção científica, formação acadêmica, orientações em andamento e concluídas, informações profissionais e trabalhos em colaboração, os conjuntos de dados descritos possibilitam diversos novos estudos com grande facilidade, tendo em vista a forma como os conjuntos de dados estão formatados.

Por fim, espera-se que com a disponibilização destes conjuntos de dados, pesquisadores de áreas distintas do conhecimento possam realizar estudos bibliométricos que visam apresentar informações relevantes para toda a comunidade. Corroborando com a importância do conjunto disponibilizado e a relevância dos dados curriculares da Plataforma Lattes.

7. Referencias

- Brito, A. G. C.; Quoniam, L.; Mena-Chalco, J. P. (2016) Exploração da Plataforma Lattes por assunto: proposta de metodologia. *TransInformação*, v. 28, n. 1, p. 77-86.
- Dias, T. M. R. (2016) Um Estudo da Produção Científica Brasileira a partir de Dados da Plataforma Lattes. Tese de Doutorado, Programa Pós-Graduação em Modelagem Matemática e Computacional, CEFET-MG.
- _____ ; Moita, G. F. (2015) A method for the identification of collaboration in large scientific databases. *Em Questão*, v. 21, p. 140-161.
- Digiampietri, L. A.; Mugnaini, R.; Alves, C. (2013) Analysis of Participation in Supervised Production of Advisors: A Case Study in Computer Science. In: *Proceedings of the Brazilian Workshop on Social Network Analysis and Mining*, Maceió.
- Ding, Y. (2011) Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Informetrics*, v. 5, n. 1, p. 187-203.
- Ferraz, R. R. N.; Quoniam, L.; Maccari, E. A. (2014) The Use of Scriptlattes tool for extraction and on line availability of academic production from a departament of stricto sensu in management. In: *Proceedings of the International Conference on Information Systems and Technology Management*, São Paulo, p. 663-679.
- Ferreira, A. A.; Gonçalves, M. A.; Laender, A. H. F. (2012) A Brief Survey of Automatic Methods for Author Name Disambiguation. *SIGMOD Record*, v. 41, n. 2, p. 15-26.
- Furtado, C. A. et al. (2015) A Spatiotemporal Analysis of Brazilian Science from the Perspective of Researchers' Career Trajectories. *PLOS ONE*, v. 10, n. 10, p. e0141528.

- Laender, A. H. F. et al. (2008) Assessing the research and education quality of the top Brazilian Computer Science graduate programs. SIGCSE Bulletin, v. 40, n. 2, p. 135-145.
- Lane, J. (2010) Let's make science metrics more scientific. Nature, v. 464, n. 7288, p. 488-489.
- Mena-Chalco, J. P. et al. (2014) Brazilian bibliometric coauthorship networks. JASIST 65(7): 1424-1445.
- Mugnaini, R.; Leite, P.; Leta, J. (2011) Fontes de informação para análise de internacionalização da produção científica brasileira. PontodeAcesso, v. 5, n. 3, p. 87-102.
- _____. et al. (2014) Comunicação científica no Brasil (1998-2012): indexação, crescimento, fluxo e dispersão. Transinformação, v. 26, n. 3, p. 239-252.
- Oliveira, E. A. et al. (2012). Comparison of Brazilian researchers in clinical medicine: are criteria for ranking well-adjusted? Scientometrics, 90(2), 429–443.
- Perez-Cervantes, E.; Mena-Chalco, J. P.; Cesar-Junior, R. M. (2012) Towards a quantitative academic internationalization assessment of Brazilian research groups. eScience 2012: 1-8.
- Revoredo, K. et al. (2012) Mining scientific literature for analysis of collaboration in research communities. In: Proceedings of the Brazilian Workshop on Social Network Analysis and Mining, Curitiba.
- Roos, D. H. et al. (2014) Brazilian scientific production in areas of biological sciences: a comparative study on the modalities of full doctorate in Brazil or abroad. Scientometrics. v. 98, n. 1, p. 415-427.
- Sidone, O. J. G.; Haddad, E. A.; Mena-Chalco, J. P. (2017) Scholarly publication and collaboration in Brazil: The role of geography. JASIST 68(1): 243-258.
- Silva, T. H. P. et al. (2016) The Impact of Academic Mobility on the Quality of Graduate Programs. D-Lib Magazine 22(9/10).
- Ströele, V.; Zimbrão, G.; Souza, J. M. (2012) Análise de redes sociais científicas: modelagem multi-relacional. In: Proceedings of the Brazilian Workshop on Social Network Analysis and Mining, Curitiba.