A integração do Arca - Repositório Institucional da Fiocruz com a Plataforma de Ciência de Dados aplicada à Saúde

São Paulo, 01 de agosto de 2019.

ICICT Instituto de Comunicação e Informação Científica e Tecnológica em Saúde

Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz

SUS

MINISTÉRIO DA SAÚDE
GOVERNO FEDERAL

Autores:
Claudete Fernandes de Queiroz – claudete.queiroz@icict.fiocruz.br,
Ana Maria Neves Maranhão - anamaranhao01@gmail.com
Luciana Danielli de Araujo - luciana.danielli@icict.fiocruz.br
Andrea F. Gonçalves do Nascimento - andrea.goncalves@icict.fiocruz.br
Raphael Belchior Rodrigues - raphael.rodrigues@icict.fiocruz.br
Éder de Almeida Freyre - eder.freyre@icict.fiocruz.br
Jefferson da Costa Lima - jefferson.lima@icict.fiocruz.br
Marcel de Moraes Pedroso - marcel.pedroso@icict.fiocruz.br

Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT)
Fundação Oswaldo Cruz - Fiocruz

Introdução

Estabelecimento de parceria entre a equipe do RI Arca e a equipe do Laboratório de Ciência de Dados da Fiocruz, que culminou no Projeto “Ciência de Dados aplicada ao Arca”, que estabeleceu os seguintes objetivos

- ✓ **Curadoria de dados:** identificação de inconsistências no preenchimento dos metadados do Arca, por meio da classificação automática utilizando *machine learning*, e consequente correção, visando qualidade das informações e dos dados extraídos, facilitando o trabalho de curadoria;

- ✓ **Recuperação da informação e visualização de dados:** oferece uma plataforma de exploração interativa para visualização e extração de dados, utilizando filtros e combinações de dados contidos no Arca, e que possam ser manipulados pelas diferentes unidades representadas no Repositório Institucional.

Ciência de Dados aplicada ao Arca

arca
Repositório Institucional da Fiocruz

Descrição:

O Arca é o Repositório Institucional da Fundação Oswaldo Cruz (Fiocruz) e sua função é reunir, hospedar, disponibilizar e dar visibilidade à produção intelectual da Instituição; visa estimular a mais ampla circulação do conhecimento, fortalecendo o compromisso institucional com o livre acesso da informação em saúde, além de conferir transparência e incentivar a comunicação científica entre pesquisadores, educadores, acadêmicos, gestores, alunos de pós-graduação, bem como a sociedade civil.

Objetivos do Projeto de Pesquisa, Inovação e Desenvolvimento Tecnológico:

- curadoria de dados: identificar inconsistências no preenchimento dos metadados do Arca, por meio da classificação automática utilizando *machine learning*, e consequente correção, visando qualidade das informações e dos dados extraídos;
- recuperação da informação e visualização de dados: oferecer uma plataforma de exploração interativa para extração e visualização de dados, utilizando filtros e combinações de dados contidos no Arca, como quantidade de produção por tipo de material, por unidade da Fiocruz, assunto, ano, entre outros, e que possam ser manipulados pelas diferentes unidades representadas no repositório institucional.

Equipe:

Claudete Fernandes de Queiroz e Luciana Danielli - Coordenadoras do Projeto

Pesquisadores e técnicos da Plataforma de Ciência de Dados aplicada à Saúde (PCDaS/Icict)

Fonte: https://bigdata.icict.fiocruz.br/ciencia-de-dados-aplicada-ao-arca

Problema detectado

- ✓ Alimentação descentralizada no Arca, sendo realizada por diversas Unidades, além do recurso de autoarquivamento, que tornou fundamental o monitoramento da qualidade dos dados preenchidos através da curadoria digital.

- ✓ Crescimento exponencial no número de depósitos, notadamente, após o estabelecimento da Política de Acesso Aberto ao Conhecimento no ano de 2014, em torno de 160%, sendo necessário e fundamental a utilização de mecanismos que facilitem a curadoria digital, a recuperação e a visualização do conteúdo disponibilizado.

