X Seminário de Extensão e Inovação
XXV Seminário de Iniciação Científica e Tecnológica
23 a 27 de Novembro | Toledo - PR
SEI 2020 SICITE UTFPR - CÂMPUS TOLEDO
UTFPR UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ CÂMPUS TOLEDO

https://eventos.utfpr.edu.br//sei/sei2020

Uma plataforma para facilitar o acesso aos dados do Repositório Brasileiro Livre para Dados Abertos do Solo

A platform to facilitate the access to data from the Brazilian Free Repository for Open Soil Data

Marcos Alexandre dos Anjos
marcosanjos@alunos;utfpr.edu.br
Discente do 6° período do Curso de Ciência da Computação. Universidade Tecnológica Federal do Paraná, Santa Helena, PR, Brasil.

Alessandro Samuel Rosa
alessandrorosa@utfpr.edu.br
Docente do Curso de Agronomia, Universidade Tecnológica Federal do Paraná, Santa Helena, PR, Brasil.

RESUMO

Neste artigo apresentamos os principais sistemas de gerenciamento de dados e metadados em repositórios para o Repositório Brasileiro Livre para Dados Abertos do Solo (FEBR). Os métodos e procedimentos consistiu numa pesquisa bibliográfica sobre sistemas de catalogação e gerenciamento de dados e metadados da pesquisa, usando o Registry of Research Data Repositories (re3data.org) como ponto de partida. No que diz respeito aos sistemas de catalogação e gerenciamento de dados e metadados, segundo o re3data.org, são três as alternativas mais populares: CKAN, DataVerse e DSpace. Todos os três são amplamente utilizados em instituições nacionais e internacionais, não havendo grandes diferenças em eles. A análise destaca as funcionalidades sobre cada software apresentando recursos para construção de um repositório de dados de pesquisa com objetivo do compartilhamento dos dados. Sendo elaborado critérios para análise dos softwares em destaque os seguintes tópicos: requisitos para instalação, finalidade e uso do software e principais limitações de cada software. Dentre os softwares analisados, concluímos que o DataVerse é o mais indicado para adoção no repositório FEBR.

PALAVRAS-CHAVE: Metadados. CKAN. DataVerse. DSpace.

ABSTRACT

In this article we present the main data and metadata management systems in repositories for the Brazilian Free Repository for Open Soil Data (FEBR). The methods and procedures consisted of a bibliographic search on research data and metadata management and cataloging systems, using the Registry of Research Data Repositories (re3data.org) as a starting point. With regard to data and metadata cataloging and management systems, according to re3data.org, there are three most popular alternatives: CKAN, DataVerse and DSpace. All three are used in national and international institutions, with no major differences between them. The analysis highlights the functionalities of each software, presenting the resources needed for building a research data repository with the objective of sharing the data. The criteria for the evaluation of the softwares were as follows: installation requirements, scope and common use of the software, and main limitations. Among the softwares evaluated, we conclude that DataVerse is the most suited to be adopted by the FEBR repository.

KEYWORDS: Metadata. CKAN. DataVerse. DSpace.

Recebido: 19 ago. 2020.
Aprovado: 01 out. 2020.

INTRODUÇÃO

A comunicação científica é a principal maneira de compartilhar o conhecimento com a sociedade, e as formas de disseminação do conh\ecimento passaram por mudanças ao longo dos anos. As tecnologias digitais ajudaram no compartilhamento dos dados científicos, o que reduziu as dificuldades de publicação e acesso aos documentos via internet (BIZER; HEATH; BERNERS-LEE, 2011). Esse desenvolvimento tecnológico impulsionou as publicações na web para a comunidade, ganhou visibilidade e tornou-se uma questão essencial para o desenvolvimento e aplicação das atividades de pesquisa científica.

Os primeiros sinais de um movimento de publicação de dados científicos abertos vieram de Paul Ginsparg, em 1991, que criou um servidor de dados do laboratório de pesquisa para que os pesquisadores pudessem depositar os dados (SOUZA, 2012). Com isso, o movimento de acesso aberto aos dados fez com que surgisse a demanda para disponibilização dos dados gerados em publicações para a sociedade. Com essa relevância do compartilhamento dos dados, uma maneira de organizar os dados científicos disponíveis são através de repositórios institucionais, que são gerenciados por sistemas de gerenciamento de publicações (WALPORT; BREST, 2011). O Registry of Data Repositories (RE3DATA, 2020) é um registro, assim lista os repositório de dados de pesquisa que abriga diversas áreas do conhecimento, sendo um registro global que oferece informações detalhadas sobre diversos repositórios de dados.

