VALIDAÇÃO AUTOMÁTICA DOS CURRÍCULOS DA PLATAFORMA LATTES À BIBLIOTECA DIGITAL BRASILEIRA DE TESES E DISSERTAÇÕES (BDTD)

AUTOMATIC VALIDATION OF LATTES PLATFORM CURRICULA TO THE BRAZILIAN DIGITAL LIBRARY OF THESES AND DISSERTATIONS (BDTD)

Thiago Magela Rodrigues Dias | Washington L. R. de Carvalho-Segundo

https://doi.org/10.21747/21836671/pagnesppk21

Resumo: Um dos principais problemas em repositórios científicos é identificar relacionamentos e vincular metadados de diferentes fontes. Este trabalho descreve brevemente os novos resultados em um esforço para se construir uma plataforma de *software* capaz de processar metadados de diferentes fontes heterogêneas. O estudo de caso para essa fase é a vinculação da rede de repositórios de teses e dissertações, a Biblioteca Digital Brasileira de Teses e Dissertações (BDTD), e a Plataforma Lattes de currículos de pesquisadores. Aplica-se uma estratégia de geração de *hashes* para registros, via regras de transformação sobre título e ano, capturando-se, em um só passo, os identificadores Lattes de autor e orientador das teses e dissertações.

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

Abstract: One of the main problems in scientific repositories is to identify relationships and link metadata from different sources. This paper briefly describes the new results in an effort to build a software platform able to process metadata from different heterogeneous sources. The case study for this phase is the linking of the thesis and dissertation repositories network, the Brazilian Digital Library of Theses and Dissertations (BDTD), and the Lattes Platform of researchers' *curricula*. A strategy for generating hashes for records is applied, through transformation rules on title and year, capturing, in one step, the Lattes identifiers of theses and dissertations authors and supervisors.

Keywords: BDTD; Metadata enrichment; Linked data, Lattes Platform.

Introdução

A Biblioteca Digital Brasileira de Teses e Dissertações (BDTD -http://bdtd.ibict.br) é uma rede de mais de cem instituições que agregam quase 670 mil teses e dissertações eletrônicas em acesso aberto. Este portal agregador utiliza o *software* de coleta provido pela rede LA Referencia (LRHarvester - https://github.com/lareferencia). 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 ela figura como o segundo maior consórcio nacional.

Por outro lado, a Plataforma Lattes (http://lattes.cnpq.br/) é uma base de dados com mais de 6,5 milhões de currículos, onde aproximadamente 340 mil se declaram como doutores e 900 mil como mestres. O pesquisador declara nesta 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.

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 Lattes através de campos específicos do esquema de metadados. Infelizmente, esta tarefa de preencher os identificadores é feita manualmente e uma pequena quantidade dos registros é preenchida de forma correta. No entanto, os identificadores Lattes são um 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 (CRIS) e grafos de informação sobre a gestão da pesquisa científica, ao modelo do que é realizado, por exemplo, no OpenAIRE Research Graph (https://zenodo.org/communities/openaire-research-graph).

Metodologia

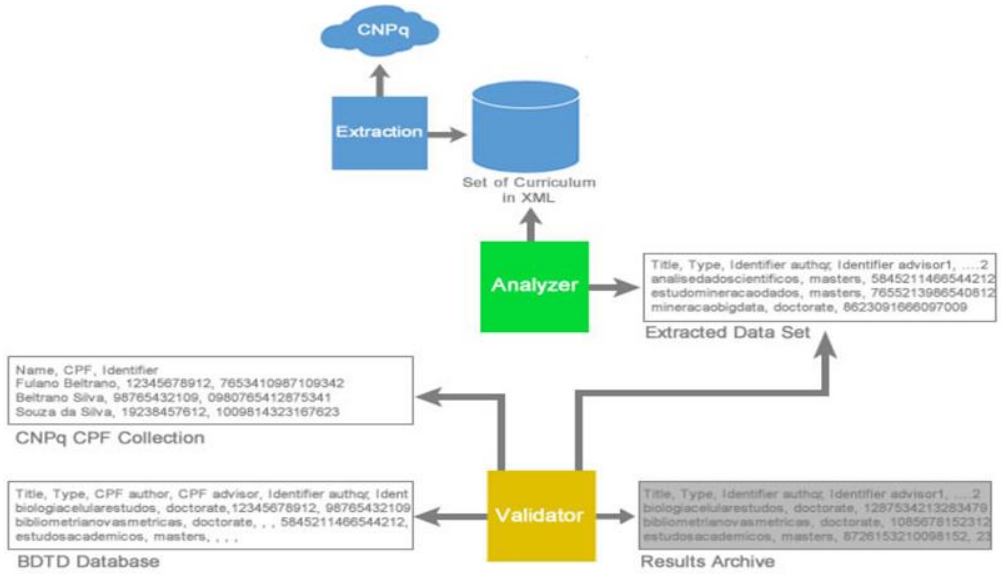
O escopo do presente trabalho é a implementação de uma estratégia de ligação automática entre os registros da BDTD e os currículos da Plataforma Lattes. A seguir é dada uma descrição da estratégia adotada.

