Utilizando o framework LattesDataXplorer para vincular automaticamente os currículos da Plataforma Lattes à Biblioteca Digital Brasileira de Teses e Dissertações (BDTD)

Thiago Magela Rodrigues Dias
Doutor em Modelagem Matemática e Computacional pelo Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) - Belo Horizonte, MG - Brasil. Professor do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET) - Divinópolis, MG - Brasil.
http://lattes.cnpq.br/4687858846001290
E-mail: thiagomagela@gmail.com

Washington Luís Ribeiro de Carvalho Segundo
Doutor em Informática pela Universidade de Brasília (UnB) – Brasília, DF – Brasil, com período sanduíche em King's College London - Londres. Coordenador do Laboratório de Metodologias de Tratamento e Disseminação da Informação (COLAB) do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) – Brasília, DF – Brasil.
http://lattes.cnpq.br/9453481318889500
E-mail: washingtonsegundo@ibict.br

Lautaro Julián Matas
Doutor em Computação.
Technical Support – La Referencia
http://lattes.cnpq.br/6322724400053732
E-mail: lmatas@gmail.com

Data de submissão: 27/09/2019. Data de aprovação no ConfOA: 12/06/2019. Data de publicação:

RESUMO

Um dos principais problemas em repositórios científicos é identificar relacionamentos e vincular metadados de diferentes fontes. Este trabalho descreve brevemente os últimos resultados de um esforço para construir uma plataforma de software comum capaz de processar metadados de diferentes fontes heterogêneas. O estudo de caso para essa fase inicial é o encadeamento da Biblioteca Digital Brasileira de Teses e Dissertações (BDTD) e dos currículos da Plataforma Lattes, aplicando uma estratégia de transformação de *string*. Os resultados preliminares mostram um caminho promissor a seguir para alcançar uma implementação de nível de produção.

Palavras-chave: Ligação de dados. Enriquecimento de metadados. BDTD. Plataforma Lattes.

Using the LattesDataXplorer framework to automatically link Lattes Platform resumes to the Brazilian Digital Library of Theses and Dissertations (BDTD)

ABSTRACT

One of the main problems with scientific repositories is identifying relationships and linking metadata from different sources. This paper briefly describes the latest results of an effort to build a common software platform capable of processing metadata from different heterogeneous sources. The case study for this initial phase is the chaining of the Brazilian Digital Library of Theses and Dissertations (BDTD) and the Lattes Platform curriculum, applying a string transformation strategy. Preliminary results show a promising way forward to achieve production level implementation.

Keywords: Data Binding. Metadata Enrichment. BDTD. Lattes Platform.

Uso del marco LattesDataXplorer para vincular automáticamente los currículums de la Plataforma Lattes con la Biblioteca Digital Brasileña de Tesis y Disertaciones (BDTD)

RESUMEN

Uno de los principales problemas con los repositorios científicos es identificar relaciones y vincular metadatos de diferentes fuentes. Este documento describe brevemente los últimos resultados de un esfuerzo por construir una plataforma de software común capaz de procesar metadatos de diferentes fuentes heterogéneas. El estudio de caso para esta fase inicial es el encadenamiento de la Biblioteca Digital Brasileña de Tesis y Disertaciones (BDTD) y el plan de estudios de la Plataforma Lattes, aplicando una estrategia de transformación de cadenas. Los resultados preliminares muestran un camino prometedor para lograr una implementación de nivel de producción.

Palabras clave: *Enlace de datos. Enriquecimiento de metadatos. BDTD. Plataforma Lattes.*

INTRODUÇÃO

A Biblioteca Digital Brasileira de Teses e Dissertações (BDTD) <http://bdtd.ibict.br> é uma rede de mais de 100 instituições que agregam mais de meio milhão de teses e dissertações eletrônicas em acesso aberto. Esse portal agregador utiliza o software de coleta provido pela rede LA Referencia (o LRHarvester). Além disso, o conteúdo da BDTD é coletado pela rede LA Referencia via oasisbr <http://oasisbr.ibict.br> (Carvalho-Segundo *et al.*, 2017), e também pela NDLTD <http://search.ndltd.org/>, onde figura como o segundo maior consórcio nacional.

A Plataforma Lattes <http://lattes.cnpq.br/> é uma base de dados com mais de 6 milhões de currículos. O pesquisador declara nessa plataforma sua formação, produção acadêmica, participação em congressos e projetos, premiações acadêmicas etc. No Brasil, ter um currículo Lattes é uma exigência para a apresentação de uma proposta de financiamento. Além disso, as agências governamentais vêm se empenhando na criação de serviços de interoperabilidade entre o ORCID, a Plataforma Lattes, repositórios científicos de acesso aberto e plataformas de financiamento.