O Repositório Brasileiro Livre de Dados Abertos do Solo (FEBR) armazena os dados de pesquisa sobre os solos brasileiros. Seu objetivo é a preservação, padronização e versionamento dos dados depositados, facilitando sua reusabilidade pela comunidade. Essa dificuldade de padronização na divulgação dos dados de pesquisas científicas nos repositórios levou a criação da Open Archives Initiative (OIA), uma iniciativa para desenvolvimento de padrões de interoperabilidade entre repositórios digitais, no qual a interoperabilidade está relacionada com os metadados que são os atributos que descrevem os dados. Esse padrão impacta diretamente na interoperabilidade entre os repositórios digitais e tem o propósito de promover o acesso simultâneo aos dados contidos em diversos repositórios, de forma a facilitar as buscas.

A informação e conhecimento científico necessita ser de livre acesso, mas é preciso que os repositórios utilizem meios que possibilitam o acesso das informações e de todo conhecimento. Esse desafio encontra se escolha de um sistema de gerenciamento de metadados para os conjuntos dados abertos para repositório dados FEBR. O levantamento das características funcionais de um software, impacta na qualidade final do produto, pois o mesmo passa pela estruturação adequada e pelo controle de qualidade das funcionalidades atendidas (SILVA, 2011). Dessa forma as características funcionais descrevem exatamente o comportamento esperado do sistema.

Neste contexto o projeto foi sugerido e submetido com o intuito de apresentar qual melhor software para gerenciamento de metadados de conjuntos de dados de pesquisa para FEBR.

MATERIAL E MÉTODOS

Os sistemas de catalogação e gerenciamento de metadados a serem avaliados foram definidos a partir de consulta ao re3data.org. Atualmente, os softwares mais utilizados para gerenciamento de repositórios de acesso aberto são DataVerse, DSpace, CKAN, MySQL, Fedora e EPrints. Dentre estes, os três mais populares (CKAN, DataVerse e DSpace) foram selecionados para uma análise mais detalhada. Os demais softwares não foram avaliados devido a sua finalidade específica de gerenciamento de banco de dados e/ou por uma questão de limitação de recursos disponíveis.

O Comprehensive Knowledge Archive Network (CKAN) foi desenvolvido pela Open Knowledge Foundation com o objetivo de compartilhamento e visualização de dados. O DataVerse foi desenvolvido pela universidade de Harvard com o objetivo de compartilhamento, preservação e exploração de conjuntos de dados. O DSpace surgiu de uma demanda da biblioteca do Massachusetts Institute of Technology (MIT) e dos laboratórios da Hewlett-Packard com o propósito de desenvolvimento de repositórios e bibliotecas digitais.

A avaliação dos sistemas de gerenciamento de metadados foi realizada em duas fases. A primeira consistiu num estudo bibliográficos ao tema e documentações oficiais dos sistemas de gerenciamento de metadados. Segunda fase incluiu a participação em eventos técnico-científicos relacionados ao tema e a consulta (por correio eletrônico) a, profissionais com experiência de trabalho com os softwares. Para esta atividade foram levantadas as seguintes características:

a) Requisitos básicos de instalação e suporte;
b) Qual a finalidade do software;
c) Depósito de conjunto de dados no software;
d) Principais limitações do software.

RESULTADOS E DISCUSSÃO

A pesquisa tem como principal objetivo analisar as características dos softwares de gerenciamento de dados de modo a escolher um software para o gerenciamento dos dados no repositório FEBR. Tendo em vista esse objetivo, foram levantados quatro características, das quais estão relacionadas aos requisitos de instalação de cada software, suporte, sua finalidade de uso e suas limitações, das quais julga serem relevantes para repositório FEBR.

Observando os resultados na tabela 1, em que os três softwares apresentam seus requisitos mínimos de hardware para o servidor e na parte de software a semelhança está relacionada ao sistema operacional que todos tem suporte para instalação no ambiente Linux. Destaca os requisitos básicos de instalação sobre cada software e suporte oferecido e tipo licença integrada ao software, esta licença está relacionado modificações e customização de cada software.

Tabela 1 – Requisitos básicos instalação e suporte para cada software.

