ANÁLISE DA ATUALIZAÇÃO DOS CURRÍCULOS LATTES

Uma análise macro sobre as datas das últimas atualizações dos currículos Lattes

Luciano Digiampietri, Rogério Mugnaini, Jesús Mena-Chalco, Karina Delgado, José Pérez-Alcázar

Eixo temático: Bases de Dados

Modalidade: Apresentação oral

1 INTRODUÇÃO

Ao longo dos últimos anos, estudos bibliométricos e cientométricos têm utilizado cada vez mais grandes volumes de informação. No Brasil existe uma base ímpar de dados bibliométricos que é a Plataforma Lattes. Nesta plataforma há mais de 3,2 milhões de registros cadastrados (http://www.cnpq.br/web/portal-lattes/dados-e-estatisticas) contendo informações sobre formação, áreas de atuação, projetos de pesquisa, produções (bibliográficas, técnicas e artísticas), participação em eventos, bancas, orientações dentre outras.

Em especial, na última década diferentes pesquisas acadêmicas têm considerado a Plataforma Lattes como principal fonte de dados. O espectro de trabalhos varia do nível macro ao micro. Por exemplo, trabalhos que visam a apresentar um panorama de toda a produção científica nacional (LEITE *et al*, 2011), incluindo a rede social acadêmia de coautorias (MENA-CHALCO *et al*, 2014), descrições sobre as informações presentes na plataforma (DIGIAMPIETRI *et al*, 2012), ferramentas para a extração e/ou mineração dos dados da plataforma (ALVES *et al*, 2011; MENA-CHALCO e CESAR JUNIOR, 2009), estudos sobre grupos específicos de pesquisadores (DIGIAMPIETRI *et al*, 2012b; ARRUDA et al, 2009; WAINER, J. e VIERA, P., 2013; COSTA, B.G. *et al*, 2013), predição de relacionamentos (LIBEN-NOWELL e KLEINBERG, 2003), entre outros.

Apesar da abundância e relevância das informações contidas na Plataforma Lattes existem diversas características que devem ser consideradas para sua utilização (CAÑIBANO e BOZEMAN, 2009). Entre elas: o fato das informações não serem validadas (i.e., os dados são inseridos pelos possuidores dos currículos sem posterior validação); muitos campos são preenchidos manualmente, acarretando em erros de digitação e/ou falta de padronização; a

frequência de atualização dos dados depende dos possuidores dos currículos e varia bastante; muitos campos são opcionais, limitando alguns tipos de análise. Este alto grau de liberdade no registro das informações curriculares é pouco estudado pelos pares. No campo da Ciência da Informação, SILVA e SMIT (2009) alertam para o comprometimento da consistência dos dados para recuperação da informação, o que acaba limitando o uso desta fonte curricular tão abrangente para uma análise mais profunda da produção científica nacional.

Em particular, a frequência de atualização dos dados é de extrema importância para trabalhos que pretendam identificar ou prever tendências na produção nacional, predição de relacionamentos ou de citações em redes acadêmicas, e recomendação de trabalhos científicos. Este artigo tem por objetivo analisar a atualização dos currículos vitae (CVs) da Plataforma Lattes, segmentando estes currículos pelas áreas de conhecimento de seus possuidores e também pela formação acadêmica máxima dos mesmos.

2 METODOLOGIA

A metodologia está estruturada em duas partes: obtenção dos dados e tratamento dos dados (que inclui a organização e o processamento automático).

2.1 Obtenção dos Dados

Para este trabalho, foram obtidos os arquivos XML de 3.187.710 CVs da Plataforma Lattes durante o mês de julho de 2013. Para a obtenção destes currículos as seguintes atividades foram executadas: (a) foi feita uma consulta no site de buscas por CVs da Plataforma Lattes de forma a solicitar a lista de todos os currículos cadastrados; esta consulta retornou múltiplas páginas Web de resposta; (b) cada uma das páginas de resposta foi copiada e os identificadores numéricos dos CVs (IDs Lattes) foram extraídos através de um *script* de computador; (c) com o identificador de cada currículo foi possível baixar cada CV Lattes em formato XML.

2.2 Tratamento dos dados

O tratamento dos dados foi dividido em três etapas: separação das informações de interesse; divisão das informações nos grupos de interesse; e cálculo de métricas.

