Proxy customizado para acesso ao web service da Plataforma Lattes

Mesailde Souza de Oliveira Matias[1], **Roniberto Morato do Amaral**[1], **Paulo Matias**[1]

[1]Universidade Federal de São Carlos (UFSCar)
Rod. Washington Luís km 235 - SP-310
São Carlos – 13.565-905 – São Carlos – SP – Brazil

{mesailde,roniberto,matias}@ufscar.br

Abstract. *The brazilian federal educational institutions have the great challenge of managing their teaching, research and extension data, transforming them into open data, according to the current governmental precepts aimed at promoting transparency, and encouraging people's participation into the improvement of public services. In addition, institutions also demand structured data to generate indicators to guide their future actions and to contribute to their decision-making processes. The Lattes Platform now aggregates information from brazilian researchers and their international partners, and therefore has great potential as a source for institutional databases and for studies of their own research groups. In this context, this work describes the development of an intelligent access proxy for the Lattes Platform, which guarantees both the integrity of the Platform servers and the autonomy of the institution over the data of its researchers.*

Resumo. *As Intituições Federais de Ensino Superior (IFES) tem o grande desafio de gerenciar seus dados de Ensino, Pesquisa e Extensão e transformá-los em dados abertos, de acordo com os atuais preceitos governamentais voltados para a promoção da transparência, e ao incentivo à participação cidadã na melhoria dos serviços públicos. Para além da transparência, as IFES também necessitam de dados estruturados para gerar indicadores que balizem suas ações futuras e possam contribuir nas tomadas de decisões de seus gestores. A Plataforma Lattes agrega hoje informações de pesquisadores do Brasil e de seus parceiros no exterior e possui grande potencial como base da dados para as IFES utilizarem no povoamento de seus sistemas e também para a utilização em estudos de seus grupos de pesquisa. Neste contexto, este trabalho descreve o desenvolvimento de um proxy inteligente de acesso à Plataforma Lattes, de modo a garantir, ao mesmo tempo, a integridade dos servidores da Plataforma, e a autonomia das IFES sobre os dados de seus pesquisadores.*

1. Introdução

A administração pública brasileira, por força de uma sociedade cada vez mais participativa e ativa no que diz respeito a ficalização dos gastos públicos, tem implementado ações para garantir maior transparência nos seus atos. Todas as instituições públicas estão submetidas à constante fiscalização dos órgãos competentes do governo de modo a responder à sociedade e,com isso, dar suporte à implementação e monitoramento das políticas públicas. No caso das Instituições Federais de Ensino Superior, esse monitoramento e transparência

tem sido cada vez mais importante não só para o conhecimento da sociedade civil, mas também para a sustentação e gestão dessas instituições, que necessitam de indicadores para balizar suas ações futuras. Os dados mais importantes das IFES giram em torno do tripé ”Ensino, Pesquisa e Extensão”, no entanto, a sistematização desses dados, bem como a disponibilização deles para a sociedade ainda é um desafio.

O Brasil possui um portal que concentra dados de pesquisadores de todo o país, trata-se da Plataforma Lattes, criado em agosto de 1999 pelo CNPq – Conselho Nacional de Desenvolvimento Cientifico e Tecnológico, que, por meio do armazenamento dos currículos de pesquisadores, docentes e discentes de todo o país, possui dados da produção científica nacional. Desde a sua criação, o Currículo Lattes tornou-se um padrão nacional compulsório no registro da vida pregressa e atual de estudantes e pesquisadores do país, sendo utilizado pelas principais universidades federais, institutos, centros de pesquisa e fundações de amparo à pesquisa dos estados como instrumento para a avaliação de pesquisadores, professores e alunos, de modo que a Plataforma Lattes constitui-se hoje como um grande repositório de dados de pesquisadores de todo o país: “A Plataforma Lattes representa a experiência do CNPq na integração de bases de dados de Currículos, de Grupos de pesquisa e de Instituições em um único Sistema de Informações. Sua dimensão atual se estende não só às ações de planejamento, gestão e operacionalização do fomento do CNPq, mas também de outras agências de fomento federais e estaduais, das fundações estaduais de apoio à ciência e tecnologia, das instituições de ensino superior e dos institutos de pesquisa. Além disso, se tornou estratégica não só para as atividades de planejamento e gestão, mas também para a formulação das políticas do Ministério de Ciência e Tecnologia e de outros órgãos governamentais da área de ciência, tecnologia e inovação.” [CNPq 2015]

Além disso, cada Currículo Lattes é atrelado univocamente à identidade civil (CPF) de seu portador, que concorda com um termo responsabilizando-se legalmente pelas informações ali contidas. Isso significa que os metadados estão agrupados por autor já em sua fonte original, ao contrário do que ocorre em outras bases. Na WoS, a identificação de autoria é realizada a partir dos nomes adotados nos trabalhos por meio de algoritmos de aprendizagem de máquina e clusterização [CRL 2015]. Na SciELO, essa identificação depende do uso consistente do nome do autor, obedecendo sempre a um mesmo formato em todas as publicações.