Software	Descrição	Licença	Suporte
CKAN	SO: Linux Debian/RedHat; HD: 25 GB*; Memória ram: 4 GB*; Processador: 4 núcleos 2 GHz*.	*Licença Affero GPLv3:* uma licença para softwares open source sendo necessário toda alteração deve sempre documentar e deixar público o documento.	Comunidade oficial no Google; Fórum stack overflow; Disponibilidade suporte pago.
DSpace	SO: Linux Debian; HD: 48 GB*; Memória ram: 4 GB*; Processador: 2 núcleos 2,8 GHz*.	*Creative Commons Attribution 4.0:* uma licença para modificação do software e podendo distribuir uma versão DSpace customizada para qualquer fim, mesmo comercial.	Comunidade oficial no Google; Site para código do projeto no GitHub; Disponibilidade suporte pago.
Data Verse	SO: Linux Debian/Windows; HD: 25 GB*; Memória ram: 8 GB*; Processador: 2 núcleos 2 GHz*.	*Apache Licence Version 2.0:* uma licença flexível para customização no software não obrigatório a documentação das alterações realizadas no software.	Disponibilidade apenas na comunidade oficial no Google; Fórum stack overflow;

Fonte: Autoria própria (2020).

De uma forma simplificada foram pesquisados o funcionamento de registro e publicação dos conjuntos de dados no repositório nos três softwares apresentam semelhanças. O usuário deve-se estar registrado na plataforma do repositório para depositar os dados numa área específica e informando qual categoria os dados pertencem. Após o usuário enviar os dados pela plataforma o mantenedor do repositório receberá os dados para verificação dos dados de forma manual e analisando se estão nas normas definidas pelo repositório. Após este procedimento, o mantenedor do repositório pode aprovar ou rejeitar os dados dessa forma mantendo uma maior integridade dos dados depositados. Quando autor dos dados enviar dos dados o software irá relacionar os metadados do autor e atribuir ao conjunto dos dados, assim mantendo histórico e controle dos dados.

Após a pesquisa dos usuários como irão depositar os dados, notou-se que os três softwares apresenta semelhanças tais como cadastro do usuário, na inserção dos conjuntos de dados, preenchimento dos metadados conforme a descrição da documentação. Lembrando que foram analisando as configurações padrão de cada software, sem nenhuma customização de sistema.

Mediante no levantamento dos dados obtidos, foi possível estabelecer um comparativo entre os softwares na tabela 2, de forma que CKAN apresenta-se

como um software desenvolvido para abertura e publicação dos dados relacionados ao governo. O ponto fraco do software não apresenta recurso para versionamento dos dados, sendo um dos pontos cruciais para FEBR.

Tabela 2 – Coleta de opiniões sobre os softwares.

Software	Escopo e Uso	Pontos Fortes	Pontos Fracos
CKAN	Escopo: Repositório de dados; Uso predominante no governo.	Plugin visualização de dados espaciais; Customização do layout.	Apresenta grau de Manutenção elevado; Sem opção para versionamento dados.
DSpace	Escopo: Repositórios dados e publicações; Uso com mesma frequência no governo e universidades.	Apresenta baixo grau de manutenção no sistema; Instalação no servidor simplificado.	Ausência plugin para dados espaciais; Ausência de gerenciamento versões dados.
Data Verse	Escopo: Repositórios dados; Uso predominante no meio universitário.	Plugin dados espaciais; Baixo grau de manutenção.	Processo de instalação requer cuidados; Limitação no tamanho dos dataset.

Fonte: Autoria própria (2020).

DSpace apresenta como um software híbrido uma característica que abrange organizações para publicação de dados e publicações de revistas ou artigos. DSpace tem maior uso nas organizações mundiais, destaque que software é usado pelos governos e universidades. Porém como DSpace adere os dois meios, apresenta limitações: ausência de plugins para visualização de dados espaciais e falta do recurso de versionamento dados sendo assim quando um conjunto de dados publicado não é possível realizar sua alteração.

DataVerse tem seu propósito para trabalhar com dados, mostrando recursos interessantes para configuração de ambiente desejada incluindo: hierarquias organizações, esquemas de metadados e versionamento de dados. Uso do software predominante no meio universitário para publicações e divulgações de dados podendo trabalhar com visualização de dados espaciais.

