IX CONFERÊNCIA INTERNACIONAL SOBRE BIBLIOTECAS E REPOSITÓRIOS DIGITAIS DA AMÉRICA LATINA
BIREDIAL-ISTEC 2019
SÃO PAULO - BRASIL 30·31 DE JULHO / 1·2 DE AGOSTO 2019

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

Claudete Fernandes de Queiroz[1], Ana Maria Neves Maranhão[2], Luciana Danielli de Araujo[3], Andrea F. Gonçalves do Nascimento[4], Raphael Belchior Rodrigues[5] , Éder de Almeida Freyre[6], Jefferson da Costa Lima[7] , Marcel de Moraes Pedroso[8]

[1] Fundação Oswaldo Cruz, Instituto de Comunicação e Informação Científica e Tecnologia em Saúde, Rio de Janeiro, Brasil. Email: claudete.queiroz@icict.fiocruz.br
[2] Fundação Oswaldo Cruz, Instituto de Comunicação e Informação Científica e Tecnologia em Saúde, Rio de Janeiro, Brasil. Email: anamaranhao01@gmail.com
[3] Fundação Oswaldo Cruz, Instituto de Comunicação e Informação Científica e Tecnologia em Saúde, Rio de Janeiro, Brasil. Email: luciana.danielli@icict.fiocruz.br
[4] Fundação Oswaldo Cruz, Instituto de Comunicação e Informação Científica e Tecnologia em Saúde, Rio de Janeiro, Brasil. Email: andrea.goncalves@icict.fiocruz.br
[5] Fundação Oswaldo Cruz, Instituto de Comunicação e Informação Científica e Tecnologia em Saúde, Rio de Janeiro, Brasil. Email: raphael.rodrigues@icict.fiocruz.br
[6] Fundação Oswaldo Cruz, Instituto de Comunicação e Informação Científica e Tecnologia em Saúde, Rio de Janeiro, Brasil. Email: eder.freyre@icict.fiocruz.br
[7] Fundação Oswaldo Cruz, Instituto de Comunicação e Informação Científica e Tecnologia em Saúde, Rio de Janeiro, Brasil. Email: jefferson.lima@icict.fiocruz.br
[8] Fundação Oswaldo Cruz, Instituto de Comunicação e Informação Científica e Tecnologia em Saúde, Rio de Janeiro, Brasil. Email: marcel.pedroso@icict.fiocruz.br

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 objetivos: melhorar a curadoria dos dados inseridos no repositório institucional, visando a qualidade das informações, e a recuperação e a visualização de dados, oferecendo uma plataforma que permite a extração de informações com potencial de uso pela gestão e pela pesquisa. No processo de curadoria foi possível identificar inconsistências no preenchimento dos metadados, utilizando classificação automática e *machine learning*, e consequente correção, de forma a 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 Elasticsearch 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.

Abstract

It presents the project developed between the Data Science Laboratory applied to Health, the Institute of Scientific and Technological Information in Health (ICICT) and the Arca - Institutional Repository of Fiocruz. The objective of the project was to improve the curation of data inserted in the institutional repository, aiming at the quality of information, and the retrieval and visualization of data, offering a platform that allows the extraction of information with potential use by management and research. In the curatorial process it was possible to identify inconsistencies in the metadata filling, using automatic classification and machine learning, and consequent correction, in order to guarantee the quality of information and data extracted. Another important factor for the realization of the project was the use of Kibana and Elasticsearch software to dynamically display data, offering an interactive exploration platform for data mining and extraction. The software allowed the use of filters and combinations of data contained in the Ark, such as production by material type, Fiocruz Units, subject, author, year and copyright so that they can be manipulated by the different units / communities represented in the Institutional Repository.

Keywords: Arca - Institutional Repository of Fiocruz. Data Science; Data Visualization. Digital Curatorship**.**

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.

Para além do desenvolvimento de pesquisas de alta qualidade e excelência, a gestão eficiente de dados permite ações de curadoria em repositórios institucionais, por exemplo, aumentando a qualidade e a confiabilidade dos registros depositados.

Na área de gestão do conhecimento em instituições de ensino e pesquisa, os Repositórios Institucionais são importantes ferramentas de gestão e não só de armazenamento e disseminação, pois permitem, com a utilização de técnicas e ferramentas adequadas, a recuperação e visualização dos dados ali contidos de forma dinâmica e objetiva, agregando-se imensurável valor às funções dos RIs.

