Comparando softwares gratuitos para criação de repositórios de dados abertos

Leonard Richard Rodrigues Rufino Campêlo
Tecnólogo em Análise e Desenvolvimento de Sistemas pelo Instituto Superior Fátima (ISF) - Brasília, DF - Brasil. Bolsista do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) - Brasília, DF – Brasil.
http://lattes.cnpq.br/1521448863751526
E-mail: leonardcampelo@ibict.br

Vanderlino Coelho Barreto Neto
Mestre em Ciências Mecânicas pela Universidade de Brasília(UnB) - Brasília, DF - Brasil. Bolsista do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) - Brasília, DF – Brasil.
http://lattes.cnpq.br/4681885854457562
E-mail: vanderlinoneto@ibict.br

Data de submissão: 27/09/2019. Data de aprovação no ConfOA: 12/06/2019. Data de publicação:

RESUMO

O estudo apresenta uma comparação entre os softwares livres para construção de repositórios Dataverse, Invenio, DSpace e CKAN. A partir de um conjunto de critérios, avalia-se em que nível essas ferramentas possuem as funcionalidades necessárias para a construção de um repositório de dados científicos.

Palavras-chave*: Dados de pesquisa. Software livre. Gestão de dados de pesquisa.*

Comparing free software for creating open data repositories

ABSTRACT

The study presents a comparison between free software for building Dataverse, Invenio, DSpace and CKAN repositories. From a set of criteria, it is evaluated at what level these tools have the functionalities necessary to build a repository of scientific data.

Keywords: *Research data. Free software. Research data management.*

Comparación de software libre para crear repositorios de datos abiertos

RESUMEN

El estudio presenta una comparación entre el software libre para construir repositorios Dataverse, Invenio, DSpace y CKAN. A partir de un conjunto de criterios, se evalúa a qué nivel estas herramientas tienen las funcionalidades necesarias para construir un repositorio de datos científicos.

Palabras clave: *Datos de investigación. Software libre. Gestión de datos de investigación.*

INTRODUÇÃO

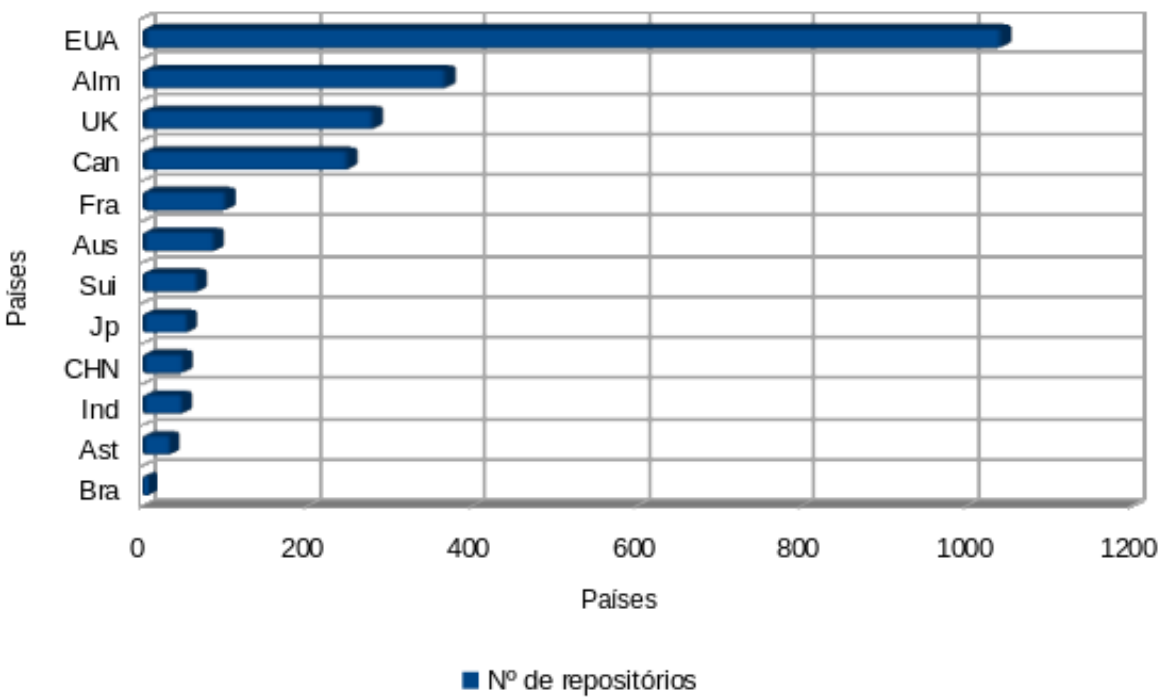
A produção de dados científicos está aumentando com o passar tempo. Esse crescimento trouxe consigo a necessidade de armazenar os dados coletados durante pesquisas científicas, para que seja possível disponibilizar o acesso a esses dados por outros pesquisadores. Com o desenvolvimento da tecnologia da informação, o armazenamento e a organização dos dados se tornou uma peça fundamental para a disseminação da informação contida nesses objetos.

