WIDaT 2018
II WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA

A CONSTRUÇÃO DO REPOSITÓRIO DE DADOS DA UFPB: a experiência com o Dataset de Arboviroses

Pollianna Marys de Souza e Silva[1]
Sandra de Albuquerque Siebra[2]
(1) UFPB. pollianna_marys@hotmail.com
(2) UFPE. sandra.siebra@gmail.com

PPGCi UFPB
Programa de Pós-Graduação em Ciência da Informação

INTRODUÇÃO

- O movimento em favor do acesso aberto (*Open Access*) surgiu a partir da crise dos periódicos e permitiu a democratização do acesso à informação, sendo os repositórios digitais uma das primeiras plataformas digitais de acesso aberto (SILVA JÚNIOR; BORGES, 2014).

- Os repositórios são ambientes digitais que possibilitam reunir dados e informações de cunho científico, administrativo, técnico, artístico, cultural, entre outros, cuja função principal é promover a visibilidade de seus objetos digitais, preservando-os por meio do gerenciamento de informação (ABREU; VIDOTTI, 2016).

- Os repositórios de dados (RD) garantem os princípios de transparência e oferecem um sistema de armazenamento seguro, além da possibilidade de se ter os dados de pesquisa disponíveis on-line, indexados, documentados, p/serem acessados, baixados, visualizados e processados por pessoas ou por sistemas, estendendo-os a uma comunidade mais ampla e conectada em rede (SAYÃO; SALES, 2016).

INTRODUÇÃO

- De fato, os RDs têm sua importância c/recurso informacional e se tornam um dispositivo de troca de experiências e compartilhamento de dados científicos, além de parte essencial das infraestruturas mundiais de pesquisa em escala global, tornando visível e aberta p/toda a sociedade uma parcela importante da atividade de pesquisa, caracterizando a chamada Open Science ou Ciência Aberta (FORMENTON, 2015).

- Os RDs, de forma diferente das publicações acadêmicas que falam por si próprias, precisam ter explícitos os seus conteúdos, para poder revelar e transmitir conhecimento no tempo e no espaço, de forma que os dados possam ser interpretados, sintetizados e reanalisados em contextos diversos, com finalidades diferentes das quais foram gerados e coletados originalmente (SAYÃO; SALES, 2016).

OBJETIVO

- Relatar a experiência de criação do 1º RD da UFPB, fazendo uso da plataforma Dataverse, com foco no Dataset de Arboviroses;

- Este RD surgiu após a realização do projeto de pesquisa - Chamada Universal - MCTI/CNPq (Número 01/2016) – intitulado “A Ciência da Informação e a Disseminação de Informações Associadas à Epidemia de Zika Vírus: uma investigação baseada na Análise de Redes Sociais”;

- A motivação p/o projeto de pesquisa surgiu após o Brasil vivenciar, em 2016, um surto de microcefalia em bebês recém-nascidos, causada pela contaminação de mulheres grávidas com o vírus zica. Sendo que para a criação do dataset decidiu-se expandir para um conjunto de patologias formado principalmente pela Zica, Dengue e Chikungunya, denominadas arboviroses. Isso porque essas são doenças com graves repercussões p/a saúde da população conforme indicadores operacionais e epidemiológicos.

OBJETIVO

Evolução da Doença no Brasil

1981-1982	1986	2018
Boa Vista - Roraima	Rio de Janeiro e Capitais do Nordeste	Em todo os estados brasileiros

(BRASIL, 2017)

PROCEDIMENTOS METODOLÓGICOS

- A pesquisa que originou esse artigo trata-se de uma pesquisa-ação, qualitativa e descritiva;

- O dataset foco desse relato é um conjunto de dados composto por posts da rede social Twitter sobre as arboviroses, que engloba a Zyca, Dengue e Chikungunya;

- A coleta de dados para compor o dataset foi realizada no período de **outubro de 2017 a março de 2018**. Os dados foram coletados por um script feito na linguagem de programação Ruby, fazendo uso de API (Application Programming Interface) disponibilizada pelo próprio Twitter. Os posts coletados foram os que apresentaram uma ou mais das seguintes palavras, desconsiderando maiúsculas e minúsculas: **zica, zika, zyca, zkv, zikav, dengue, dengue hemorrágica, chikungunya, chicungunya, arbovirose, arboviros e/ou microcefalia**;