Metodologia de trabalho

- ✓ Formalização de parceria entre Laboratório de Ciência de Dados e o Arca – Repositório Institucional da Fiocruz;
- ✓ Verificação das áreas que deveriam ser extraídas para compor a página de visualização de dados no Arca, como: ano de publicação, assunto, unidade/comunidade, tipologia, autor e direito autoral;
- ✓ Extração dos registros no DSpace, referentes as coleções de artigos, dissertações e teses (tipologias mandatórias);
- ✓ Estabelecimento de critérios para as variantes das palavras (plural e singular, sinônimos e homônimos), através da criação de uma tabela de equivalência visando reunir num universo delimitado os assuntos que apareciam com maior frequência no Arca;
- ✓ Identificação de inconsistências no preenchimento de alguns metadados, como, por exemplo, registros com mais de uma URI, que precisavam ser corrigidos;
- ✓ Disponibilização de uma página no Arca para visualização dos dados gerais extraídos, através de um *dashboard* com os metadados definidos (ano de publicação, assunto, unidade/comunidade, tipologia, autor e direito autoral).

Resultados e Discussões

- ✓ Implantação de uma rotina sistêmica no trabalho de curadoria dos dados no Arca, de forma que os gestores das Comunidades pudessem visualizar as informações a partir da extração dos registros relevantes;

- ✓ Identificação das inconsistências no preenchimento dos metadados, utilizando os sistemas Kibana e Elasticsearch para a classificação automática e correção dos dados, de forma padronizada;

- ✓ Criação de uma nuvem de tags com os assuntos mais indexados no Arca, destacando assim, a importância da indexação e do papel do Bibliotecário na gestão das informações;

- ✓ Realização de um trabalho colaborativo, promovendo a melhoria na qualidade dos metadados armazenados, a visualização de uma quantidade significativa de informações e a garantia de uma recuperação mais precisa;

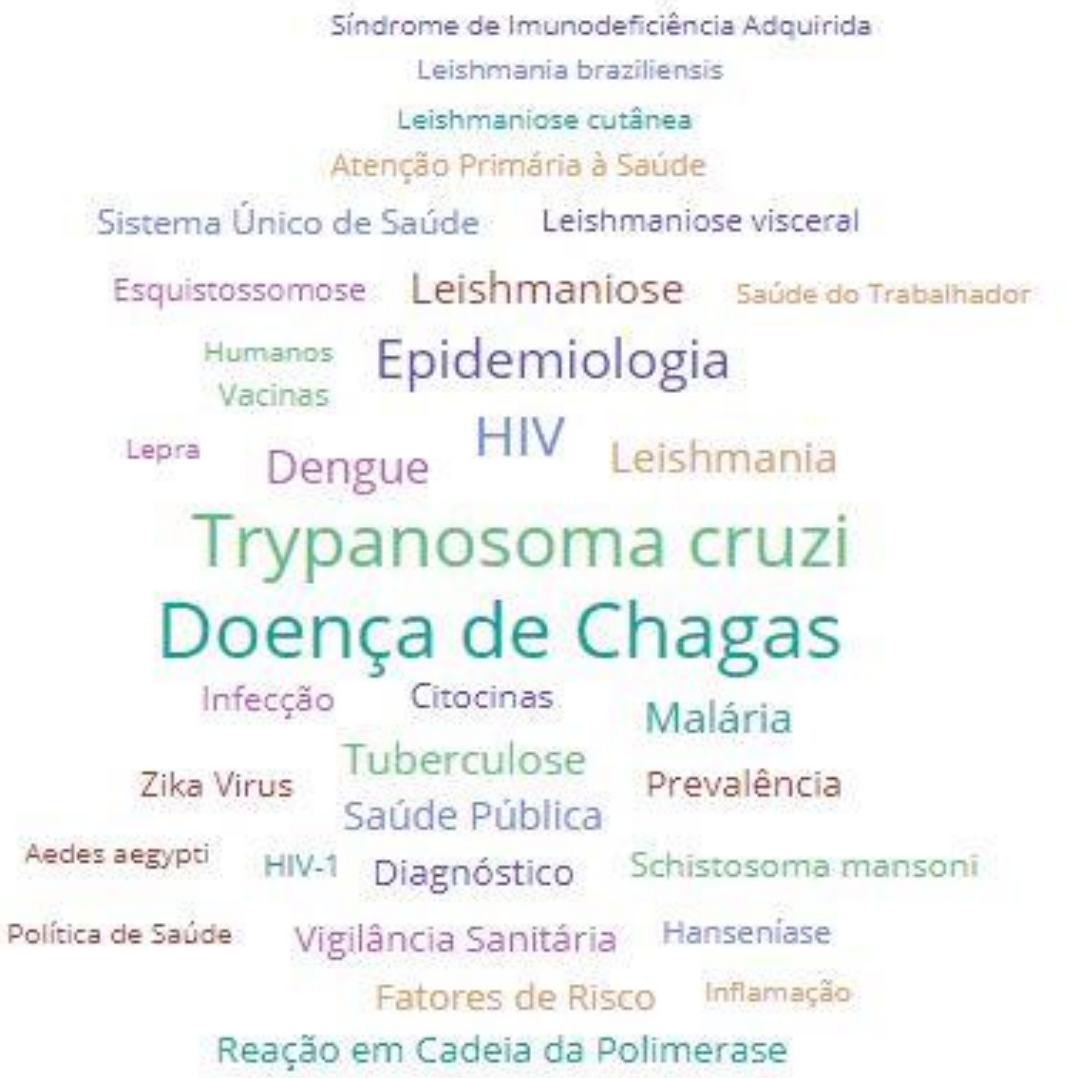
- ✓ Apresentação do resultado da parceria na página do Arca e nas reuniões e palestras ministradas.