2.2.1 Separação das informações de interesse

Neste trabalho três tipos de informação dos CVs Lattes foram consideradas: (a) a data da última atualização (oriunda das informações gerais de cada CV); (b) as grandes áreas de atuação; e (c) as formações acadêmicas/titulações. Foi desenvolvido um *script* para extrair as informações de interesse de cada currículo.

2.2.2 Divisão das Informações nos Grupos de Interesse

Além da análise conjunto da atualização de todos os CVs da Plataforma Lattes, também foram identificados grupos, segundo as grandes áreas de atuação e a maior formação presente em cada currículo. Cada CV pode manter registro de zero ou mais grandes-áreas de atuação, permitindo que determinado CV faça parte de mais de um grupo. Segundo a CAPES, São nove as grandes-áreas do conhecimento: Ciências Agrárias; Ciências Biológicas; Ciências da Saúde; Ciências Exatas e da Terra; Ciências Humanas; Ciências Sociais Aplicadas; Engenharias; Linguística, Letras e Artes; e Outros/Multidisciplinar. Também foi criado um grupo adicional formado por CVs que não continham esta informação. Quanto às formações acadêmicas/titulações foram definidas 12 opções, das quais este estudo considerou seis: *Ensino Fundamental/Primeiro Grau*, *Ensino Médio/Segundo Grau*, *Curso Técnico Profissionalizante*, *Graduação*, *Mestrado/Mestrado Profissionalizante* e *Doutorado*. Além destes, foram criados dois grupos adicionais: um contendo os CVs que não apresentavam nenhuma destas formações/titulações e outro contendo os possuidores de bolsa de produtividade em pesquisa do CNPq (os CVs destes pesquisadores compõem tanto na conta do grupo *Doutorado* quanto do grupo *Produtividade*). Para a criação destes grupos só foram consideradas as formações/titulações máximas entre as seis destacadas (independente de estarem concluídas ou em andamento). Já que os possuidores de bolsa produtividade frequentemente atualizam seus currículos (ao menos) logo antes do pedido da bolsa optou-se por selecionar os CVs daqueles que possuiam bolsa produtividade em 2010.

2.2.3 Cálculo de Métricas

Tanto para o conjunto de todos os CVs quanto para cada uma das divisões feitas foram calculadas as seguintes métricas: quantidade média de dias desde a última atualização; valor da mediana de dias desde a última atualização; distribuição das atualizações por meses; e quantidade de dias desde a última atualização para cada decil (cada decil representa 1/10 do total do conjunto de dados).

3 RESULTADOS

Os resultados são apresentados em três subseções. Na primeira, são apresentados alguns dados gerais envolvendo todos os CVs analisados. Em seguida apresenta-se uma análise da atualização dos CVs considerando-se as grandes-áreas de atuação. Por fim, é descrita uma análise considerando-se as formações/titulações dos currículos registrados no conjunto de dados.

3.1 Dados Gerais

Dos 3.187.710 de CVs examinados, baixados da Plataforma Lattes em julho de 2013, as datas de atualização variaram de 22/08/1997 (CV com data de atualização mais antiga) a 30/07/2013 (currículo que foi atualizado no dia que foi baixado), este intervalo de datas corresponde a pouco mais de 16 anos (194 meses). Na média os CVs foram atualizados 860 dias antes de terem sido baixados (correspondendo, na média, a 16/03/2011), já a mediana é de 486 dias (correspondendo a 24/03/2012). É possível observar que, considerando esse conjunto de dados, mais da metade dos currículos foram atualizados pela última vez há mais de um ano.

A Figura 1 indica a porcentagem de currículos que foram atualizados dentro de períodos mensais. Só são exibidos os 36 primeiros meses dentro dos quais 70,96% dos currículos foram atualizados. 5,4% dos CVs foram atualizados há, no máximo, um mês da data que foram baixados. A curva cinza apresenta os valores acumulados. É possível observar que 41,90% dos CVs foram atualizados nos últimos 12 meses e 60,16% nos últimos 24 meses.