Para DSpace podemos configurar com os mesmos recursos do DataVerse, mas como objetivo do DSpace não é trabalhar especificamente com dados, então o software apresenta limitações no controle de versionamento dos dados. CKAN apresenta limitações comparado aos outros softwares, mas ele se apresenta uma alternativa quando usado como serviço para divulgação dos dados, mas com a submissão e preservação digital sendo mantida por outro software.

Para ajudar na escolha do software para o repositório FEBR teve a participação nos eventos: “Escola de Outono” promovido pelo Programa de Pós-Graduação em Ciência da Informação da Universidade Federal do Rio de Janeiro (PPGCI/IBICT-UFRJ). Este evento foi em formato de webinar que ocorreu em 19/06/2020 com o tema “Interoperabilidade e Tecnologias Aplicadas a Repositórios de Dados de Pesquisa”. A gravação do evento está disponível na plataforma (YouTube) no canal PPGCI/IBICT-UFRJ. A segunda participação foi uma reunião Dataverse Community Meeting 2020, realizado pela Universidade

Harvard nos dias 17, 18 e 19 de junho com vários webinar recursos e casos de uso.

As participações nestes webinars com os profissionais que trabalham com planejamento da preservação dos dados, contribuiu para um esclarecimento sobre funcionamento da arquitetura sobre cada software e o seu principal objetivo. Portanto podemos concluir que os softwares para o gerenciamento de metadados tem papel importante para preservação digital dos dados. Dessa forma o software DataVerse que apresentou características mais adequadas para o gerenciamentos dos metadados para FEBR. Conclusão semelhante foi alcançada pelo projeto de pesquisa Rede Nacional de Pesquisa Brasileira, desenvolvido pela Universidade Federal do Rio Grande do Sul (FURG), com participação da Rede Nacional de Ensino e Pesquisa (RNP) e Instituto Brasileiro de Informação em Ciências e Tecnologia (PAVÃO et al., 2018).

CONCLUSÃO

Dentre os três sistemas de gerenciamento e catalogação de metadados de conjuntos de dados de pesquisa mais populares (DSpace, CKAN e DataVerse), o mais recomendado para adoção pelo FEBR é o DataVerse. Apesar de os três sistemas serem bastante similares em sua estrutura e funcionamento, o DataVerse se destaca por apresentar funcionalidades específicas para o gerenciamento de dados de pesquisa. As duas principais são (a) o registro de versão dos conjuntos de dados e (b) o uso de esquema de metadados concebido para dados de pesquisa. A próxima etapa do projeto deve envolver a instalação do software para realizar testes de modo a avaliar aspectos relacionados a sua manutenção e usabilidade.

AGRADECIMENTOS

Este trabalho foi financiado pela da UTFPR/PROREC como financiadora do PIBEX, na forma de bolsa de projeto de extensão (Edital 01/2019 – PROREC/UTFPR). Os autores são gratos à Gabriel Panca Santos (UTFPR), Maiara Pusch (UNICAMP) e Taciara Horst-Heinen (UFSM) pelos comentários em uma versão preliminar do artigo. Os autores também são gratos à Augusto Herrmann (Ministério do Planejamento), Debora Pignatari Drucker (Embrapa Informática Agropecuária) pelas informações prestadas sobre os sistemas de de gerenciamento de metadados em repositórios.

REFERÊNCIAS

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data: The story so far. In: **Semantic services, interoperability and web applications: emerging concepts.** [s.l.] IGI Global, 2011. p. 205–227.

PAVÃO, C. M. G. et al. **Acesso aberto a dados de pesquisa no Brasil: soluções tecnológicas para compartilhamento de dados no Brasil**: relatório. 2018.

SILVA, J. DA. **Uma Abordagem para Especificação de Requisitos Dirigida por Modelos Integrada ao Controle de Qualidade de Aplicações Web.** PhD Thesis—[s.l.] Tese de Doutorado. UFRJ/COPPE/Programa de Engenharia de Sistemas e ..., 2011.

SOUZA, M. F. DE. C**omunicação da informação científica em novos espaços de memória.** Master's Thesis—[s.l.] Universidade Federal de Pernambuco, 2012.

RE3DATA. Home. [S.l.], 2020. Disponível em: https://www.re3data.org/. Acesso em: 20 jul. 2020.

WALPORT, M.; BREST, P. Sharing research data to improve public health. **The Lancet**, v. 377, n. 9765, p. 537–539, 2011.