Eixo Temático: Repositórios Institucionais e Temáticos

Tipo de Trabalho: Comunicação Oral

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

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

Fundação Oswaldo Cruz, Instituto de Comunicação e Informação Científica e Tecnologia em Saúde, Rio de Janeiro, Brasil.

Resumo

Apresenta o projeto desenvolvido entre o Laboratório de Ciência de Dados aplicada à Saúde, do Instituto de Informação Científica e Tecnológica em Saúde (ICICT) e o Arca – Repositório Institucional da Fiocruz. O projeto teve como objetivo melhorar a curadoria de dados, através da identificação de inconsistências no preenchimento dos metadados, utilizando classificação automática e *machine learning*, e consequente correção, visando assim, garantir a qualidade das informações e dos dados extraídos. Outro fator importante para a realização do projeto foi a utilização do software Kibana e do Elastsearch para a visualização de dados de forma dinâmica, oferecendo uma plataforma de exploração interativa para extração e mineração de dados. O software permitiu a utilização de filtros e combinações de dados contidos no Arca, como produção por tipo de material, Unidades da Fiocruz, assunto, autor, ano e direito autoral de forma que possam ser manipulados pelas diferentes unidades/comunidades representadas no Repositório Institucional.

Palavras-chave: Arca - Repositório Institucional da Fiocruz. Ciência de Dados; Visualização de Dados. Curadoria Digital.

Introdução

Ciência de Dados é um campo que objetiva reunir um conjunto de estratégias, ferramentas e técnicas que combina métodos tradicionais de análise com algoritmos sofisticados para processar grandes volumes de dados em formatos diversos - dados estruturados, semiestruturados e não estruturados. Esse processo de análise, no âmbito da Ciência de Dados, envolve fases como coleta e ingestão; pré-processamento; análise exploratória; mineração de dados; e pós-processamento (PEDROSO, 2017).
Segundo Sayão e Sales (2015)

> o reconhecimento do potencial informacional dos dados de pesquisa para a ciência contemporânea transformou a visão que os caracterizava como simples subprodutos dos processos de pesquisa. Atualmente os pesquisadores, as instituições acadêmicas e as agências de fomento à pesquisa começam a compreender que esses dados, se devidamente tratados, preservados e gerenciados, podem constituir uma fonte inestimável de recursos informacionais para a pesquisa científica e para o ensino da ciência.

Nessa direção, uma gestão eficiente dos dados é fundamental para o desenvolvimento de pesquisas de alta qualidade e excelência, além de servir como suporte às ações de curadoria que objetivam aumentar a confiabilidade e a qualidade dos registros depositados em repositórios institucionais. Nesse sentido, destacamos o trabalho realizado pela Plataforma de Ciência de Dados aplicada à Saúde da Fiocruz, que é:

> fruto de projeto de pesquisa e desenvolvimento tecnológico do Laboratório de Informação em Saúde do Instituto de Comunicação e Informação Científica e Tecnológica em Saúde da Fundação Oswaldo Cruz (Lis/Icict/Fiocruz), que disponibiliza para a comunidade científica e gestores um serviço online de armazenamento, gestão e análise de dados em saúde, possibilitando o uso de estratégicas como análise visual, mineração de dados, big data, aprendizagem de máquina, dentre outras[1] (FUNDAÇÃO..., 2019).

O trabalho desenvolvido pelo Laboratório foi percebido pela equipe do RI Arca, que tem relatado um crescimento significativo no povoamento dos seus dados nos últimos anos, com um aumento consecutivo de aproximadamente 30%. Sendo assim, trabalhar com um grande volume de informação, requer habilidades e técnicas que se destacam pela capacidade de gerenciar grandes ou complexos sistemas, promovendo a qualidade das informações, consistência dos metadados, interação e integração entre bases de dados.

Desta forma, podemos afirmar que os Repositórios Institucionais são importantes ferramentas de gestão e não só de armazenamento e disseminação, pois permitem a recuperação e visualização dos dados ali contidos de forma dinâmica e objetiva, agregando imensurável valor às funções dos RIs.

[1] https://bigdata.icict.fiocruz.br/.

Ciente desta importância, foi estabelecida uma parceria entre a equipe do RI Arca[2] e a equipe do Laboratório de Ciência de Dados da Fiocruz, que culminou no Projeto "Ciência de Dados aplicada ao Arca[3]", que estabeleceu os seguintes objetivos:

✓ 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, facilitando o trabalho de curadoria iniciado em 2015;

✓ Recuperação da informação e visualização de dados: oferecer 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.

