e-LATTES: UM NOVO ARCABOUÇO EM LINGUAGEM R PARA ANÁLISE DO CURRÍCULO LATTES

Ricardo Barros Sampaio
Fundação Oswaldo Cruz, Rio de Janeiro, RJ, Brasil
rsampaio.br@gmail.com

Antonio de Abreu Batista Junior
Universidade Federal do ABC, Santo André, SP, Brasil
antonio.batista@ufma.br

Jesús P. Mena-Chalco
Universidade Federal do ABC, Santo André, SP, Brasil
jesus.mena@ufabc.edu.br

1 INTRODUÇÃO

A Plataforma Lattes é um sistema de informação curricular, disponibilizado pelo CNPq, que permite o registro curricular dos pesquisadores brasileiros. Atualmente, a plataforma conta com mais de cinco milhões de currículos cadastrados[1]. Por sua riqueza de informações e sua crescente confiabilidade e abrangência, o Currículo Lattes (CV Lattes) se tornou elemento indispensável e compulsório à análise de mérito e competência dos pleitos de financiamentos na área de ciência e tecnologia.

No entanto, os CVs Lattes foram projetados para mostrar informação pública e individual de cada usuário. Muitas vezes, realizar uma compilação ou sumarização de produções bibliográficas para um grupo de usuários cadastrados de médio ou grande porte (e.g. grupo de professores, departamento de pós-graduação) requer um grande esforço mecânico que muitas vezes é suscetível a falhas (MENA-CHALCO; CESAR JUNIOR, 2009).

A fim de superar esta limitação, um número de ferramentas de mineração de dados tem sido propostas a fim de prover a extração e visuali-

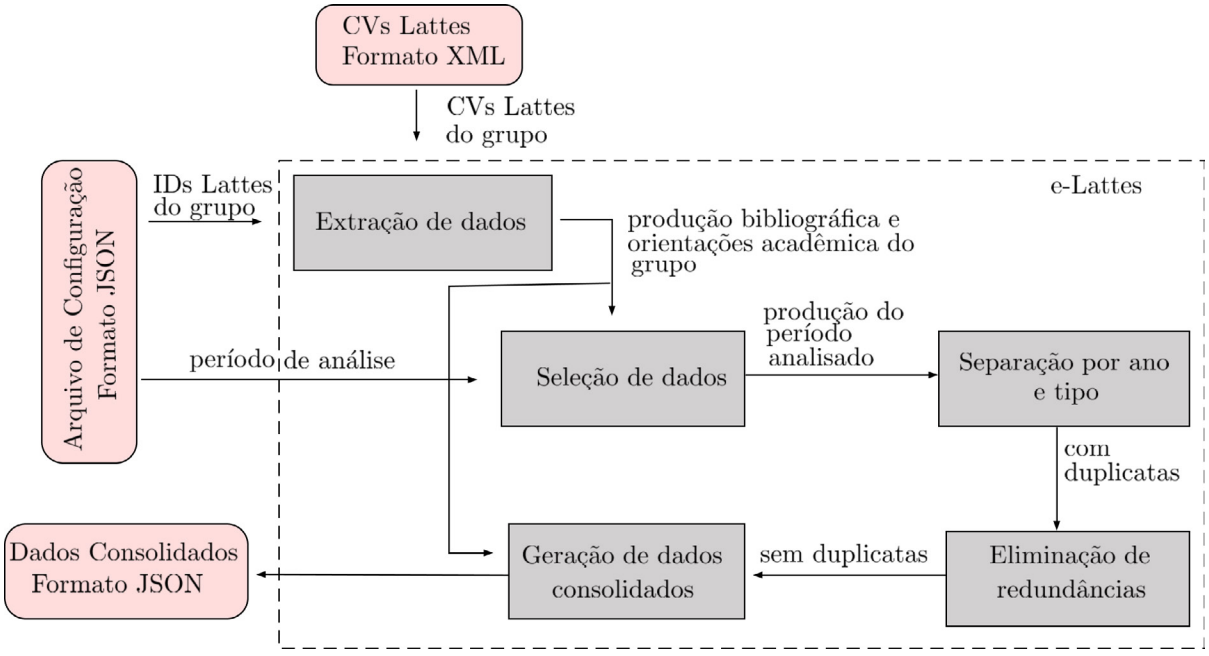
1 A Plataforma Lattes está disponível em: <http://lattes.cnpq.br>. Acesso em: fev. 2018.

zação de conhecimento a partir de CVs cadastrados na Plataforma Lattes (ALVES; YANASSE; SOMA, 2011; MENA-CHALCO; CESAR JUNIOR, 2009). Entretanto, aquelas disponíveis na linguagem de programação R (CAMPELO, 2017; PERLIN, 2017), apesar de boas ferramentas, elas são incompletas, lidando somente com publicações científicas. A linguagem R possui ótimo suporte para análises estatística e gráficos o que torna a linguagem potencialmente ideal para tarefas de mineração de dados. Aqui é apresentado um novo ferramental computacional, em forma de pacote R, para apoiar a exploração de dados curriculares da plataforma Lattes. Nós descrevemos os elementos inovadores que compõe a ferramenta e apresentamos como estudo de caso a sua aplicação considerando os pesquisadores da Fundação Oswaldo Cruz (FIOCRUZ).

