Avaliação de softwares livres e de código aberto para gestão de metadados no Repositório Brasileiro Livre para Dados Abertos do Solo

Marcos Alexandre dos Anjos
Universidade Tecnológica Federal do Paraná - UTFPR Santa Helena/PR, Brasil
ORCID 0000-0002-6842-9853

Alessandro Samuel-Rosa
Universidade Tecnológica Federal do Paraná - UTFPR Santa Helena/PR, Brasil
ORCID 0000-0003-0877-1320

***Resumo* — Neste artigo apresentamos os resultados obtidos na avaliação dos softwares de gerenciamento de metadados no Repositório Brasileiro Livre para Dados Abertos do Solo (FEBR). Os métodos e procedimentos consistiram numa pesquisa bibliográfica sobre sistemas de catalogação e gerenciamento de dados e metadados da pesquisa, usando o Registry of Research Data Repositories (re3data.org). No entretanto, os sistemas de catalogação e gerenciamento de dados e metadados, segundo o re3data.org, são três as alternativas mais populares: CKAN, DataVerse e DSpace. Os primeiros resultados se destacam pelas funcionalidades de 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, principais limitações de cada software e instalação.**

***Abstract* — In this paper we present results obtained from 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). With regard to data and metadata cataloging and management systems, according to re3data.org, there are three most popular alternatives: CKAN, DataVerse and DSpace. 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.**

Metadados, CKAN, DataVerse, DSpace.

I. Introdução

O avanço das tecnologias digitais impulsionou o compartilhamento dos dados de pesquisas científicas, organizando o acesso às informações de maneira descentralizada [1]. Em meados dos anos 90, surgiram os primeiros registros de compartilhamento de dados científicos abertos. A partir disso, o professor Paul Ginsparg desenvolveu um servidor de dados para seu laboratório de pesquisa, possibilitando aos demais pesquisadores do laboratório o compartilhamento de seus dados [2]. Assim, outros pesquisadores passaram a discutir o tema, iniciando um movimento de acesso aberto aos dados.

Nos últimos anos, a relevância do compartilhamento de dados científicos levou pesquisadores a organizarem e padronizarem o compartilhamento de dados. Surgiram então softwares capazes de gerenciar esses dados [2]. Um exemplo de plataforma que disponibiliza informações de repositórios é a *Registry of Data Repositories* (re3data.org), que reúne muitas áreas do conhecimento e representa um registro global que fornece informações detalhadas a respeito de diversos softwares para gerenciamento de dados.

A ciência do solo, especificamente, apresenta um déficit na organização e padronização dos dados das pesquisas. Muitas vezes, os dados são compartilhados através de arquivos em formato PDF, tais arquivos dificultam a reusabilidade dos dados, aumentando significativamente o esforço de extração desses dados.

No Repositório Brasileiro Livre para Dados Abertos do Solo (FEBR) estão contidos dados, obtidos através de pesquisa, sobre os solos brasileiros. Seu objetivo é a preservação, padronização e versionamento dos dados depositados, sempre pensando em estratégias para facilitar o reuso dos dados pela comunidade.

Dessa maneira o FEBR tem como objetivo buscar por softwares de gerenciamento de dados que utilizam o padrão *Open Archives Initiative* (OAI). Este padrão implica diretamente na interoperabilidade entre os repositórios digitais, o que facilita a leitura dos metadados realizada pelos algoritmos dos buscadores, facilitando as buscas.

Tendo em vista que os softwares para gerenciamento dos metadados apresentam as características, diferenças e fragilidades destes softwares de forma a facilitar e justificar a escolha do software.

O objetivo deste artigo é apresentar os resultados parciais da avaliação de softwares de gerenciamento de metadados para o FEBR, listando suas respectivas potencialidades e lacunas identificadas.

II. Metodologia

Os softwares avaliados foram definidos a partir de consulta ao re3data.org. A figura 1 apresenta os três softwares livres e de código aberto mais utilizados para a gestão de metadados em repositórios até novembro de 2020. São eles: CKAN, DSpace e DataVerse. Os demais softwares listados na figura 1 não foram analisados pela limitação de recursos assim como os softwares apresentam outros

objetivos. No caso MySQL é um software para gerenciamento de banco de dados, mas seu uso pode ser justificado pela facilidade de desenvolver soluções próprias para gerenciamento dos dados.