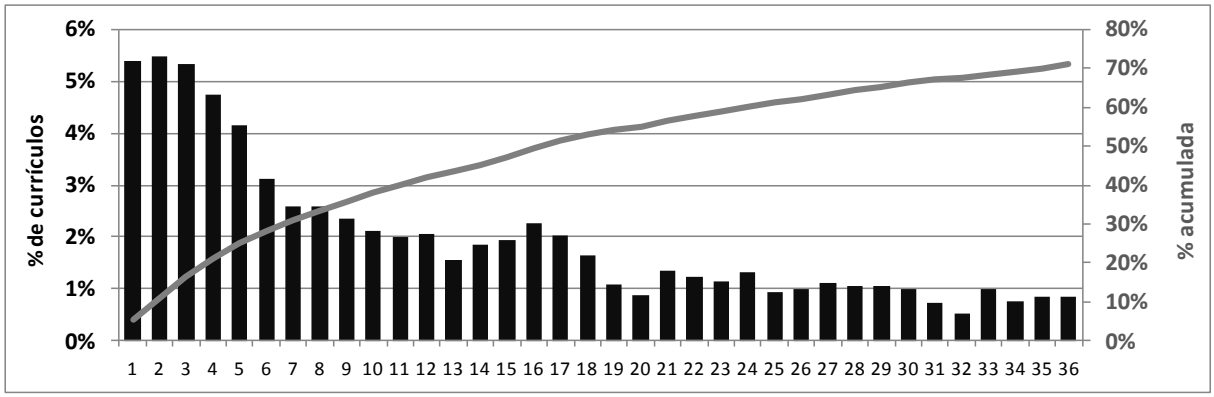

Figura 1 – Porcentagem de CVs atualizados ao longo do tempo

3.2 Grandes Áreas

A Tabela 1 mostra os decis da variável dias desde a última atualização. Por exemplo, ao se analisar os 10% CVs mais atualizados em Ciências Biológicas estaremos olhando para CVs atualizados até 28 dias antes da data em que os CVs foram baixados. Ao se analisar 50% de todos os currículos que declararam esta mesma grande área estaremos olhando para CVs atualizados há até 247 dias antes da data em que foram baixados. A coluna *Total* contém o número total de currículos pertencente ao respectivo grupo.

Tabela 1 – Número de dias desde a última atualização dos CVs por grande-área

Grande Área	Total CVs	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Sem Grande Área	1.006.088	124	328	637	978	1173	1425	1718	2121	2631	5782
Ciências Agrarias	141.730	31	63	103	168	278	441	661	1043	2066	5779
Ciências Biológicas	175.384	28	59	97	151	247	398	613	1001	1944	5783
Ciências da Saúde	459.825	45	95	158	257	397	530	741	1093	1921	5780
Ciências Exatas e da Terra	266.693	40	78	130	218	344	501	747	1187	2251	5796
Ciências Humanas	417.334	37	78	130	201	316	478	670	975	1783	5779
Ciências Sociais Aplicadas	439.330	47	99	168	275	418	561	776	1146	2009	5795
Engenharias	192.495	44	90	151	258	403	566	851	1450	2633	5814
Linguística, Letras e Artes	157.597	40	83	137	218	333	487	684	1012	1834	5599
Outros/Multidisciplinar	337.017	49	86	125	179	242	330	420	469	708	5780
Mundo Lattes	3.187.710	55	112	196	333	486	715	1045	1523	2293	5814

Na Tabela 1 é possível observar que os CVs da área de Ciências Biológicas são aqueles atualizados mais recentemente (exceto pela última coluna). Já os currículos que não têm nenhuma área de atuação declarada são aqueles mais desatualizados (exceto novamente pela última coluna). A última coluna da tabela representa a quantidade de dias passados desde a última atualização do currículo com data de atualização mais antiga de cada grupo. Por isto, esta informação não é muito representativa.

3.2 Formações Acadêmicas/Titulações

A Tabela 2 está organizada da mesma maneira que a Tabela 1, porém os CVs foram agrupados de acordo com sua maior formação acadêmica. É interessante notar que mais de dois terços dos CVs tem a graduação (completa ou em andamento) como maior formação. Dentre os grupos, aqueles de maior formação são atualizados mais recentemente para a grande maioria das colunas. É possível observar, por exemplo, que ao se analisar metade dos

CVs dos doutores (ou doutorandos) da plataforma estamos lidando com a última atualização realizada em 114 dias. Já para os 20% dos currículos mais atualizados dos graduados (ou graduandos) a última atualização foi realizada há 141 dias.

Tabela 2 – Número de dias desde a última atualização dos CVs por formação.