Nos anos 1990 teve início a era dos repositórios de documentos digitais, que tinham como objetivo difundir a produção científica de certo nicho ou instituição. Mais recentemente, o desenvolvimento da tecnologia sobre objetos digitais foi agregada à proposta de Ciência Aberta, com a disponibilização não apenas de documentos científicos, mas também dos dados coletados durante as pesquisas.

Entre as iniciativas pioneiras para construção de repositórios de dados de pesquisa existentes, destaca-se o software de repositório Dataverse, desenvolvido em um projeto apoiado por uma equipe de pesquisadores da Universidade de Harvard (DATA, 2019). O Dataverse foi criado com o objetivo de resolver os problemas de compartilhamento de dados com a criação de tecnologias e incentivar os pesquisadores e editores a compartilhar seus dados (DATA, 2019).

O mais conhecido diretório de registro de instâncias repositórios de dados de pesquisa é o Re3data (RE3DATA, 2019). Atualmente, existem mais de 2.300 repositórios registrados nesse diretório, o qual permite a aplicação de filtros de busca e geração de alguns gráficos de estatísticas. Uma das análises possível é a quantidade de repositórios registrados em cada país.

Figura 1 – Gráfico dos números de repositórios registrados por países (2019)

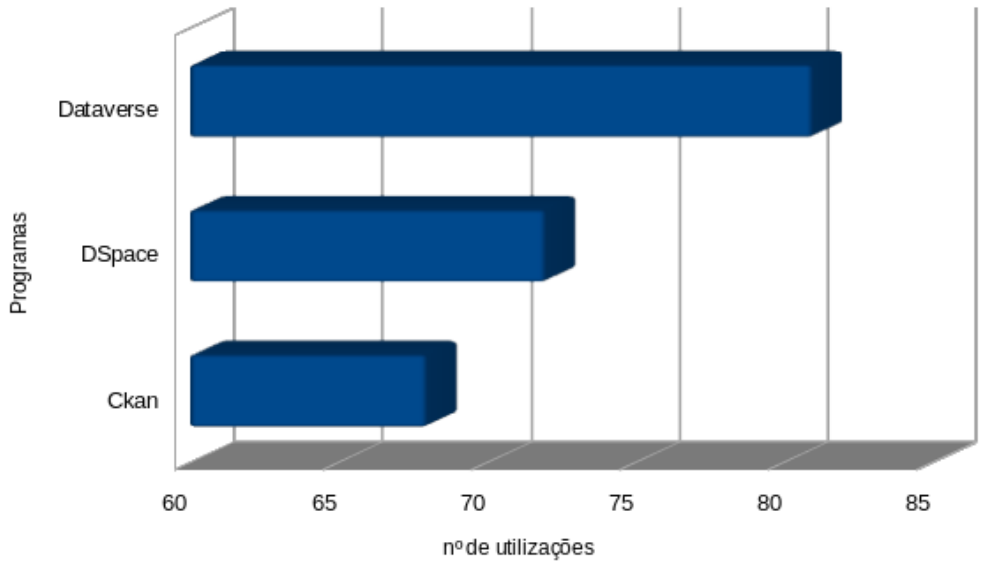


Na figura 1, verifica-se que os Estados Unidos detêm mais de 1.040 repositórios, seguidos da Alemanha, com aproximadamente 370, e do Reino Unido, com mais de 280 repositórios cada. O Brasil tem aproximadamente 8 repositórios registrados no Re3data (RE3DATA, 2019).

O presente estudo apresenta os resultados da avaliação dos softwares livres Dataverse, Invenio, DSpace e CKAN, com a exibição do nível de adequação das funcionalidades dentro de cada um dos critérios adotados. Estes softwares foram escolhidos tomando como referência a sua relevância na comunidade, tendo em vista o número de repositórios criados com o respectivo software.

Como se verifica na figura 2, o software Dataverse segue como um dos mais utilizados, seguido pelo DSpace e CKAN, nessa ordem. O software Invenio foi selecionado dada a sua relevância em algumas comunidades de usuários. A escolha dos softwares também tomou como base os pré-requisitos necessários para implementação de um repositório, os protocolos existentes de interoperabilidade, os padrões de descrição de metadados e os formatos de arquivo suportados.