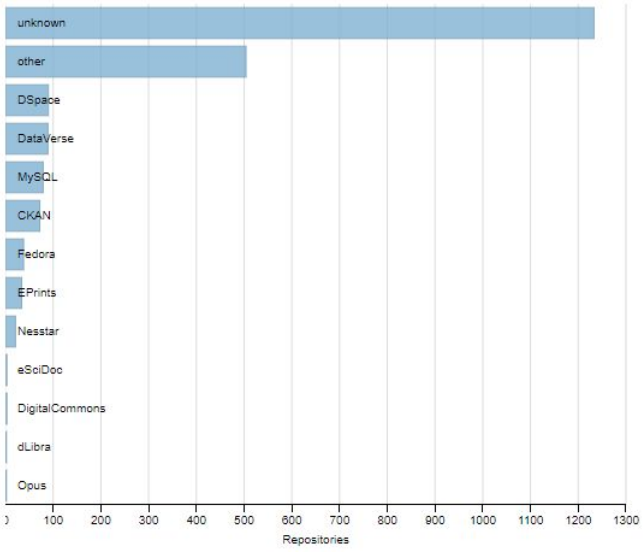

Figura.1 Softwares de gerenciamento de metadados em repositórios listados no re3data.org em novembro de 2020.

A. Revisão de Literatura e Consulta a Especialistas

Neste assunto relativamente novo são encontrados poucos trabalhos na literatura relacionados ao tema gerenciamento de metadados. Foram avaliados os trabalhos encontrados na bibliografia e documentação oficial de cada software [3] [4] [5]. Assim a tomada decisão foi complementada com a participação em eventos como: "Escola de Outono" promovido pelo Programa de Pós-Graduação em Ciência da Informação da Universidade Federal do Rio de Janeiro (PPGCI/IBICT-UFRJ), cujo tema foi "Interoperabilidade e Tecnologias Aplicadas a Repositórios de Dados de Pesquisa". A segunda participação consistiu em uma reunião Dataverse Community Meeting 2020, promovida pela Universidade Harvard. Onde a participação neste eventos permitiu uma consulta com os especialistas da área para maiores esclarecimentos sobre o tema.

Isso permitiu selecionarmos algumas características a serem avaliadas nos softwares analisados para uma melhor tomada de decisão. São elas:

a) Requisitos básicos de software e hardware;
b) Finalidade do software;
c) Limitações de cada software;
d) Instalação e lacunas detectadas.

B. Testes Operacionais

Para a realização da instalação a Universidade Tecnológica Federal do Paraná (UTFPR), câmpus Santa Helena disponibilizou um servidor com as seguintes especificações: máquina virtual (VM) com sistema operacional Linux Ubuntu Server 20.04 (LTS), disco rígido (HD) 50 gigabytes, processador com 4 núcleos e 8 gigabytes de memória RAM. Esta configuração básica possibilitou a realização deploy do software e demais testes necessários. Todos os processos de instalação estão documentados em em arquivo de texto com detalhes necessários como versões e comandos para instalação das dependências. Sendo considerado uma boa prática a documentação da instalação dos pacotes e dependências que são instaladas no servidor. Desta forma, é estabelecido um maior controle sobre as atividades e eventuais inconsistências no sistema. Com o auxílio da documentação também são evitados eventuais problemas relacionados à atualizações no sistema, evitando conflitos entre versões das dependências.

III. Resultados Esperados e Parciais

A pesquisa exploratória foi concluída e culminou em uma sugestão para o FEBR no que diz respeito ao software mais adequado para o gerenciamento de metadados. Os parâmetros considerados para a seleção estão diretamente relacionados às características levantadas: configuração básica de software e hardware para o servidor, finalidade do software e limitações do software de gerenciamento de metadados.

Observando os dados na tabela 1, foi realizado o levantamento das configurações necessárias de software e hardware. O DataVerse se destaca pela maior demanda por memória RAM. Isso ocorre porque o modelo da arquitetura do software permite criação e o gerenciamento de vários DataVerses a partir do software instalado. No caso do software DSpace requer espaço maior para armazenamento em disco rígido pelo fato que precisamos configurar além do DSpace uma outra dependência para realizar o gerenciamento dos dados isso justifica o tamanho do disco rígido.

TABELA I
REQUISITOS MÍNIMOS DE SOFTWARE E HARDWARE

CKAN	**HD:** 25 GB **Memória ram:** 4 GB **Processador:** 4 núcleos de 2 GHz
DSpace	**HD:** 48 GB **Memória ram:** 4 GB **Processador:** 2 núcleos de 2,8 GHz
DataVerse	**HD:** 25 GB **Memória ram:** 8 GB **Processador:** 2 núcleos de 2 GHz

Reconhecendo os requisitos mínimos de instalação dos softwares analisados, são apresentadas na tabela 2 as finalidades de cada software bem como uma visão geral a respeito das organizações nas quais os mesmos são geralmente empregados.