Estratégia baseada em transformação de *string*. Nesta estratégia, o objetivo principal é minimizar ao máximo o custo computacional necessário para a comparação de títulos, contrapondo-se a comparação de *strings* via força-bruta que geralmente é 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.

LattesDataXplorer. Para se realizar a análise apresentada neste trabalho, foi utilizado o *framework* LattesDataXplorer (Dias 2016) na coleta dos dados curriculares da Plataforma Lattes. Esse *framework* abrange todo um conjunto de técnicas e métodos para coletar, selecionar, processar e analisar dados contidos em um determinado currículo armazenado na Plataforma. Logo, no intuito de propor um processo que contemplasse todos os passos necessários para realizar a validação automática dos dados contidos nos currículos cadastrados na Plataforma Lattes e BDTD, um arcabouço de componentes foi proposto (Fig. 1).

Inicialmente, o módulo **Extraction** (originalmente do LattesDataXplorer) é executado para a extração dos currículos registrados na Plataforma Lattes. Nesta 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 informacional que os currículos oferecem.

Fig. 1 – ***Estrutura para validação automática de dados da Plataforma Lattes e banco de dados BDTD***

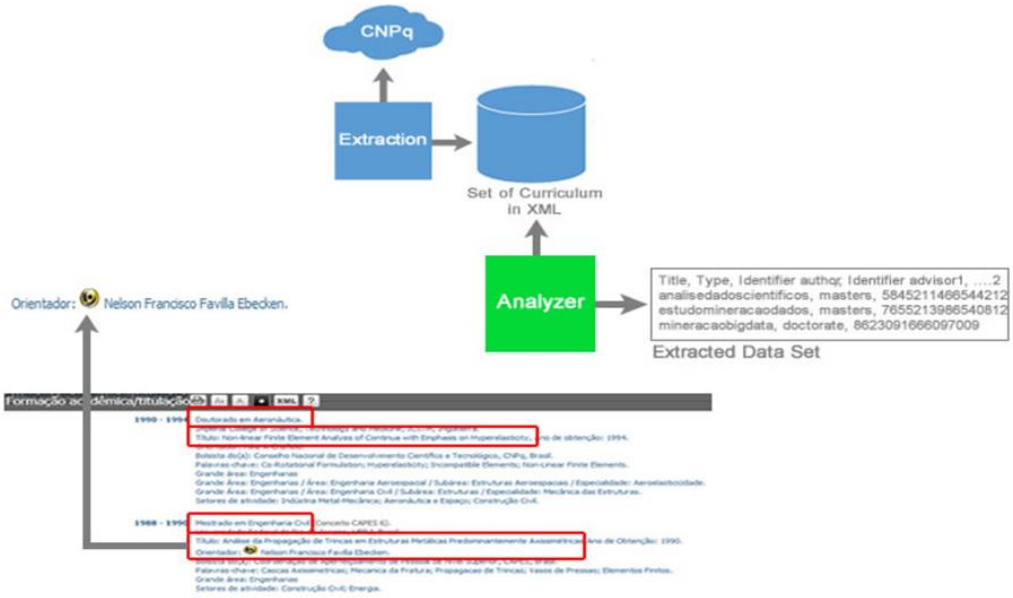


Fonte: Os autores.

Como pode ser observado, após a extração dos dados, o módulo **Analyzer** é responsável por extrair todas as informações de interesse e caracterizar um dicionário contendo informações de cada uma das formações acadêmicas (mestrado e doutorado) e suas informações complementares, a saber, título, tipo (tese ou dissertação), orientando e orientadores. Utilizando-se a estratégia de comparação de strings, os títulos de publicações são transformados em chaves (hashes) do dicionário gerado, incrementando-o a medida que novas chaves são geradas. 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.