Maior Formação/Titulação	**Total CVs**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
Sem Formação Declarada	108.758	621	793	1202	2021	2290	2442	2584	2804	3202	5814
Ensino Fundamental Primeiro Grau	15.508	123	176	434	561	697	783	841	921	1156	4157
Ensino Médio Segundo Grau	121.554	112	196	368	485	621	697	781	849	1114	4284
Curso Técnico Profissionalizante	30.940	205	552	678	781	878	1044	1329	1762	2604	5786
Graduação	2.203.076	70	141	247	384	523	791	1138	1561	2239	5796
Mestrado	428.264	34	74	124	187	299	478	747	1195	2093	5787
Doutorado	279.610	19	34	55	79	114	167	281	533	1314	5780
Produtividade	13.787	10	18	22	32	42	58	78	111	134	2534

Já para os possuidores de bolsa produtividade, 80% dos currículos atualizados mais recentemente foram atualizados há menos de quatro meses da data em que os currículos foram baixados (111 dias). Por outro lado, os CVs que não possuem nenhuma formação declarada são aqueles que foram atualizados há mais tempo.

4 CONSIDERAÇÕES FINAIS

Alguns dos principais fatores considerados na análise de dados são: completude, corretude e atualização dos dados (CAÑIBANO e BOZEMAN, 2009). Apesar dos dados contidos na Plataforma Lattes serem de grande valia para pesquisas bibliométricas e cientométricas nenhum destes fatores é garantido, pois os três dependem dos usuários que estão cadastrando seus currículos e do recorte utilizado para selecionar os CVs Lattes. Mesmo com estas limitações, a quantidade e a riqueza da informação disponível são tão grandes que justificam sua ampla utilização.

Neste trabalho analisamos a data da última atualização dos CVs considerando-se as diferentes áreas de atuação e maior formação acadêmica. Com a análise apresentada pretende-se deixar mais claro quais grupos de CVs mantêm atualizados (visão macro) de forma a auxiliar os trabalhos futuros na seleção dos grupos de currículos a serem utilizados e/ou dos recortes a serem feitos.

REFERÊNCIAS

ARRUDA, D. *et al*. Brazilian computer science research: Gender and regional distributions. *Scientometrics*, v. 79, p. 651-665, 2009.

ALVES, A. D. *et al*. LattesMiner: a multilingual DSL for information extraction from lattes platform. *In Proceedings of SPLASH'11, SPLASH '11 Workshops*, New York, NY, USA, p. 85–92, 2011.

CAÑIBANO C., e BOZEMAN B. Curriculum vitae method in science policy and research evaluation: the state-of-the-art. *Research Evaluation*, v. 18, n. 2, p.86-94, 2009.

COSTA, B.G. *et al*. Scientific collaboration in biotechnology: the case of the northeast region in Brazil. *Scientometrics*, v. 95, p. 571-592, 2013.

DIGIAMPIETRI, L. A. *et al*. Minerando e Caracterizando Dados de Currículos Lattes. *In Proceedings of the Brazilian Workshop on Social Network Analysis and Mining (BraSNAM - CSBC 2012)*, 2012

_____ . Dinâmica das Relações de Coautoria nos Programas de Pós-Graduação em Computação no Brasil. *In Proceedings of the Brazilian Workshop on Social Network Analysis and Mining (BraSNAM - CSBC 2012)*, 2012.

LEITE, P. *et al*. A new indicator for international visibility: exploring Brazilian scientific community. *Scientometrics*, v. 88, p. 311-319, 2011.

LIBEN-NOWELL, D. e KLEINBERG, J. The link prediction problem for social networks. *In Proceedings of the twelfth international conference on Information and knowledge management - CIKM '03*, New York, New York, USA, p. 556, 2003.

MENA-CHALCO, J. P. *et al*. Brazilian bibliometric coauthorship networks. *Journal of the Association for Information Science and Technology*, 2014.

MENA-CHALCO, J. P. e 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.

SILVA, F. e SMIT, J. W. Organização da informação em sistemas eletrônicos abertos de Informação Científica & Tecnológica: análise da Plataforma Lattes. Perspect. ciênc. inf., Belo Horizonte , v. 14, n. 1, 2009 .

WAINER, J. e VIERA, P. Correlations between bibliometrics and peer evaluation for all disciplines: the evaluation of Brazilian scientists. *Scientometrics*, v. 96, p. 395-410, 2013.