TABELA II
FINALIDADE E USO DE CADA SOFTWARE

CKAN	**Finalidade:** Repositório de dados **Uso:** Predominante no governo para visualização de dados.
DSpace	**Finalidade:** Repositório de dados e publicações **Uso:** Na mesma frequência no governo e nas universidades
DataVerse	**Finalidade:** Repositório de dados **Uso**: Predominante no meio acadêmico

A tabela 3 apresenta os principais pontos considerados em cada software. Uma vez conhecidos os requisitos mínimos, sua finalidade e uso predominante dos softwares em questão, é possível levantar os principais pontos destacados.

TABELA III
PONTOS FORTES E FRACOS DE CADA SOFTWARE

	Pontos Fracos	**Pontos Fortes**
CKAN	Não apresenta opção para versionamento de dados. Falta uma estrutura para organização dos datasets.	Conta com plugins para visualização de dados espaciais. Apresenta compatibilidade com distribuição Linux/Windows.
DSpace	Não apresenta capacidade para versionamento de dados. DSpace não foi desenvolvido para dados de pesquisa.	Compatibilidade com diferentes distribuições Linux. Personalização em esquemas de metadados. Apresenta um software sólido com baixa intensidade de manutenção.
Data Verse	Recomendação Linux CentOS para instalação. Limitação no tamanho do upload de dataset.	Conta com plugins para dados espaciais; Compatibilidade e facilidade no versionamento e gerenciamentos dos DataVerses.

Dessa maneira, foram obtidas três análises dos softwares apresentados com um ponto comum entre eles que apresentam suporte para instalação em ambiente Linux. O primeiro software, CKAN, caracterizou-se como um software simples e mais direcionado para abertura e visualização de dados. Este software é comumente encontrado em aplicações do governo, que se utilizam deste software para a abertura de dados. Apesar disso, ele não apresentou suporte para o versionamento dos dados. Entretanto, ele ainda representa uma opção, se utilizado em conjunto com outro software capaz de realizar tal função.

O segundo software analisado, DSpace, contém dados e publicações de forma integrada. Todavia, essa integração deixou o software com limitações, incluindo a ausência do versionamento dos dados. Entretanto existe uma solução que recentemente que foi o desenvolvimento de um repositório para o versionamento chamado de Dryad. Neste repositório Dryad fica armazenado todos arquivos são realizados uploads assim como ele aceita um identificador persistente único. Sua desvantagem é que quando realizamos uma alteração este arquivo recebe um novo identificador. Por último, podemos personalizar nosso esquema de metadados de acordo com as necessidades da organização.

O terceiro e último software analisado, DataVerse, se apresentou como um software robusto não somente capaz de trabalhar com os dados como também de realizar seu versionamento. Dataverse assim como DSpace foram desenvolvidos para ser uma solução voltada para pesquisa científicas, gestão dos metadados, atribuição de identificadores únicos persistentes e possibilidade de interoperabilidade [6]. O Dataverse apresenta-se como uma solução que apresenta fácil acesso para reutilizar os datasets, assim como disponibiliza uma *Application Programming Interface* (API), para depositar e consultar os conjuntos de dados depositados.

De modo geral, o DataVerse apresentou pontos relevantes para o gerenciamento dos dados do FEBR.

A. DataVerse

Dentre as versões disponíveis do software Dataverse, a versão 4.19 se faz bastante presente em muitas organizações, enquanto a versão 5 ainda se mostra pouco aderente neste sentido. Desta forma, a versão 5, considerada a mais recente, foi determinada para a realização dos testes no FEBR. Em novembro recebemos a nova atualização do software, para versão 5.1.1. Esta versão apresenta diversas melhorias e correções de erros. O protocolo de instalação foi realizado conforme as normas e recomendações da documentação oficial do software.

O processo de instalação durou cerca de 50 horas, incluindo a configuração da máquina virtual e a instalação do DataVerse e suas dependências. O processo de instalação está completamente documentado para a distribuição Linux Ubuntu Server 20.04. Desta forma, futuros trabalhos podem se utilizar deste recurso, aplicando menores esforços durante os próximos processos.