Uma vez que o currículo Lattes organiza e armazena informações sobre as atividades desenvolvidas pelos pesquisadores brasileiros em ciência e tecnologia, podemos então vislumbrá-lo como fonte de dados para prover autonomia e controle das IFES sobre seus próprios indicadores, que poderão ser utilizados estrategicamente por seus gestores, além de colaborar no povoamento de informações dos sistemas das instituições, tanto os acadêmicos, como os sistemas voltados para gestão de Recursos Humanos.

Visto o potencial da Plataforma Lattes para o gerenciamento de dados de Ensino, Pesquisa e Extensão das IFES, este trabalho teve por objetivo desenvolver uma solução inteligente de extração dos dados da plataforma, de modo que estes dados ficassem disponíveis não só para utilização nos sistemas de gestão da instituição, como também para os grupos de pesquisa dentro dos interesses de cada grupo.

2. Método

O presente trabalho foi desenvolvida com o apoio do NIT - Materiais[1] e da SIn/UFSCar (Secretaria Geral de Informática) e utiliza como procedimento metodológico a pesquisa-ação, uma das principais formas de abordagem qualitativa [Terence and Filho 2006].

Em geral, a pesquisa-ação é um procedimento apropriado quando a pergunta da pesquisa refere-se a descrever e executar uma série de ações ao longo do tempo em determinado grupo, comunidade ou organização, de modo a compreender, enquanto membro de um grupo, como e por que a sua ação pode mudar ou melhorar o funcionamento de alguns aspectos de um sistema [Coughlan and Coghlan 2002]. Essa abordagem foi escolhida pois tem como foco a "pesquisa em ação", tendo como ideia principal a utilização do método científico para estudar a resolução de importantes problemas sociais ou organizacionais, diretamente em conjunto com aqueles que sofrem esses problemas, ou seja, trata-se de uma abordagem participativa cujo principal objetivo é fazer com que a ação seja mais efetiva, enquanto simultaneamente constrói-se um corpo de conhecimento científico [Coughlan and Coghlan 2002].

A unidade-caso utilizada neste trabalho é a Universidade Federal de São Carlos – UFSCar. Fundada em 1968, e atualmente formada pelos campi São Carlos, Araras, Sorocaba e Lagoa do Sino (em Buri), é uma instituição de grande relevância no cenário nacional. Pode ser considerada a primeira Universidade Federal do Estado de São Paulo, uma vez que a Escola Paulista de Medicina (apesar de ter sido federalizada em 1956) foi elevada ao título de Universidade (UNIFESP) somente em 1994. Hoje, além do Instituto Federal de São Paulo (IFSP), é a única Instituição Federal de Ensino Superior presente no interior do Estado.

O objetivo principal desta pesquisa foi prover uma sistemática inteligente de extração de dados dos pesquisadoes por meio da Plataforma Lattes (CNPq), uma vez que praticamente todo pesquisador brasileiro já está familiarizado com essa ferramenta, devido ao Currículo Lattes ser uma das principais formas de avaliação para concessão de bolsas e projetos de pesquisa por meio das agências de fomento do país, o que culmina no fato já mencionado neste trabalho sobre esta plataforma constituir-se hoje como o principal repositório de dados dos pesquisadores de todo o país.

Os principais procedimentos realizados durante o desenvolvimento da pesquisa foram:

- Prospecção de uma metodologia de consulta ao *web service* da plataforma Lattes via protocolo SOAP, utilizando as linguagens Java e Python;
- Desenvolvimento de um proxy[2] para compartilhar o acesso aos *web services* da plataforma Lattes;

[1] O NIT-Materiais é um núcleo de pesquisa ligado ao departamento de Engenharia de Materias da UFSCar, que atua na pesquisa de prospecção tecnológica e inteligência competitiva, suas metodologias, ferramentas e aplicações para suporte ao desenvolvimento sustentável de empresas, arranjos empresariais e instituições públicas `http://www.nit.ufscar.br`

[2] Servidor intermediário

3. Resultados

3.1. A Plataforma Lattes - Extração de dados por meio de *web service*

Em abril de 2015, a Plataforma Lattes passou a incluir um CAPTCHA[3] para dificultar o acesso automático aos currículos [de Fausto 2015], citando em sua página principal uma preocupação com a "publicação indevida [de espelhos dos currículos] por sites não autorizados".

Mesmo antes da inclusão do CAPTCHA pelo CNPq (dezembro de 2014), ao procurar outras alternativas que não sobrecarregassem os servidores da Plataforma Lattes, verificou-se que esta dispõe de um convênio no qual fornece às Instituições de Nível Superior um acesso direto, por meio de um *web service*[4], para consulta de currículos em formato XML. Esses dados são completos — incluem todas as informações digitadas pelos pesquisadores ou buscadas automaticamente pelo Lattes. Como os dados consultados pelo *web service* estão em formato bruto, e são disponibilizados de forma oficial justamente com a finalidade de acesso automatizado por parte das instituições, o risco de sobrecarregar os servidores do CNPq torna-se praticamente nulo, por este motivo, investimos nossos esforços nesta solução.