2 SOLUÇÃO TECNOLÓGICA

O e-Lattes é composto por cinco módulos. A Figura 1 ilustra os relacionamentos entre eles. Cada módulo é responsável por uma funcionalidade do pacote R e produz resultados intermediários que são insumos para outros módulos.

FIGURA 1 - MÓDULOS IMPLEMENTADOS NO ARCABOUÇO E-LATTES.



Fonte: Elaborado pelos autores, 2018.

O módulo de Extração de dados tem como funcionalidade a extração de dados dos CVs dos pesquisadores, em formato XML, e a sua representação no computador. O módulo extrai os dados gerais (e.g., nome, endereço profissional, resumo), além de toda a produção bibliográfica e acadêmica dos CVs. A estrutura de dados contendo estas informações é utilizada no módulo de Seleção de dados, mas pode ser utilizada por outros módulos. O módulo de Seleção de dados permite filtrar, por período, as publicações científicas e orientações acadêmicas. A separação por ano e tipo (e.g., periódicos, eventos, orientações concluídas) é feito pelo módulo de Separação. O módulo de Eliminação de redundâncias usa o algoritmo de casamento de strings (*Levenshtein*) para eliminação de duplicatas. Por exemplo, artigos científicos com títulos aproximados que são encontrados nos CVs analisados são considerados apenas diferentes instâncias de um artigo. Por fim, as publicações e orientações únicas, separadas por ano e tipo, servem como entrada para o módulo de Geração de dados consolidados. Atualmente, com esses dados pode-se gerar diversos indicadores bibliométricos, entre eles o perfil acadêmico dos pesquisadores, a interseção de áreas, a senioridade acadêmica, a produção bibliográfica e orientações acadêmicas do grupo, por exemplo.

O e-Lattes está sendo projetado para tratar arquivos em formato JSON tanto para a entrada de parâmetros de execução quanto para a saída de dados. Esse formato permite uma ampla utilização desses dados em diferentes plataformas de programação atual.

3 EXPERIMENTO DE USO DA FERRAMENTA

O universo de CVs Lattes, objeto do presente estudo, foi o de funcionários concursados da FIOCRUZ, conforme dados disponibilizados pela própria instituição. A Tabela 1 mostra indicadores bibliométricos selecionados do grupo FIOCRUZ e das cinco maiores unidades. Todos os processamentos foram realizados em um computador com sistema operacional Linux / Ubuntu com quatro unidades de processamento Pentium 4, frequência 3 GHz e 80GB de RAM.

TABELA 1 - INFORMAÇÕES PARA O GRUPO FIOCRUZ E CINCO MAIORES UNIDADES GERADAS PELO E-LATTES

Número total de CVs	Unidade	Número de publicações	Número de orientações	Tempo estimado de processamento computacional
4.454	FIOCRUZ	40.624	14.647	20 horas
678	IOC	12.322	3.477	40 minutos
671	ENSP	7.116	3.316	20 minutos
616	IFF	2.458	889	2 minutos
299	INI	2.624	737	4 minutos
228	BIO	415	120	1 minuto

Fonte: Elaborado pelos autores, 2018.