No Brasil, os registros da BDTD possuem um esquema de metadados mais rico que os repositórios padrão de publicações científicas. Por exemplo, autores, orientadores, coorientadores e membros de banca podem anexar seus identificadores dos currículos da Plataforma Lattess através de campos específicos do esquema de metadados. Infelizmente, a tarefa de preencher os identificadores é feita manualmente e pequena quantidade dos registros é preenchida corretamente. No entanto, os identificadores são importante elemento para a construção de métricas e análise de dados nos repositórios. Outro aspecto importante é que essas estratégias de vinculação são um passo em direção à construção de Sistemas de Informações de Pesquisa Corrente (os ecossistemas CRIS).

METODOLOGIA

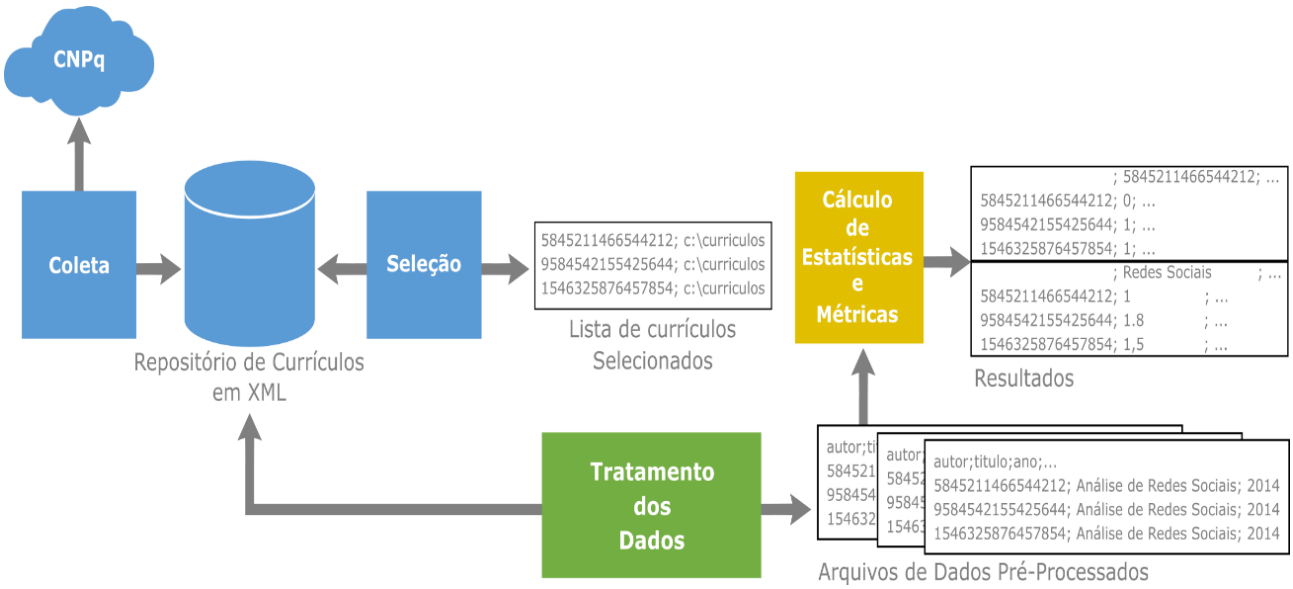
O escopo do presente trabalho é a implementação de uma estratégia de ligação automática entre os registros do BDTD e os currículos da Plataforma Lattes. A seguir são listadas as estratégias adotadas.

Metadados de publicação automática e vinculação de currículos. Como essa é uma iniciativa colaborativa, um dos principais objetivos era fornecer uma plataforma comum para que os desenvolvedores pudessem contribuir com diferentes estratégias e comparar resultados usando os mesmos conjuntos de metadados. Nessa fase, duas estratégias de diferentes grupos de pesquisa estão sendo integradas. Em ambos os casos, a ideia principal é usar títulos, autores e outros campos de metadados para inferir as relações entre um registro da BDTD e os currículos de autores e orientadores na Plataforma Lattes. A seguir, uma estratégia de transformação de *string* é apresentada.