Até o presente momento, esta pesquisa encontra-se na etapa de configurações internas do DataVerse, realizando os ajustes finais. Em breve serão iniciados os testes na aplicação, que já encontra-se disponível para acesso. Entretanto, pelo fato de o estudo encontrar-se em processo

de implantação, neste momento os dados são acessíveis apenas para os desenvolvedores do FEBR. Os testes de usabilidade serão aplicados para usuários cadastrados na plataforma, permitindo que cada usuário carregue um ou mais conjuntos de dados.

B. Comparação com Outros Estudos

Outros dois grupos de pesquisa estão estudando os sistema de gerenciamento de metadados apresentando resultado semelhante foi alcançado pelo grupo de pesquisa Rede Nacional de Pesquisa (RNP), desenvolvida pela Universidade Federal do Rio Grande do Sul (UFRGS) com objetivo que identifica soluções tecnológicas para a construção de repositório para Acesso Aberto a Dados de Pesquisa (AADP), concluíram que melhor software para gerenciamento dos metadados é Dataverse [7]. Em relatório publicado pelo Ministério da Transparência, Fiscalização e Controladoria-Geral da União o documento apresenta Implantação da infraestrutura federada piloto de repositórios de dados da pesquisa, apresenta as organizações envolvidas que estão sendo implantado o repositório Dataverse RNP, Lattes Data (CNPq), IBICT, Embrapa, ainda se encontra processo de implantação e testes [8].

IV. Conclusão

Dentre os três sistemas avaliados, o mais recomendado para adoção no FEBR é o DataVerse. De modo geral, todos mostraram-se bastante similares em sua estrutura e funcionamento. Contudo, o DataVerse se destaca por apresentar funcionalidades específicas para o gerenciamento de dados de pesquisa.

Como trabalho futuro serão realizadas análises envolvendo desempenho do sistema operacional, medição do tempo dos processos, do desempenho do banco de dados com avaliação do desempenho do hardware versus softwares. Assim como envolverá parte de testes e gerenciamento dos arquivos de modo a deixar a aplicação escalável.

Agradecimentos

Este trabalho foi financiado pela da UTFPR/PROREC na forma de bolsa iniciação à extensão (Edital 01/2019 – PROREC/UTFPR). Os autores são gratos à Gabriel Panca Santos (UTFPR), Bruna Finardi (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 gerenciamento de metadados em repositórios.

Referências

[1] M. F. de Souza, “Comunicação da informação científica em novos espaços de memória”, Master’s Thesis, Universidade Federal de Pernambuco, Recife, PE, 2012.

[2] M. Walport e P. Brest, “Sharing research data to improve public health”, *The Lancet*, vol. 377, nº 9765, p. 537–539, 2011.

[3] “Full table of contents — CKAN 2.10.0a documentation”. https://docs.ckan.org/en/latest/contents.html (acessado ago. 18, 2020).

[4] “All Documentation - DSpace Documentation - LYRASIS Wiki”. https://wiki.lyrasis.org/display/DSDOC/All+Documentation (acessado ago. 18, 2020).

[5] “Dataverse Documentation v. 5.1.1 — Dataverse.org”. https://guides.dataverse.org/en/latest/ (acessado nov. 20, 2020).

[6] C. G. Pavão, E. N. Borges, L. Ciuffo, e L. A. B. Azambuja, “Acesso Aberto a Dados de Pesquisa no Brasil: Práticas e Soluções Tecnológicas.”, Simpósio Internacional Network Science, Rio de Janeiro, RJ, 2018. Acessado: set. 20, 2020. [Online]. Disponível em: http://networkscience.com.br/wp-content/uploads/2018/11/IISINS_EScience_DadosAbertos_Artigo_AcessoAberto.pdf.

[7] C. M. G. Pavão, R. F. Gabriel Junior, S. A. de S. Vanz, E. N. Borges, e L. A. B. Azambuja, “Acesso aberto a dados de pesquisa no Brasil: soluções tecnológicas para compartilhamento de dados no Brasil: relatório”, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, 2018. [Online]. Disponível em: https://www.lume.ufrgs.br/handle/10183/185126.

[8] P. R. B. BERTIN, “Relatório de Status de Execução: Estabelecer mecanismos de governança de dados científicos para o avanço da ciência aberta no Brasil”, Ministério da Transparência, Fiscalização e Controladoria-Geral da União Parceria para Governo Aberto, Brasília, DF, mar. 2020. Acessado: ago. 10, 2020. [Online]. Disponível em: https://www.gov.br/cgu/pt-br/governo-aberto/a-ogp/planos-de-acao/4o-plano-de-acao-brasileiro/compromisso-3-docs/rse_3_-10mar2020.pdf.