

Práticas de compartilhamento de dados pelos pesquisadores PQ da área de Física e Astronomia do CNPQ

RENÉ FAUSTINO GABRIEL JUNIOR

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

rene.gabriel@ufrgs.br

TATYANE GUEDES MARTINS DA SILVA

THIAGO MAGELA RODRIGUES DÍAS

RESUMO

Compartilhar os dados de pesquisa significa promover a transparência, promover a colaboração científica, evitar a duplicação desnecessária de esforços e possibilitar a reprodutibilidade das pesquisas. Faz-se a importância de estudar as práticas de compartilhamento de dados de pesquisa pelas diversas áreas do conhecimento de forma a ter um entendimento de suas práticas e a cultura do compartilhamento de dados. O uso dos identificadores persistentes nas publicações científicas, repositórios científicos e nos conjuntos de dados está possibilitando o cruzamento de informações nas diversas bases de dados. O uso de robôs de coleta também tem facilitado o manuseio desses identificadores em diversas bases de dados.

Práticas de compartilhamento de dados pelos pesquisadores PQ da área de Física e Astronomia do CNPQ

Rene Faustino Gabriel Junior (1), Tatyane Guedes Martins da Silva (2), Thiago Magela Rodrigues Dias (3)

Introdução

Compartilhar os dados de pesquisa significa promover a transparência, promover a colaboração científica, evitar a duplicação desnecessária de esforços e possibilitar a reprodutibilidade das pesquisas. Faz-se a importância de estudar as práticas de compartilhamento de dados de pesquisa pelas diversas áreas do conhecimento de forma a ter um entendimento de suas práticas e a cultura do compartilhamento de dados.

O uso dos identificadores persistentes nas publicações científicas, repositórios científicos e nos conjuntos de dados está possibilitando o cruzamento de informações nas diversas bases de dados. O uso de robôs de coleta também tem facilitado o manuseio desses identificadores em diversas bases de dados.

Objetivo

Identificar as práticas de compartilhamento de dados por pesquisadores PQ da área de física e astronomia do CNPQ

Dados Abertos CAPES

Plataforma Lattes

ScholarXplorer

Metodologia

Para identificar as práticas de compartilhamento de dados de pesquisa por pesquisadores das áreas de física e astronomia parte-se do arquivo BR-CAPES-COLSUCUP-DOCENTE-2020-2021-11-10 disponibilizado no Portal Dados Abertos CAPES. Extraí os pesquisadores da área de avaliação 3 (física e astronomia) que tiveram ou tem bolsa produtividade PQ do CNPQ resultando em uma lista de 1.078 pesquisadores.

Para possibilitar a análise foram coletadas das produções científicas, foi necessário a utilização de uma API para coleta de dos arquivos XML no Lattes Extrator de cada pesquisador. Esta coleta resultou na identificação de 107.444 artigos, sendo 19.377 (18,0%) desses trabalhos sem DOI. Com este corpus de 88.607 (82,0%) trabalhos foi construído um script em Python, para consulta via API na base do ScholarXplorer. Para não sobrecarregar o servidor, as consultas foram realizadas com intervalos de 1 seg. cada, resultando aproximadamente 24 horas para finalizar as consultas. Todos os dados foram armazenando em uma base de dados MySQL.

Com os dados pode-se verificar a relação entre os DOIs, o que o ScholarXplorer chama de "Source" e "Target" com seu "Relationship". Os resultados dessas relações foram inseridas em um dataframe para serem analisados.

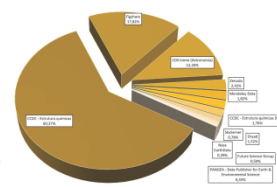
Resultados

Com o processamento dos dados foi possível verificar que os DOIs "target" coletados no ScholarXplorer podem ser divididos em dois grandes grupos, um referente aos conjuntos de dados de pesquisa e um segundo relacionado a publicações pré-print e suporte a publicação científica tradicional das revistas.

Também foi possível observar que os pesquisadores da física utilizam o arXiv. Sendo um repositório de arquivo para preprints eletrônicos de artigos científicos nos campos da matemática, física, ciência da computação, biologia quantitativa e estatística que podem ser acessados via internet. Em muitos campos da matemática e da física, quase todos os artigos estão no arXiv. No estudo, observou-se que 88,0% dos dados em preprint recuperados estavam relacionados a este repositório, ou seja, muito utilizados pelos pesquisadores da física.



Observa-se que na Figura 1, 20,40% dos DOIs dos artigos tem relacionamento com algum dataset ou repositório. Os dados foram gerados



Com base nos DOIs dos compartilhamentos permitiram identificar as fontes dos dados de pesquisas compartilhados, observou-se que a área de física e astronomia tiveram o maior depósito de dados no Cambridge Crystallographic Data Centre (CCDC) especializada em dados de química estrutural e desenvolvimento de materiais e ciências da vida para uso em descoberta farmacêutica.

O segundo repositórios de dados mais frequente é o Figshare, repositório científico online, aberto e gratuito que permite aos pesquisadores acadêmicos preservar e compartilhar todos os dados e resultados gerados em pesquisa, de documentos e procedimentos operacionais, padrões e protocolos. O Figshare foi criado em 2011 e é um dos vários negócios de portfólio apoiados pela Digital Science3, uma subsidiária da Springer Nature.

Com 12,26% dos depósitos de dados, o Centre de Données astronomiques de Strasbourg (CDS home), é um centro de dados que recolhe e distribui informações astronômicas. Foi fundado em 1972 com o nome de Centre de Données Stellaires. Os serviços on-line atualmente fornecidos pelo CDS são um banco de dados de objetos astronômicos, serviço para catálogos astronômicos, e atlas celeste interativo e banco de dados de imagens. Também foram identificados depósitos nos repositórios de dados do Zenodo (2,4%), Mendley Data (1,82%), Dryad (1,52%).

Considerações finais

Este estudo foi utilizado para avaliar a viabilidade e confiança nas informações do cruzamento de dados entre a publicação científica e os conjuntos de dados. Entretanto, observou-se que a prática de registrar DOIs em repositórios de preprint é recorrente pelos pesquisa da área da Física. Considera-se que o objetivo do estudo foi desenvolvido, entretanto necessita de refinamento na metodologia para melhor identificação das variáveis e as especificidades de cada área do conhecimento.