Estratégia baseada em transformação de *string*. Nessa estratégia, o objetivo principal é diminuir ao máximo o custo computacional necessário para a comparação de títulos, contrapondo-se à comparação de *stings* via força bruta geralmente adotada em outras estratégias. O processo inicial é baseado na análise de cada um dos títulos de orientação e formação acadêmica (mestrado e doutorado) registrados na Plataforma Lattes, gerando uma chave para um dicionário com os títulos encontrados, vinculando a cada uma dessas chaves um identificador único.

Estrutura LattesDataXplorer. Para realizar a análise apresentada neste trabalho, o framework LattesDataXplorer (Dias 2016) foi usado para coletar os dados curriculares da Plataforma Lattes. Esse framework abrange todo um conjunto de técnicas e métodos para coletar, selecionar, processar e analisar dados contidos em determinado currículo armazenado na Plataforma Lattes. Uma visão geral do LattesDataXplorer é mostrada na figura 1.

Figura 1 – Visão geral do LattesDataXplorer



Fonte: Dias (2016).

Inicialmente, o **módulo coleta** é executado para a extração dos currículos registrados na Plataforma Lattes. Nessa etapa, uma solicitação é feita diretamente à plataforma, na qual o currículo é extraído e armazenado em formato XML. Após o armazenamento local dos currículos baixados, é possível manipular os dados com flexibilidade e explorar todo o potencial que os currículos oferecem.

A fim de analisar grupos específicos dos perfis, como aqueles compostos por professores de programas de pós-graduação ou de uma instituição em particular, o componente chamado Seleção é usado para a composição de subgrupos com base nas informações presentes em seus registros.

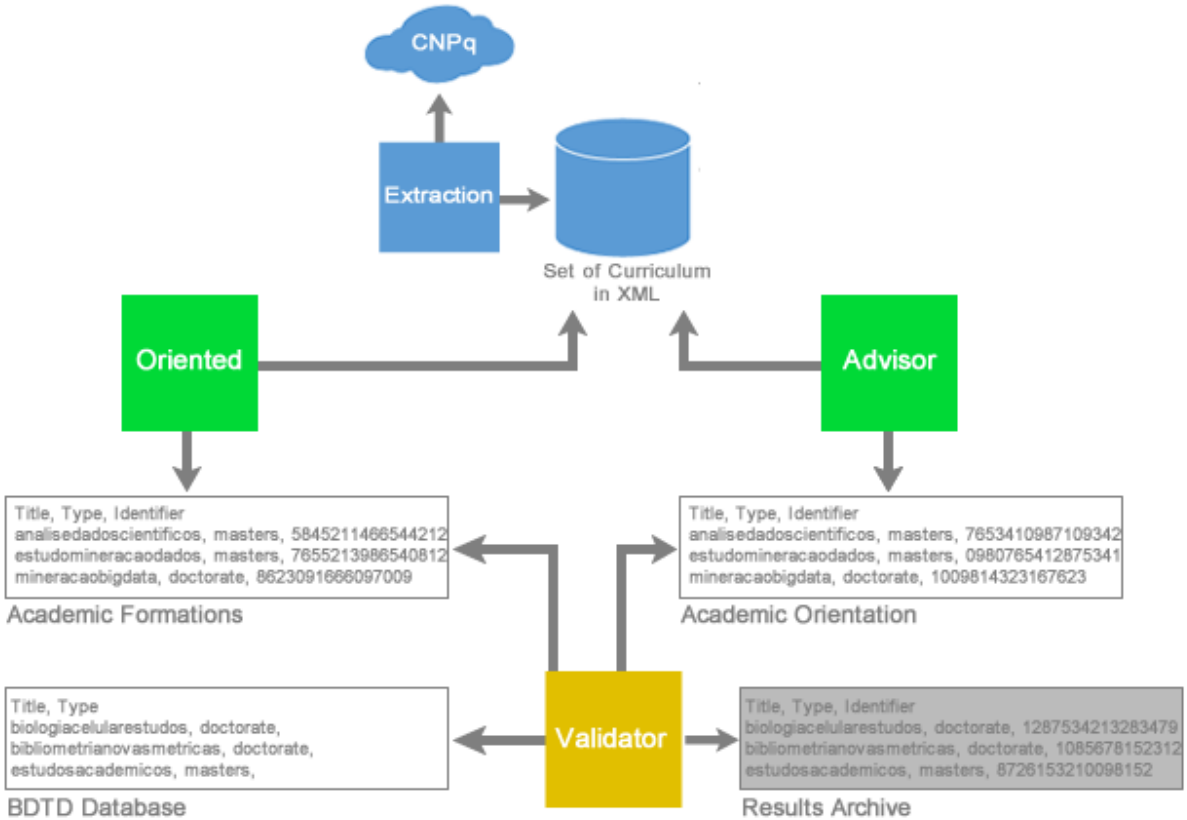
Para se executar a tarefa de seleção, a linguagem de consulta XPath (XML Path Language) é usada. A linguagem XPath permite construir expressões para processar um documento XML de maneira muito semelhante ao uso de expressões regulares. Portanto, é possível agrupar um conjunto de perfis com base em determinados parâmetros predeterminados.