Arca – Repositório Institucional da Fiocruz

Dashboard com dados gerais do Arca - Maio 2019
Coleções Mandatórias (Artigos, Dissertações e Teses)

Fonte: https://www.arca.fiocruz.br/terms/visualizacaodedados.jsp

Nuvem de Tags – Assuntos mais Indexados no Arca - maio 2019



Fonte: https://www.arca.fiocruz.br/terms/visualizacaodedados.jsp

Referências

FUNDAÇÃO OSWALDO CRUZ. **Ciência de Dados aplicada à Saúde**. Rio de Janeiro, 2019. Disponível em: https://bigdata.icict.fiocruz.br/Apresenta%C3%A7%C3%A3o. Acesso em 20 mar. 2019.

FUNDAÇÃO OSWALDO CRUZ. **Sobre o Arca**. Rio de Janeiro, 2019. Disponível em: https://www.arca.fiocruz.br/terms/sobre.jsp. Acesso em 10 abr. 2019.

MARANHÃO, Ana Maria Neves; DE QUEIROZ, Claudete Fernandes; RODRIGUES, Raphael Belchior. Curadoria Digital de Dados no Arca - Repositório Institucional da Fiocruz: relato de experiência. **RECIIS - Revista Eletrônica de Comunicação, Informação & Inovação em Saúde**, Rio de Janeiro, v. 11, p. 1-4, nov. 2017. Suplemento. Disponível em: https://www.arca.fiocruz.br/handle/icict/23725. Acesso em: 02 abr. 2019.

PEDROSO, Marcel de Moraes; LIMA, Jefferson da Costa; ASSEF NETO, Vinicius Belchior. Ciência de Dados aplicada ao Arca: desenvolvimento e disponibilização de ferramentas para recuperação da informação no Repositório Institucional da Fundação Oswaldo Cruz. **RECIIS - Revista Eletrônica de Comunicação, Informação & Inovação em Saúde**, Rio de Janeiro, v. 11, p. 1-5, nov. 2017. Suplemento. Disponível em: https://www.arca.fiocruz.br/handle/icict/23717. Acesso em: 02 abr. 2019.

SAYÃO, Luis Fernando; SALES, Luana Farias. **Guia de Gestão de Dados de Pesquisa para Bibliotecários e Pesquisadores**. Rio de Janeiro: CNEN/IEN, 2015. 90 p.

Nossos sinceros agradecimentos!

Equipe
Arca – Repositório Institucional da Fiocruz

Contato: repositorioarca@fiocruz.br
Telefones: (55 21) 3865-3271 / 3285

Fonte: Google

Instituto de Comunicação e Informação Científica e Tecnológica em Saúde

www.facebook.com/fiocruz.icict

twitter.com/@Icict_fiocruz

www.youtube.com/videosaudefio

www.icict.fiocruz.br

ICICT Instituto de Comunicação e Informação Científica e Tecnológica em Saúde

Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz

SUS

MINISTÉRIO DA SAÚDE GOVERNO FEDERAL