O Repositório Institucional da Fiocruz – Arca tem como objetivo reunir e disponibilizar em um único local a produção intelectual produzida na Instituição e foi estabelecido como principal instrumento para realização da política de acesso aberto institucional. A Política de Acesso Aberto ao Conhecimento[1], implantada em 2014, estabelece como mandatório o depósito de teses, dissertações e artigos publicados, promovendo, também, desta forma, o alinhamento da Fiocruz com o movimento internacional de acesso aberto.

O Repositório atua, ainda, como uma rica fonte de informação para o desenvolvimento de novas linhas de pesquisas em todas as áreas da saúde pública, corroborando assim, com a missão da Fiocruz perante a sociedade:

> Produzir, disseminar e compartilhar conhecimentos e tecnologias voltados para o fortalecimento e a consolidação do Sistema Único de Saúde (SUS) e que contribuam para a promoção da saúde e da qualidade de vida da população brasileira, para a redução das desigualdades sociais e para a dinâmica nacional de inovação, tendo a defesa do direito à saúde e da cidadania ampla como valores centrais (FUNDAÇÃO OSWALDO CRUZ, 2019).

O Repositório foi criado em 2007, sendo lançado oficialmente como institucional em 2011. Está organizado em Comunidades que correspondem às Unidades Técnico-Científicas da Fiocruz e é mantido pelo Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT).

O RI utiliza o software livre DSpace[2] e sua alimentação é descentralizada, realizada pelas diversas unidades e seus respectivos setores e atores: bibliotecas, pesquisadores, etc., cabendo ao ICICT e a equipe técnica do Arca a gestão central.

Após o estabelecimento da Política de Acesso Aberto, o RI tem apresentado crescimento exponencial no número de depósitos realizados, em torno de 30%, conforme Tabela 1.

[1] https://portal.fiocruz.br/sites/portal.fiocruz.br/files/documentos/portaria_-_politica_de_acesso_aberto_ao_conhecimento_na_fiocruz.pdf

[2] No momento, estamos utilizando a versão 4.7, mas realizando testes para mudar para a versão 6.3.

Tabela 1: Crescimento Anual Arca (Atualizado até abril 2019).
Fonte: Fiocruz

No campo das Ciências de Dados, o ICICT, unidade técnico-cientifica que atua nas áreas de comunicação e informação para a saúde, na Fiocruz, vem desenvolvendo uma série de pesquisas e estudos, como o desenvolvido pelo Laboratório de Informação em Saúde (LIS), que criou e disponibiliza uma Plataforma para processamento de grandes volumes de dados.

A Plataforma de Ciência de Dados aplicada à Saúde da Fiocruz 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), em parceria com o Laboratório Nacional de Computação Científica (LNCC), 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[3] (FUNDAÇÃO OSWALDO CRUZ, 2019).

Dentro deste contexto, a principal proposta da plataforma é "coletar, processar e analisar informações por meio da Ciência de Dados que permitirá planejar, monitorar e avaliar políticas públicas e serviços de saúde em tempo real, gerando indicadores de alerta e painéis de monitoramento bastante específicos"[4] (FUNDAÇÃO OSWALDO CRUZ, 2019).

[3] https://bigdata.icict.fiocruz.br/
[4] A Plataforma irá proporcionar aos pesquisadores, docentes e discentes do Icict e de outras instituições acesso facilitado e qualificado a grandes quantidades de microdados.

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, percebeu-se a importância na formatação de uma parceria entre o RI-Arca e a Plataforma de Ciência de Dados Aplicada à Saúde, para definir estratégias e ferramentas que auxiliassem na coleta, transformação e análise dos dados disponibilizados no Repositório.

Essa parceria resultou no projeto "Ciência de Dados aplicada ao Arca" (Figura 1), com 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.

Figura 1: Fonte: https://bigdata.icict.fiocruz.br/ciencia-de-dados-aplicada-ao-arca

Metodologia

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

Foram desenvolvidas as seguintes etapas/atividades:

- estabelecimento das áreas que comporiam a página de visualização de dados, *dashboard*, do RI Arca, como: ano de publicação, assunto, unidade/comunidade, tipologia, autor e direito autoral;
- extração de todos os registros da base de dados, em formato xml, referente as coleções de teses e dissertações, dos programas de pós-graduação da Fiocruz, e dos artigos científicos publicados, tipologias mandatórias da Política de Acesso Aberto ao Conhecimento da Instituição[5];
- após a extração dos registros, estabelecimento de critérios visando a melhor visualização das informações, reunindo variantes das palavras – plural e singular, sinônimos e homônimos, e o corte em um número de frequência mínimo, visando reunir num universo delimitado os assuntos que apareciam com maior frequência no RI Arca;
- identificação de inconsistências no preenchimento de metadados, como, por exemplo, registros com mais de uma URI;
- emissão e envio de relatórios aos responsáveis pela alimentação nas diversas unidades institucionais para que realizassem as correções necessárias; (Figura 2);
- realizadas as correções, realização de nova exportação, seguindo os mesmos critérios para a verificação dos acertos descritos.