PROCEDIMENTOS METODOLÓGICOS

- Os posts identificados, mais de um milhão, foram baixados no formato JasonB e categorizados considerando 3 grupos: **Zyca, Dengue e Chikungunya**. Em seguida, foi criado um banco de dados modelado para as normas do Dataverse, onde os dados extraídos foram efetivamente armazenados. Ressalta-se que, p/poder contextualizar e adicionar valor aos dados, foi criado um conjunto de metadados descritivos p/sintetizar os diferentes contextos em que as mensagens foram utilizadas e compreender c/os usuários da rede social discutiram/escreveram sobre a temática;

- O Dataverse foi escolhido c/plataforma por ser uma das mais popularmente utilizadas p/a criação de RDs.

RESULTADOS

- O RD da UFPB foi disponibilizado em 2018, no endereço **https://dataverse.ufpb.br/dataverse/root;**

- A equipe inicial de pesquisadores em uma primeira etapa precisou se familiarizar com o que é e c/funciona um RD, o que foi sanado com a pesquisa bibliográfica. Posteriormente, sobre a plataforma Dataverse;

- Como toda documentação da plataforma estava em inglês, a equipe se dedicou a traduzir e destacar as principais partes da documentação, antes de começar a implantação da plataforma na instituição;

- Uma vez a plataforma instalada, ela foi populada com o dataset de arboviroses extraído do twitter;

RESULTADOS

- Para um melhor gerenciamento de acesso e segurança, buscou-se restringir o acesso aos dados a pesquisadores previamente cadastrados, cujas credenciais de acesso podem ser feitas a partir do cadastro individual do pesquisador no Dataverse, utilizando uma conta do Gmail;

- Assim, optou-se por permitir acessar o arquivo restrito do dataset apenas após login e senha c/as credenciais de acesso adequadas ao sistema e, durante download do arquivo do dataset, o sistema deve solicitar a aceitação de um termo de acesso e uso;

RESULTADOS

- O repositório criado a partir da plataforma Dataverse possui a identificação especificada abaixo:

Dataverse	Departamento de CI
Identifier	https://dataverse.ufpb.br/dataverse/dci
Category	Department
Affiliation	CCSA - UFPB
Description	Este repositório concentra dados de pesquisas do Departamento de Ciências da Informação (DCI) do Centro de Ciências Sociais Aplicadas (CCSA) da Universidade Federal da Paraíba (UFPB).

RESULTADOS

- No Dataverse, os campos de metadados são escolhidos para uso em cada conjunto de dados (dataset) a serem adicionados (THE DATAVERSE PROJECT, 2018). Considerando a natureza multidisciplinar das pesquisas do Departamento de CI, os pesquisadores optaram pelo uso de campos de metadados gerais, em um padrão que pudesse ser aplicado em dados das diversas áreas do conhecimento. Assim, os principais campos a serem preenchidos no Dataverse para o dataset são: título, autor(es), informações para contato, descrição resumida da pesquisa, data da coleta dos dados no RD, tópicos da pesquisa, palavras chave, entre outros. Esses metadados tanto contextualizam e descrevem os dados armazenados, como facilitam a sua recuperação;

- A preservação e acesso a longo prazo são garantidos no Dataverse pela identificação persistente, que protege os documentos digitais com mecanismos que preveem a obsolescência dos dados - migração dos dados para um software mais recente e a prescrição que consiste em guardar o conjunto de bytes para serem consultados quando for necessário. (DATAVERSE PROJECT, 2018).

CONSIDERAÇÕES FINAIS

- As arboviroses são enfermidades tropicais endêmicas, que merecem receber atenção especial dos profissionais de saúde que atuam na atenção básica e na vigilância em saúde e dos gestores das esferas federal, estadual e municipal.