Metodologia

A metodologia proposta envolveu primeiramente a formalização de uma parceria e desenvolvimento de um Projeto entre o Laboratório de Ciência de Dados e o Arca – Repositório Institucional da Fiocruz, visando o estabelecimento de critérios e procedimentos que atendessem uma demanda pelo gerenciamento e visualização de dados contidos no RI.

Em seguida, a equipe definiu algumas etapas para o desenvolvimento do projeto. A primeira foi verificar quais as áreas que deveriam ser extraídas para compor a página de visualização de dados no RI Arca, como: ano de publicação, assunto, unidade/comunidade, tipologia, autor e direito autoral. A partir dessa informação, iniciou-se a extração dos registros no DSpace – sistema utilizado pelo Repositório, para arquivos no formato xml (padrão Dublin Core), referente as coleções de teses e dissertações, dos programas de pós-graduação da Fiocruz e dos artigos científicos, tipologias mandatórias da Política de Acesso Aberto ao Conhecimento da Instituição[4].

Após a extração dos registros, foram estabelecidos alguns critérios pela equipe, como reunir variantes das palavras – plural e singular, sinônimos e homônimos visando a criação de uma tabela de equivalência cujo propósito seria reunir num universo delimitado os assuntos que apareciam com maior frequência no RI Arca.

Outra etapa realizada foi a identificação de inconsistências no preenchimento de alguns metadados, como, por exemplo, registros com mais de uma URI, que precisavam ser corrigidos. Esses erros foram transformados em relatórios e encaminhados para cada Gestor de Comunidade do Repositório. Após a correção, foi realizada uma nova exportação seguindo os mesmos critérios para a verificação dos acertos descritos. Esse

[2] 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 (FUNDAÇÃO... 2019).
[3] https://bigdata.icict.fiocruz.br/ciencia-de-dados-aplicada-ao-arca
[4] https://portal.fiocruz.br/sites/portal.fiocruz.br/files/documentos/portaria_-_politica_de_acesso_aberto_ao_conhecimento_na_fiocruz.pdf

procedimento se tornou sistêmico tanto para a equipe do RI Arca quanto para as Comunidades, promovendo assim, um trabalho mais ágil e cooperativo em Rede.

Com a finalização dessas etapas, foi disponibilizada uma página no Repositório para a visualização dos dados gerais extraídos, através de um *dashboard* com os metadados definidos (ano de publicação, assunto, unidade/comunidade, tipologia, autor e direito autoral), conforme apresentado na Figura 1.