RESULTADOS

Neste trabalho, foram utilizados seis milhões de currículos coletados em janeiro de 2019. Todos os currículos foram analisados, e verificou-se se havia algum registro de formação acadêmica completa e seus respectivos orientadores acadêmicos (em mestrado e doutorado). Para cada formação ou supervisão encontrada em um currículo específico, os títulos desses registros passaram por um processo de tratamento e então foram inseridos em um dicionário. Dicionários específicos foram criados para autores (orientados) e orientadores (figura 2).

Como pode ser observado, após a caracterização dos dicionários de formação acadêmica e orientações acadêmicas, são gerados dois arquivos, cada um correspondendo a cada dicionário. Nessa estratégia, cada currículo é revisado apenas uma vez sem a necessidade de realizar comparações com outros currículos. Portanto, a única comparação feita é verificar se a chave resultante do título processado existe no dicionário. Com isso, o custo computacional para análise é da ordem de O(N), permitindo que grandes quantidades de orientações acadêmicas ou formações acadêmicas sejam realizadas em poucos minutos.

Figura 2 – Estrutura para cruzamento de dados da Plataforma Lattes e banco de dados BDTD



Fonte: Elaborada pelos autores.

Posteriormente, um componente chamado *Validator* é responsável por verificar no banco de dados do BDTD se a chave transformada com o mesmo tratamento dos títulos aplicados na Plataforma Lattes existe nos dicionários caracterizados. Se a chave já existir no dicionário de formações acadêmicas, isso significa que o autor em questão foi encontrado e seu identificador da Plataforma Lattes é incorporado à base da BDTD. Se a chave é encontrada no arquivo de orientações acadêmicos, isso significa que o orientador desse trabalho foi encontrado.

Assim, com as inserções do autor e orientador, o arquivo de resultados é gerado contendo o banco de dados BDTD original, embutido com os identificadores de autor e orientador.

Após a construção do arquivo de resultados, utilizaram-se os dados de um subconjunto da coleção BDTD (87.341 registros) que possuem o identificador Lattes atribuído. Esse subconjunto foi utilizado como prova de controle no cálculo de erro da estratégia adotada. Esses dados foram utilizados para calcular a precisão e a revocação da estratégia.

No cálculo da precisão foi possível obter uma porcentagem de 100% de acerto, o que expressa que nos casamentos sugeridos todos eram verdadeiros, mostrando que o algoritmo é confiável quando sugere uma vinculação. Em relação à revocação que indica o percentual recuperado pela estratégia no conjunto possível, a taxa de acerto foi de 76,7%. Esse percentual é visto como um bom resultado, considerando que outras estratégias, com maior custo computacional, têm semelhante comportamento, o que se traduz como resultado importante para um esforço inicial na tentativa de identificar autores e orientadores.

CONCLUSÕES

Assim, a estratégia de transformação de *strings* apresentada revelou-se uma importante tentativa de identificar autorias e orientações, com baixo custo computacional e passível de aplicação em grandes bases de dados.

Essa solução pode ser uma alternativa interessante para a primeira tentativa de realizar a vinculação, principalmente quando se considera a precisão de suas identificações e sua taxa de revocação. Após a implementação dessa estratégia, os registros que não puderam ser identificados (aproximadamente 23%) poderão ser analisados com outros algoritmos que exigem custo computacional maior.

É importante ressaltar que a mesma estratégia pode ser adotada com outros tipos de documentos, como artigos de conferência (ou de periódicos), a fim de realizar vinculações que anteriormente não eram possíveis de ser caracterizadas.

REFERÊNCIAS

CARVALHO-SEGUNDO, W. *et al. The LA Referencia Software and the Brazilian Portal of Scientific Open Access Publications (oasisbr)*. Brasília: IBICT, 2017.

DIAS, T. M. R. *Um Estudo Sobre a Produção Científica Brasileira a partir de dados da Plataforma Lattes.* 2016. 181 f. Tese (Doutorado em Modelagem Matemática e Computacional) - Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2016.