Cabe ressaltar que este último procedimento se tornou sistemático e foi incorporado as tarefas rotineiras, tanto para a equipe do RI Arca, quanto para as Comunidades que compõem o RI, promovendo assim, um trabalho mais ágil e cooperativo em Rede, com o estabelecimento de estratégias de coleta, gestão e correção dos metadados.

Repositório Institucional Arca
Casa de Oswaldo Cruz
Identificação de erros nos metadados na coleção – Artigos de Periódicos
Anos verificados 2010 a 2019

Identificador do Documento (handle)	Título do documento	Campo do metadado com problemas	Problema detectado	Solução
http://www.arca.fiocruz.br/handle/icict/12101	La "cultura de la sobrevivencia" y la salud pública internacional en América Latina: la Guerra Fría y la erradicación de enfermedades a mediados del siglo XX	dc.subject.en	As palavras chaves estão repetidas 4 vezes: international health; Cold War; Latin America; malaria; eradication	Remover as palavras repetidas
		dc.subject.es	As palavras-chaves estão sem padrão começando algumas com maiúscula e outras com minúscula	Padronizar as palavras-chaves, conforme descrito no manual de preenchimento de metadados
		dc.date.issue	Falta de padronização na data 15-Jan-2015	Corrigir para 2015
http://www.arca.fiocruz.br/handle/icict/10740	Uma abordagem arquivística: os documentos de um laboratório das ciências biomédicas	dc.identifier.citation	Falta de padronização v.19, n.1, jan.-mar. 2012, p.303-323.	Corrigir para v. 19, n. 1, p. 303-323, jan./mar. 2012.
http://www.arca.fiocruz.br/handle/icict/1474	Gestão do Conhecimento: ainda um obscuro objeto de desejo?	dc.identifier.citation	Falta de padronização SANTOS, P. X.; REIS, M. E. A. Gestão do conhecimento: ainda um obscuro objeto de desejo? RECIIS: Revista Eletrônica de Comunicação, Informação & Inovação em Saúde, Rio de	Corrigir para SANTOS, Paula Xavier; REIS, Maria Elisa Andries dos. Gestão do conhecimento: ainda um obscuro objeto de desejo? RECIIS: Revista Eletrônica de Comunicação, Informação & Inovação em Saúde,

Figura 2: Exemplo de Relatório após finalização da curadoria

[5] https://portal.fiocruz.br/sites/portal.fiocruz.br/files/documentos/portaria_-_politica_de_acesso_aberto_ao_conhecimento_na_fiocruz.pdf

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 3.