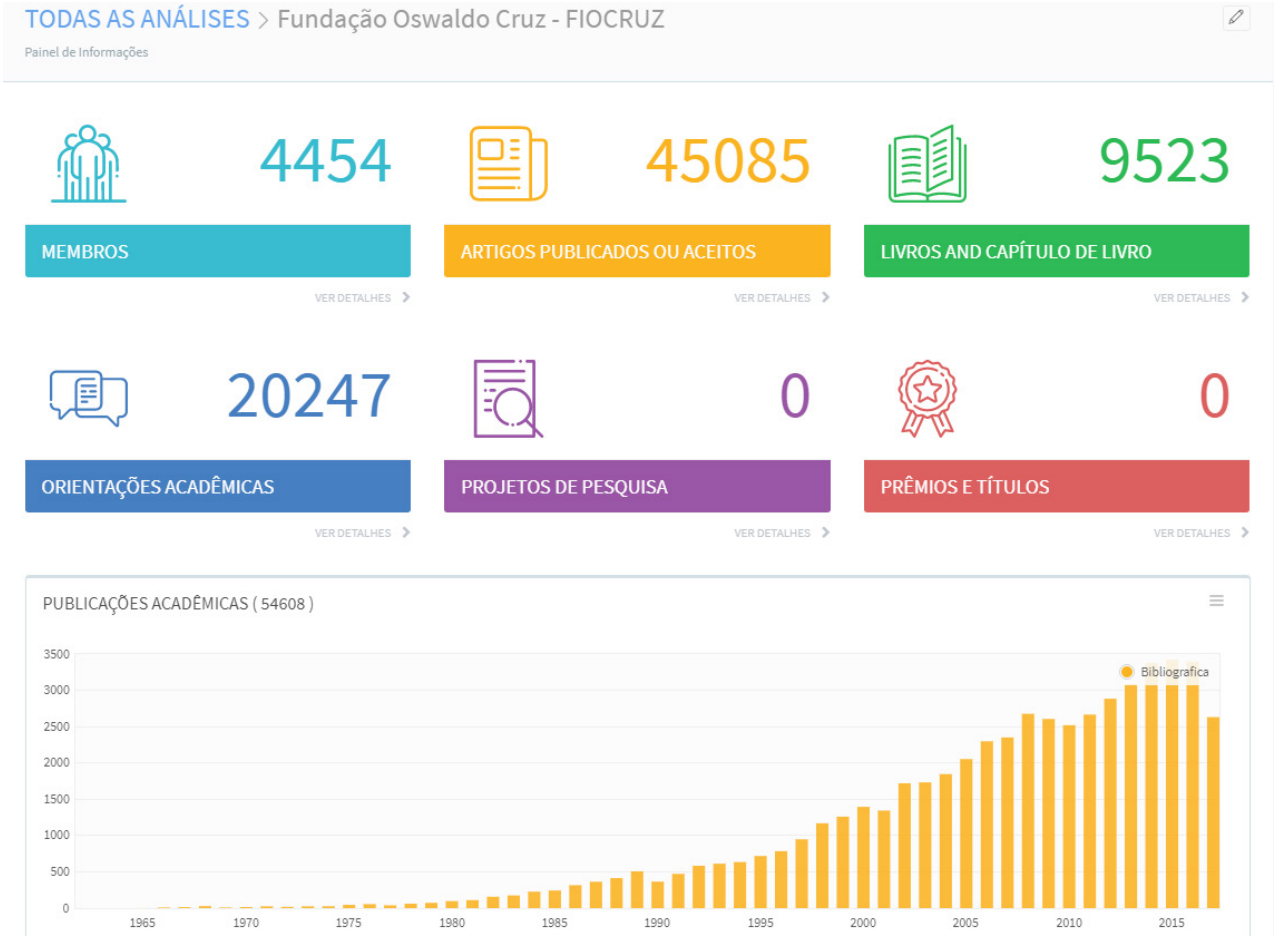
Os dados apresentados na Tabela 1 estão relacionados a toda a FIOCRUZ, que é composta por 22 unidades técnico-científicas, e mais cinco dessas unidades, sendo elas o Instituto Oswaldo Cruz (IOC), a Escola Nacional de Saúde Pública (ENSP), o Instituto Nacional Fernandes Figueira (IFF), o Instituto Nacional de Infectologia Evandro Chagas (INI) e o Instituto de Tecnologia em Imunobiológicos (Bio-Manguinhos ou BIO). A análise foi realizada para o período 1962-2017 incluindo a produção de todos os funcionários com CV Lattes calculando a sua produção a partir da data de entrada na instituição. As publicações contabilizadas correspondem a artigos em revistas científicas e as orientações concluídas no nível de mestrado e doutorado.

O tempo estimado de processamento para a FIOCRUZ foi de 20 horas. No entanto, esse tempo pode ser reduzido para menos da metade (7 horas) se calcularmos o mesmo universo mas com a data de início da análise em 2011. Para o IOC, maior unidade técnico científica da instituição, o tempo de processamento foi de 40 minutos e o de BIO menor instituição em número de publicações apresentada na Tabela 1, foi de 1 minuto.

4 CONSIDERAÇÕES FINAIS

O e-Lattes é uma ferramenta que está em desenvolvimento (http://elattes.com.br). A Figura 2 apresenta a tela com dados descritivos da análise da Fiocruz. Planejamos considerar nas próximas versões: (i) redes de coautoria científica, (ii) medidas de multidisciplinaridade de pesquisadores, e (iii) informações de geolocalização. A solução proposta e de acesso aos gestores e pesquisadores em geral está disponível em ambiente WEB de acesso aberto e de fácil manuseio dos CVs Lattes.

FIGURA 2 - TELA INICIAL DE ANÁLISE DO E-LATTES



Fonte: Elaborado pelos autores, 2018.

REFERÊNCIAS

ALVES, A.; YANASSE, H.; SOMA, N. Sucupira: a system for information extraction of the Lattes platform to identify academic social networks. In: IBERIAN CONFERENCE ON INFORMATION SYSTEMS AND TECHNOLOGIES, 6., 2011, Chaves. Proceedings... Chaves: CISTI, 2011. p. 1-6.

CAMPELO, F. ChocoLattes: Processing Data from Lattes CV Files, 2017. Disponível em: <https://CRAN.R-project.org/package=ChocoLattes>. Acesso em: 5 jan. 2018.

PERLIN, M. GetLattesData: Reading Bibliometric Data from Lattes Platform, 2017. Disponível em: <https://CRAN.R-project.org/package=GetLattesData>. Acesso em: 5 jan. 2018.

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**, Porto Alegre, v. 15, n. 4, p. 31-39, 2009.