3.2. Um proxy customizado para acesso ao *web service* da Paltaforma Lattes

A Plataforma Lattes é um banco de dados público mantido pelo governo que contém os currículos de pesquisadores brasileiros, e que pode ser acessado por qualquer um por meio de um navegador web. Metadados brutos dos currículos em formato XML também podem ser obtidos, mas o download automatizado (sem CAPTCHA) desses dados só é permitido oficialmente por meio de um serviço SOAP[5], que é disponibilizado somente para instituições brasileiras de pesquisa e ensino superior. No entanto, cada instituição só pode solicitar a liberação de acesso para um único endereço IP, por esse motivo, no âmbito deste trabalho, foi criado o cnpqwsproxy[6], um proxy cacheante[7] baseado em *OpenResty*[8] para os *web services* SOAP do CNPq – Plataforma Lates.

Dentre os principais benefícios alcançados pelo proxy, destacam-se:

- Permite que a instituição gerencie sua própria listagem interna de endereços IP que podem acessar o serviço web.
- Assegura que múltiplos aplicativos da mesma instituição acessando o serviço web não causem uma sobrecarga significativa nos servidores do CNPq, fazendo cache das respostas sempre que possível.

[3] *Completely Automated Public Turing test to tell Computers and Humans Apart*: Teste de Turing público completamente automatizado para diferenciação entre computadores e humanos, é um teste bastante utilizado na Internet para assegurar que não sejam realizados acessos automatizados a servidores, e geralmente é composto por imagens distorcidas de letras e números ou sons de difícil compreensão.

[4] Um *web service* é um conjunto de métodos (*web methods*) logicamente associados e chamados através de um servidor HTTP.

[5] *Simple Object Access protocol* – SOAP, é um protocolo de comunicação baseado em XML para troca de informações estruturadas na implementação de *web services*.

[6] `https://github.com/nitmateriais/cnpqwsproxy`

[7] *Cache* é uma cópia local, ou próxima de quem requisita os dados, das informações mais acessadas dentre um conjunto de dados que está distante ou cujo acesso é lento.

[8] *OpenResty* é uma plataforma composta pelo nginx `http://nginx.org`, LuaJIT `http://luajit.org`, e alguns módulos de extensão.

- Preserva a compatibilidade com quaisquer aplicativos existentes. Mudar o endereço do serviço web no arquivo WSDL[9] ou sobrescrever a resposta do servidor DNS[10] usando o arquivo `/etc/hosts` (em um sistema compatível com UNIX) é suficiente para fazer com que um aplicativo preexistente utilize o proxy.

4. Conclusão

O servidor intermediário (proxy) desenvolvido no âmbito deste trabalho permite que múltiplos aplicativos tenham acesso aos dados extraídos da Plataforma Lattes. Esses dados podem, por exemplo, servir de base para um Sistema de Gestão de Informação Científica (CRIS – Current Research Information System), fornecendo informações a respeito de projetos, publicações, agências financiadoras, dentre outras, que poderão, uma vez sistematizadas, contribuir para a tomada de decisão dos gestores institucionais [Amante et al. 2014].

Este projeto também contribuirá para o desenvolvimento de um sistema de apoio à progressão docente, uma vez que, sistematizando os dados extraídos do Lattes, pode-se criar automaticamente os perfis de publicação de cada docente no sistema de gestão da instituição, aproveitando os dados inseridos pelo docente na Plataforma, evitando retrabalho na organização dessas informações.

Todos os scripts e ferramentas criados neste trabalho foram referenciados no decorrer do texto com seus respectivos endereços na Internet. O conjunto de ferramentas utilizadas nesta sistemática pode ser encontrado na página da organização NIT–Materiais no GitHub[11].

Nosso objetivo em compartilhar este proxy faz parte da ideologia de dados abertos do governo e também, uma vez que as IFES adotem o uso deste software, zela pela autonomia dessas intituições sobre seus próprios dados e pela redução do tráfego na troca de dados entre as universidades e a plataforma Lattes, evitando o congestionamento dos servidores.

Referências

Amante, M. J., Lopes, S., Marçal, B., and Segurado, T. (2014). *Cardernos BAD*, (2):83–93.

CNPq (2015). A criação – portal cnpq.

Coughlan, P. and Coghlan, D. (2002). Action research for operations management. *International Journal of Operations & Production Management*, 22(2):220–240.

CRL (2015). Database: Web of science.

de Fausto, S. S. (2015). Captcha nos CVs Lattes.

Terence, A. C. F. and Filho, E. E. (2006). Abordagem quantitativa, qualitativa e a utilização da pesquisa-ação nos estudos organizacionais. In *XXVI Encontro Nacional de Engenharia de Produção*, Fortaleza. ABEPRO.

[9] *Web Services Description Language*: Linguagem para Descrição de Serviços Web.
[10] *Domain Name System*: Sistema de Nomes de Domínio.
[11] `https://github.com/nitmateriais`