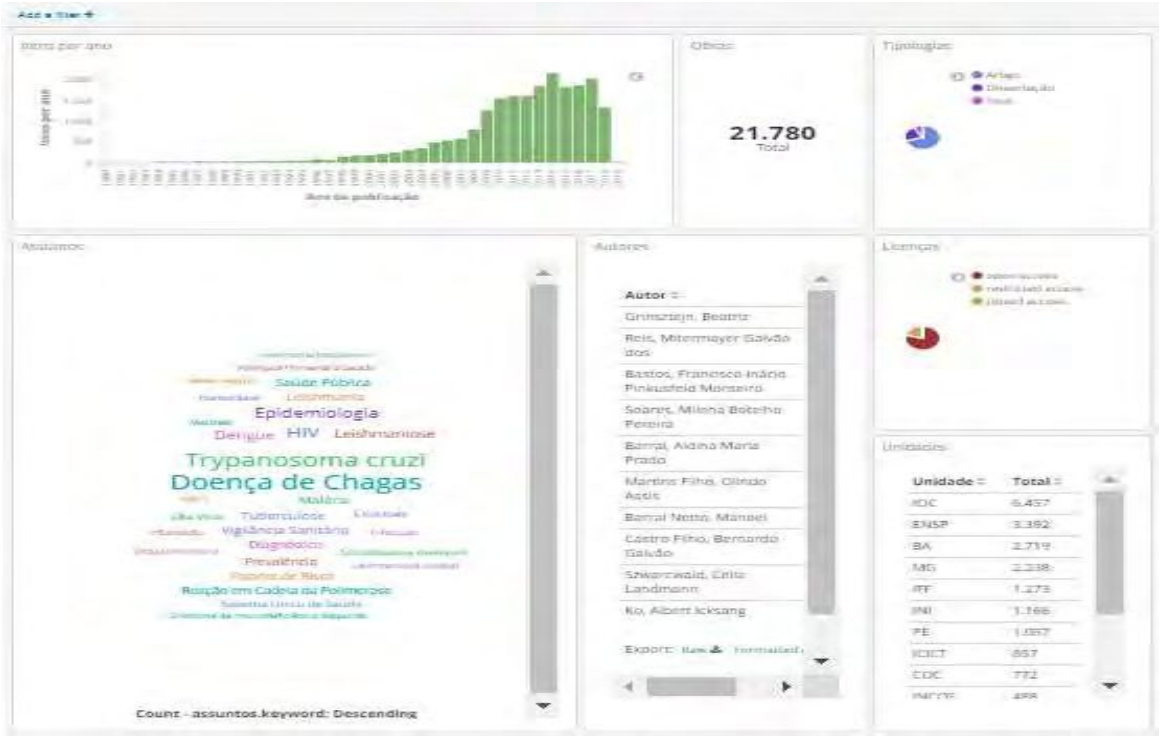

Figura 3: Dashboard com dados gerais do RI - Arca (abril 2019)

Na Figura 3 é possível visualizar as informações de forma dinâmica, permitindo uma visão geral da produção Institucional, como também selecionar 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% (período 2014-2018), 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 (Figura 4).

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, oferecendo aos gestores das Comunidades do Repositório, ferramentas que facilitassem e agilizassem os acertos em um tempo mais viável.

