Pecha Kucha

DOI:10.21680/2447-7842.2023v9n2ID33837

Integração de repositórios de dados abertos para certificação de produção científica

Integration of open data repositories for scientific production certification

Washington Luís Ribeiro Carvalho Segundo [1]

Thiago Magela Rodrigues Dias [2]

Tales Henrique José Moreira [3]

Vivian Santos Silva [4]

Jesús Pascual Mena-Chalco [5]

Submetido em: 17/04/2023	Aprovado na ConfOA: 14/06/2023	Publicado em: 04/12/2023

Resumo: Nos últimos anos, diversas iniciativas voltadas para a criação de sistemas que gerenciam a produção acadêmica de uma instituição, país ou área do conhecimento, têm recebido atenção de diferentes áreas. Tais sistemas são conhecidos pela sigla CRIS (*Current Research Information Systems*) e destinam-se a agregar informações de diversas bases de dados a fim de fornecer relatórios, dados consolidados para que os pesquisadores possam analisar, bem como ser objeto de certificação da informação científica e tecnológica. Portanto, este trabalho apresenta o processo de integração de dados da Plataforma BrCris com o objetivo

[1] Doutor e Mestre em Informática pela Universidade de Brasília, com Estágio de Doutorado Sanduíche no King's College London.
[2] Doutor e Mestre em Modelagem Matemática e Computacional pelo Centro Federal de Educação Tecnológica de Minas Gerais.
[3] Mestre em Modelagem Matemática e Computacional pelo Centro Federal de Educação Tecnológica de Minas Gerais.
[4] Doutora em Ciência da Computação pela Universität Passau e Mestre em Informática pela Universidade Federal do Rio de Janeiro.
[5] Doutor e Mestre em Ciências da Computação pela Universidade de São Paulo.

de fornecer dados confiáveis, para que seja possível realizar o processo de certificação de informações sobre a comunidade científica brasileira. Essa estratégia se apresenta como um importante mecanismo de agregação de dados, fornecendo informações confiáveis e importantes sobre dados científicos, especialmente sobre a produção científica e tecnológica brasileira. Portanto, a utilização dos dados agregados da Plataforma BrCris para o processo de certificação proporcionará diversos estudos bibliométricos que a priori seriam extremamente complexos em sua elaboração.

Palavras-chave: integração de dados; certificação; produção científica; BrCris.

Abstract: In recent years, several initiatives aimed at creating systems that manage the academic production of an institution, country or area of knowledge have received attention from different areas. Such systems are known by the acronym CRIS (Current Research Information Systems) and are intended to aggregate information from various databases in order to provide reports, consolidated data for researchers to analyze, as well as to be the object of certification of scientific information and technological. Therefore, this work presents the data integration process of the BrCris Platform with the objective of providing reliable data, so that it is possible to carry out the process of certifying information about the Brazilian scientific community. This strategy presents itself as an important data aggregation mechanism, providing reliable and important information on scientific data, especially on Brazilian scientific and technological production. Therefore, the use of aggregated data from the BrCris Platform for the certification process will provide several bibliometric studies that a priori would be extremely complex in their elaboration.

Keywords: data integration; certification; scientific production; BrCris.

1 INTRODUÇÃO

A produção científica brasileira tem crescido significativamente e, tendo em vista as especificidades de diferentes campos disciplinares, heterogênea quanto à tipificação de sua produção tanto em termos quantitativos quanto qualitativos. E o resultado dessa produção se materializa na forma de artigos em periódicos, teses e dissertações, além de diversos produtos como: softwares, patentes, obras e instalações artísticas, entrevistas e projetos cinematográficos.

A partir desse cenário, começaram a surgir iniciativas voltadas para a criação de sistemas que gerem a produção acadêmica de uma instituição, país ou área do conhecimento. Tais sistemas são conhecidos pela sigla CRIS (*Current Research Information Systems*) e têm como objetivo agregar informações de diversas bases de dados a fim de fornecer relatórios e dados consolidados para que os pesquisadores da área possam analisar como ocorre a produção em seus países ou áreas de atuação.

Assim, o BrCris visa estabelecer um modelo único de organização da informação científica de todo o ecossistema de pesquisa brasileira. Entre os agentes desse ecossistema estão pesquisadores, projetos, infraestruturas, laboratórios e instituições de pesquisa, financiadores, além de resultados de pesquisa expressos principalmente por publicações científicas, teses, dissertações, conjuntos de dados científicos, software e patentes.

Portanto, com a integração dos dados em um repositório de dados padronizado e devidamente avaliado, pode-se realizar todo um processo de certificação de dados originários de outras fontes ainda não validadas, proporcionando uma visão real da produção científica e tecnológica brasileira.

2 DESENVOLVIMENTO

O BrCris concentra um amplo ecossistema de dados de diversas fontes, como dados curriculares de indivíduos, de organizações, programas de pós-graduação, publicações, periódicos científicos, entre outros, exigindo todo esforço para

processar os dados de interesse. Nesse contexto, tendo em vista as diversas fontes de dados que irão compor o BrCris, faz-se necessária a transformação dos dados para um formato padronizado, sendo necessária a transformação com base em um modelo de dados.

Com os dados coletados e já deduplicados, classificados e categorizados, eles podem ser posteriormente adaptados e validados, estabelecendo relações com registros de outras fontes. Um registro coletado na fonte “A” tem um atributo comum com o registro coletado na fonte “B”, podendo ser estabelecida uma ligação entre ambos, com certo grau de confiabilidade. Os outros atributos de registro podem ser mesclados para resultar em um único registro enriquecido, eliminando as réplicas. Um esquema de validação pode ser criado para descartar registros malformados, redundantes, inconsistentes ou ambíguos.

3 CONSIDERAÇÕES

No contexto deste trabalho, o primeiro modelo de certificação testado é a integração da Plataforma Lattes (Lane, 2010) com o Oasis.br (https://oasisbr.ibict.br/). Por meio de desdobramentos no âmbito do Projeto BrCris, foi possível criar um mecanismo inteligente de identificação de teses e dissertações declaradas nas sessões de formação acadêmica e orientações concluídas de um determinado Currículo cadastrado na Plataforma Lattes, que também constavam no cadastro agregado pelo Oasisbr.

Desta forma, a Oasisbr torna-se o “terceiro de confiança” neste processo, sem a necessidade da preexistência de um identificador persistente explicitamente atribuído à tese ou dissertação.

Todo o processo de certificação é baseado em estratégias computacionais testadas e validadas em diversos estudos, por meio da análise de informações autodeclaradas, comparadas com informações inseridas em repositórios, bibliotecas digitais e portais agregados pelo Oasisbr. O selo de certificação Oasisbr é exibido próximo aos títulos das teses ou dissertações no currículo do usuário.

As vantagens do processo de certificação são muitas. Por meio da certificação, é possível verificar que os trabalhos científicos, orientações, participação em bancas, entre outros elementos de fontes autodeclaradas são realmente verdadeiros, evitando informações falsas.

REFERÊNCIA

Lane, J. (2010). Let's make science metrics more scientific. *Nature*, 464(7288), 488-489.