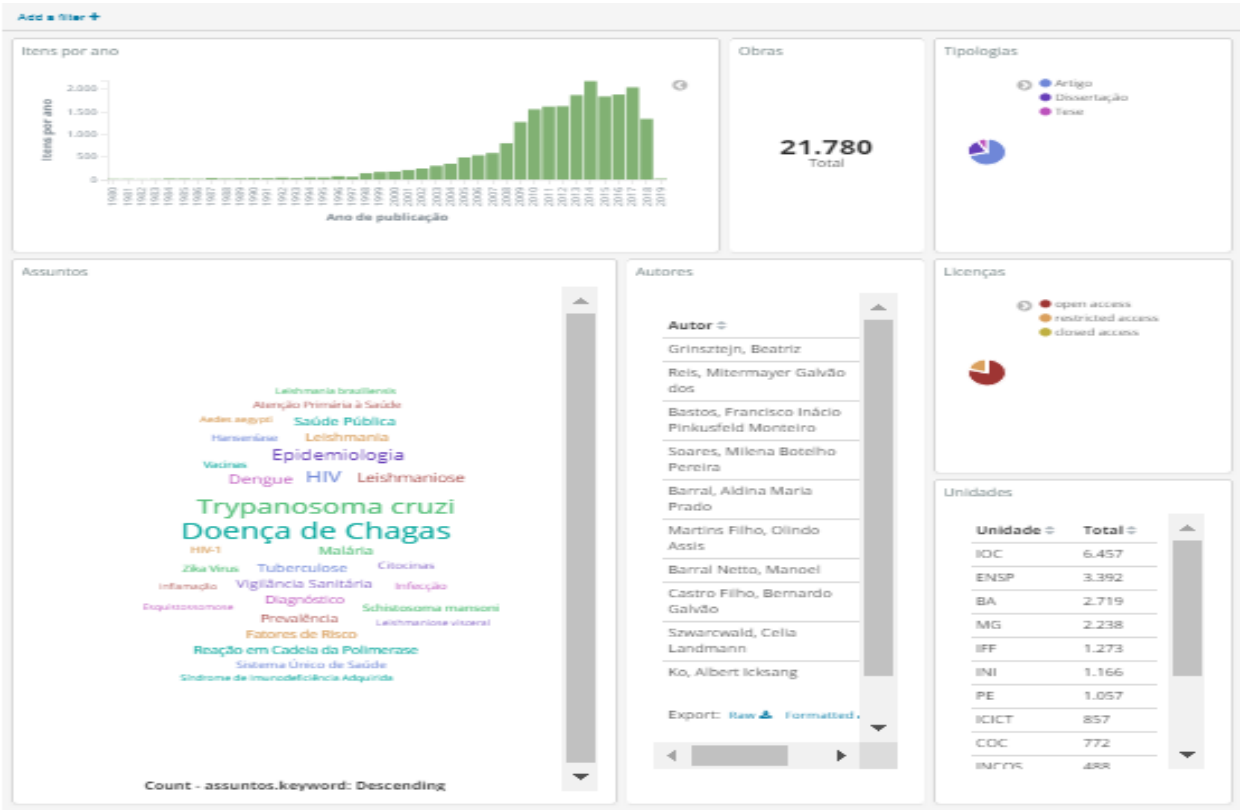

Figura 1: Dashboard com dados gerais do RI - Arca (abril 2019)

As informações obtidas foram visualizadas de forma dinâmica, permitindo não só a visualização geral, como também seria possível escolher uma Comunidade e visualizar quantos documentos foram publicados sobre determinado assunto, em um determinado ano.

No projeto foram utilizados os softwares Elasticsearch e Kibana - o primeiro trabalha com grandes volumes de dados e fornece uma API para a realização de análises dos dados recuperados, e o segundo é um *plugin*, que fornece recursos de visualização para os conteúdos indexados.

Problema

Tendo em vista que a alimentação no RI Arca é descentralizada, sendo realizada através das diversas Unidades e de suas Bibliotecas, além do recurso de autoarquivamento, se tornou fundamental o monitoramento da qualidade dos dados preenchidos através da curadoria digital. Dentro deste contexto, em 2015, estabeleceu-se um plano de ação para dar início ao trabalho de curadoria digital no RI Arca que tinha como principal objetivo firmar padrões visando a organização das informações e dos objetos digitais dentro do RI (MARANHÃO; QUEIROZ; BELCHIOR, 2017).

O 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%, tornou necessário e fundamental a utilização de mecanismos que facilitassem a curadoria digital, a recuperação e a visualização do conteúdo disponibilizado, permitindo assim, obter um panorama da produção científica institucional, tendo em vista que os RIs são, também, instrumentos de gestão.

Justificativa

A parceria firmada entre o RI Arca e a equipe do Laboratório de Ciência de Dados ajudou a complementar uma lacuna que existia no que se refere a curadoria digital e na gestão dos registros disponibilizados. Além disso, foi possível abordar de forma prática grandes quantidades de dados em diferentes formatos por meio de estratégias e técnicas relacionadas a Ciência de Dados.

Com o estabelecimento de diretrizes e procedimentos, foi viável criar uma interface amigável para a visualização dos dados contidos no RI Arca. Outro fator importante para a realização do projeto foi que seria possível estabelecer estratégias para a coleta, gestão e correção dos metadados descritos nas diferentes tipologias.

Resultados e Discussões

Implantação de uma rotina sistêmica no trabalho de curadoria dos dados no RI Arca, de forma que os gestores das Unidades técnico científicas da Fiocruz pudessem visualizar as informações a partir da extração dos registros relevantes. Também foi possível identificar as inconsistências no preenchimento dos metadados, utilizando os sistemas Kibana e Elasticsearch para a classificação automática e correção dos dados, de forma padronizada.

O sistema também possibilitou apresentar uma nuvem de tags com os assuntos mais indexados no Repositório[5], destacando assim, a importância da indexação e do papel do Bibliotecário na gestão das informações relevantes para o campo da Saúde e Pesquisa dentro da Fiocruz (Figura 2).

[5] É importante lembrar que RI Arca não reproduz necessariamente o que a Fiocruz produz, mas sim o que está depositado.

Figura 2: Nuvem de Tags – Visualização de Dados
(Assuntos mais Indexados no RI - Arca (abril 2019)

Podemos afirmar, portanto, que a aplicação do Projeto de Ciência de Dados pode propiciar, de forma colaborativa, a melhoria na qualidade dos metadados armazenados, promover a visualização de uma quantidade significativa de informações e garantir a recuperação mais precisa para os usuários do RI.

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.