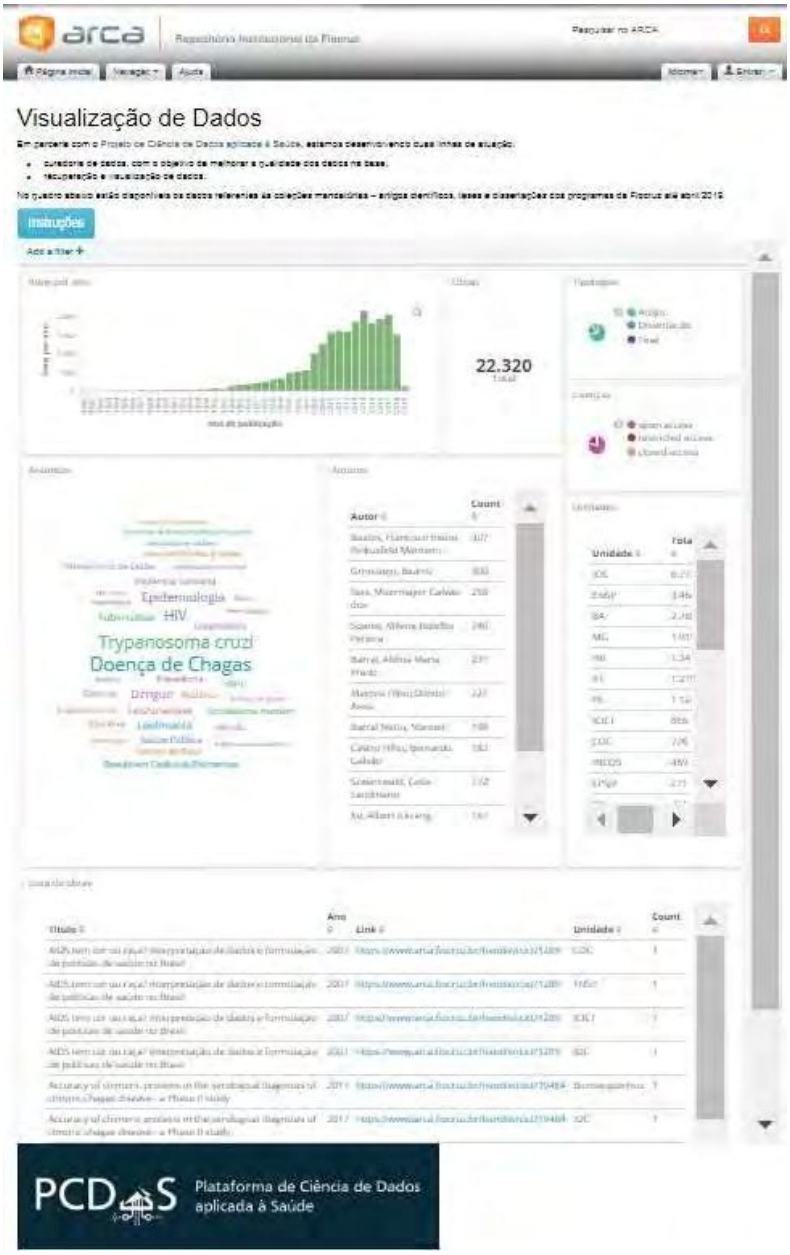

Figura 4: Visualização de Dados no Arca (abril 2019)

Resultados e Discussões

Implantação de uma rotina sistemática 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[6], 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 5).

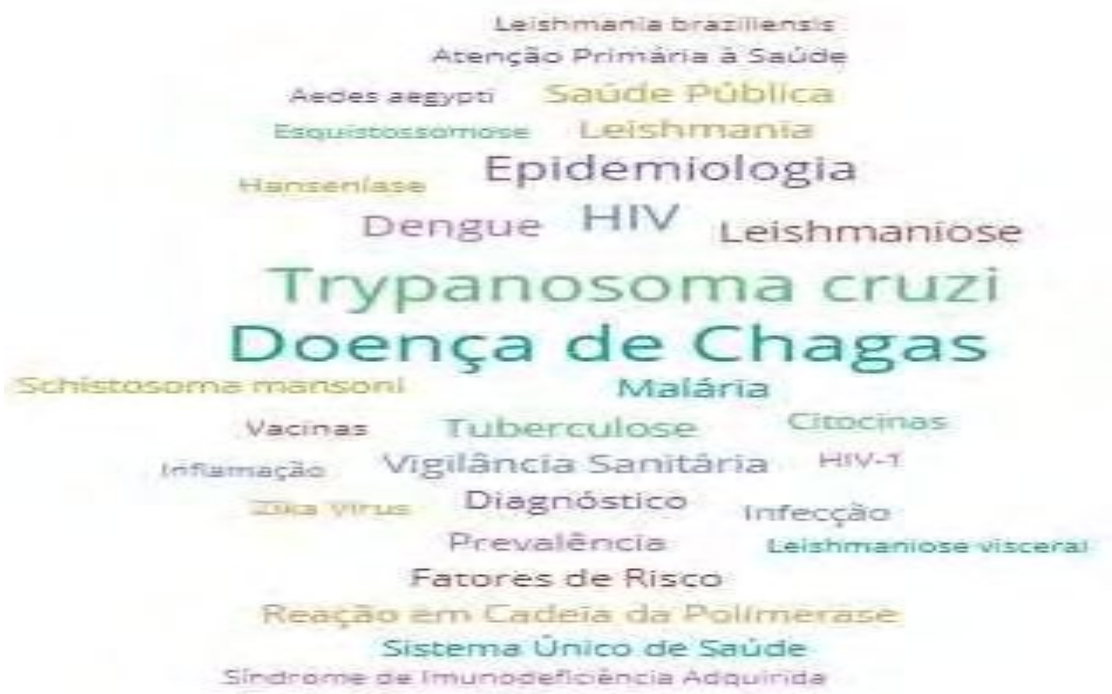

Figura 5: Nuvem de Tags – Visualização de Dados
(Assuntos mais Indexados no RI - Arca (abril 2019)

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.

Outro fator importante para a implantação do projeto foi que através da integração dos sistemas DSpace, Kibana e Elasticsearch, foi possível oferecer soluções de análise, mineração e visualização de dados, através da extração de uma considerável quantidade de registros, proporcionando assim, o acesso aos resultados estruturados e quantificados numa interface mais amigável para o usuário final.

Referências

FUNDAÇÃO OSWALDO CRUZ. **Ciência de Dados aplicada à Saúde**. Rio de Janeiro, 2019. Disponível em:

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

<https://bigdata.icict.fiocruz.br/Apresenta%C3%A7%C3%A3o>. Acesso em 20 mar. 2019.

FUNDAÇÃO OSWALDO CRUZ. **Perfil institucional**. Rio de Janeiro, 2019. Disponível em: <https://portal.fiocruz.br/perfil-institucional>. 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.