- Elas incapacitam ou matam milhões de pessoas e representam uma necessidade médica importante. Assim, o estudo sobre a disseminação de informações sobre essas enfermidades, tanto nos círculos acadêmicos e científicos formais, como nos informais, apresenta inúmeras oportunidades de investigação para pesquisadores. É nesta vertente que a proposta do Dataset Arboviroses da UFPB está direcionada, contribuindo com dados brutos para o campo científico e para a sociedade diante da urgência da temática.

- Podendo-se investigar: **O que a população sabe sobre as arboviroses? Que tipo de dúvidas possuem? Como a informação tem chegado até a sociedade? Que tipo de queixas são realizadas sobre as doenças em questão? Que localidades mais discutem sobre arboviroses?** Além de ser possível mapear casos de relatos de morte, surtos e agravamento das doenças por meio dos posts coletados.

CONSIDERAÇÕES FINAIS

- Espera-se, como trabalhos futuros, poder relatar os usos feitos do Dataset de Arboviroses tanto pelos administradores do Dataverse, como por pesquisadores cadastrados na plataforma.

REFERÊNCIAS

ABREU, J. P.; VIDOTTI, S. A. B. G. Curadoria Digital nos Contexto dos Repositórios Digitais. In: Encontro Internacional de Dados, Tecnologia e Informação, 2., 2016. Marília. **Anais…** Marília: UNESP, 2016.

BALDISSERA, A. Pesquisa-Ação: uma metodologia do conhecer e do “agir” coletivo. **Sociedade em Debate**, Pelotas, v. 7, n. 2, p. 5-25, agosto, 2001. Disponível em: <http://revistas.ucpel.edu.br/index.php/rsd/article/viewFile/570/510>. Acesso em: 02 ago. 2018.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Coordenação-Geral de Desenvolvimento da Epidemiologia em Serviços. **Guia de Vigilância em Saúde**: volume 2. Brasília: Editora do Ministério da Saúde, 2017.

DATAVERSE PROJECT. **About Dataverse**. Disponível em: <https://dataverse.org/>. Acesso em: 08 jun. 2018.

FORMENTON, D. **Identificação de Padrões de Metadados para Preservação Digital**. São Carlos: UFSCar, 2015. 102 f. Dissertação de Mestrado - Universidade Federal de São Carlos. Disponível em: <https://repositorio.ufscar.br/bitstream/handle/ufscar/7221/DissDF.pdf?sequence=1>. Acesso em: 02 jun. 2018.

REFERÊNCIAS

GOMEZ, M. N. G. O Domínio das Informações em Saúde. In: PINTO, V. B.; CAMPOS, H. H. (Org). **Diálogos Paradigmáticos Sobre Informação para a Área da Saúde.** Fortaleza: Edições UFC, 2013.

MICHEL, M. H. **Metodologia e Pesquisa Científica em Ciências Sociais**. 2 ed. São Paulo: Atlas, 2009.
RICE, R; SOUTHALL, J. **The Data Librarian's Handbook.** London: Facet, 2016. 169p.

SAYÃO, L. F.; SALES, L. F. Algumas Considerações Sobre os Repositórios de Dados de Pesquisa. **Informação & Informação**, Londrina, v. 21, n. 2, p. 90 – 115, maio/agosto, 2016. Disponível em: <http:www.uel.br/revistas/informacao/90>. Acesso em: 12 jul. 2018.

SIEBRA, S. A.; BORBA, V. R.; MIRANDA, M. J. K. F. O. Curadoria digital: um termo interdisciplinar. In: XVII Encontro Nacional de Pesquisa Em Ciência da Informação (ENANCIB), 17., 2016, Salvador. **Anais...** Salvador, BA: UFBA, 2016. Disponível em: <http://www.ufpb.br/evento/lti/ocs/index.php/enancib2016/enancib2016/paper/view/4107/2559>. Acesso em: 2 jul. 2018.

SILVA JUNIOR, L. P.; BORGES, M. M. Preservação digital no Repositório Científico de Acesso Aberto de Portugal. **Rev Eletrônica de Comun. Inf. Inov. Saúde**, v. 8, n. 4, p. 567-574, out./dez. 2014. Disponível em:<http://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/911>. Acesso em: 04 ago. 2018.