O módulo **Analyzer**, percorre uma só vez duas seções de cada currículo, Formação Acadêmica e Orientações Concluídas, otimizando assim o número de consultas realizadas. O modo de análise das formações acadêmicas pode ser visualizado na Fig. 2. Como mostra esta Fig. 2, em cada uma das formações de interesse do currículo que está sendo analisado é verificado se seu título transformado em chave está no dicionário. Além disso, o identificador do aluno (que é o detentor do currículo) é inserido juntamente com a chave (título transformado). Caso o identificador do orientador daquela formação esteja implicitamente vinculado ao trabalho (em que seu identificador está inserido) o identificador do orientador também é inserido juntamente à chave do dicionário. Esta estratégia faz com que tanto o identificador do orientando como o identificador do orientador sejam inseridos juntamente com a chave (título e ano da formação) no dicionário. Posteriormente, a seção de Orientações Concluídas é também analisada.

Fig. 2 –*Análise das formações acadêmicas*



Fonte: Os autores.

Cada uma das Orientações Concluídas (mestrado e doutorado) é analisada. Após a transformação do título da orientação com o método já descrito, ela se torna uma chave e esta chave é inserida no dicionário juntamente com o identificador do orientador (detentor do currículo em análise), caso a chave já exista no dicionário, apenas o identificador do orientador é vinculado a esta chave. Caso o identificador do aluno daquela orientação, esteja implicitamente vinculado ao trabalho (em que seu identificador está inserido) o identificador do orientado também é inserido juntamente a esta chave no dicionário. Como dito anteriormente, esta estratégia faz com que tanto o identificador do orientador como o identificador do orientando sejam inseridos juntamente com a chave (título e ano da formação) no dicionário.

Posteriormente, um componente chamado **Validator** é responsável por verificar no banco de dados 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 deste 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.

Além disso, para os registros da BDTD que ainda não foram inseridos, mas que possuem o CPF, tanto do aluno como do orientador, é realizada uma consulta, em um dicionário, obtido junto ao CNPq, que possui um mapeamento entre CPF e identificador na Plataforma Lattes. Neste caso, o CPF é substituído pelo identificador correspondente. Gerando desta forma o arquivo de resultados.

Resultados

Após a construção do arquivo de resultados, utilizou-se os dados de um subconjunto da coleção BDTD (aproximadamente 90 mil registros) que possuem o Identificador Lattes atribuído. Este subconjunto foi utilizado como uma prova de controle no cálculo de erro da estratégia adotada. Estes dados foram utilizados para se calcular a precisão e a revocação da estratégia.

No cálculo da **precisão**, realizado, por enquanto, apenas com o conjunto de doutores, foi possível obter uma percentagem 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 a **revocação** que indica o percentual recuperado pela estratégia no conjunto possível, a taxa de acerto foi de **87,7%**. Esse percentual é apresentado como um resultado relevante considerando que outras estratégias, com maior custo computacional, têm semelhante comportamento. O valor de revocação obtido apresenta uma melhoria ao apresentado no trabalho em (DIAS *et al*., 2019).

Conclusão

A nova estratégia apresentada mostrou-se ser uma importante tentativa de identificar autorias e orientações, com um 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.

Referências bibliográficas

CARVALHO-SEGUNDO, W. [et al.]
2017 *The LA Referencia Software and the Brazilian portal of scientific open access publications (oasisbr)*. Open Repositories, 2017.

DIAS, T. M. R.
2016 Um Estudo sobre a produção científica brasileira a partir de dados da Plataforma Lattes. 2016. Tese de Doutorado em Modelagem Matemática e Computacional - Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG).

DIAS, T. M. R.; CARVALHO-SEGUNDO, W.; MATAS, L.
2019 Utilizando o framework LattesDataXplorer para vincular automaticamente os currículos da Plataforma Lattes à Biblioteca Digital Brasileira de Teses e Dissertações (BDTD). *Ciência da Informação*. [Em linha]. 48:3 (2019). [Consult. abr. 2020]. Disponível em: http://revista.ibict.br/ciinf/article/view/5003.

Thiago Magela Rodrigues Dias | thiagomagela@cefetmg.br
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Brasil

Washington L. R. de Carvalho-Segundo | washingtonsegundo@ibict.br
Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), Brasil