Figura 2 – Gráfico do registrado de utilização dos softwares (2019)



METODOLOGIA

Os pontos avaliados foram escolhidos com base em Martins, Silva (2017), Rodrigues (2004) e Pavão *et al.* (2018), sendo eles uma variedade de critérios com foco em aspectos técnicos. Esses pontos são infraestrutura, organização e controle de conteúdo, descoberta de conteúdo, ferramentas de relatórios, recursos sociais e notificação, interoperabilidade, autenticação, acessibilidade, preservação e curadoria digital, escalabilidade, linguagem de programação em que o software foi desenvolvido, suporte a banco de dados, suporte a serviços e manutenção, formato de arquivo de suporte, serviços web.

ANÁLISE E DISCUSSÃO DOS RESULTADOS

O Dataverse foi desenvolvido em uma aplicação Java EE que é compilada em um arquivo War e implementado em um servidor de aplicações (Glassfish). Esse software para repositórios de dados utiliza o banco de dados relacional (PostgreSQL) e o motor de busca (Apache-Solr) (DATA, 2019). Tem-se uma ferramenta capaz de representar estruturas organizacionais hierárquicas, à medida que dataverses podem conter outros dataverses (AMORIM *et al.,* 2017) e (PRINCIPE *et al.,* 2018) .

Cada dataverse contém datasets, que são as entidades que representam conjuntos de dados na forma de estudo.

As estruturas de um dataset são adequadas para representar dados de pesquisa, sendo compatíveis com padrão de metadados DDI e DDI 2.5. Esses esquemas de metadados visam descrever dados no contexto de pesquisa científica, e envolvem tanto metadados descritivos, quando metadados estruturais, que caracterizam os arquivos que compõem o pacote (PAC2). No caso de conjuntos de dados tabulares, ocorre também a descrição das estruturas das variáveis (PAC2) (ROCHA *et al.,* 2018).

Já no software Invenio, tem-se uma ferramenta construída com base no framework Flask na linguagem de programação Python. Nota-se que essa é uma linguagem que vem ganhando muito espaço no desenvolvimento em softwares para análise de dados (INVENIO, 2019).

O software Invenio é construído com muitos módulos individuais menores, que é a maneira como o Flask permite criar aplicativos modulares, os chamados blueprints. Em geral, um aplicativo Flask é executado por meio de três interfaces diferentes: a WSGI, a CLI e o Celery. Em relação aos motores de busca, o Invenio é o único que faz uso do Elasticsearch, enquanto os demais utilizam o Apache-Solr.

No DSpace, tem-se uma ferramenta que opera em vários níveis: com um servlet Java (que é um contêiner de servlet sobre o Apache-Tomcat), e com tarefas agendadas e operações sob demanda. Muitas dessas operações são executadas na interface de linha de comando (CLI), também conhecida como prompt do Unix. A interface web utilizada pelo DSpace pode ser a JSPUI ou a XMLUI, sendo a última baseada em programação sobre XML (DSPACE, 2019).

Sucintamente, a constituição do DSpace é orientada à disseminação de publicações científicas, e não há relação explícita com outras entidades, como por exemplo pessoas e projetos (ROCHA *et al.,* 2018). Esse software não foi desenvolvido com a finalidade de implementação de repositórios de dados de pesquisa, porém, com adaptações sobre sua lógica de comunidades e coleções, é possível representar as instituições produtoras ou custodiadoras de conjuntos de dados.

No Dspace também não existem interfaces de visualização de documentos do tipo planilha, imagem ou apresentação (AMORIM *et al.*, 2017). Sua tecnologia baseada em Java exige profissionais especializados para realização de customizações e modificações de seu código fonte, quando utilizado para o armazenamento e organização de dados científicos.

O CKAN é um software desenvolvido para a criação de portais de dados abertos governamentais. É uma ferramenta que auxilia o gerenciamento e a disseminação de coleções de dados (CKAN, 2019). No Brasil, é um software amplamente utilizado por instituições governamentais na disponibilização de recursos digitais como planilhas, arquivos CSV, XML, PDF e RDF, JSON, entre outros formatos.

No CKAN, pode-se armazenar o recurso internamente ou simplesmente endereçá-lo com um link ao hospedeiro remoto do recurso digital. Um conjunto de dados pode conter um número arbitrário de objetos digitais. Infelizmente o CKAN não é nativamente compatível com a comunicação via OAI-PMH, que é um protocolo amplamente utilizado para promover a interoperabilidade entre repositórios de arquivos abertos (AMORIM *et al.*, 2017).

A partir das informações coletadas durante a instalação e avaliação dos softwares, obteve-se a tabela 1, que contém uma descrição de acordo com os critérios de avaliação adotados.

Tabela 1 − Informações dos softwares de repositório

Softwares	Dataverse	Invenio	DSpace	CKAN
Desenvolvido na linguagem	JAVA	Flask (Python)	JAVA	Python
Ferramenta de busca	Apache-Solr	Elasticsearch	Apache-Solr	Apache-Solr
Interface gráfica	PrimeFaces e Bootstrap	WSGI, e Aipo	JSPUI, XMLUI	ckan plugins
Banco de Dados	PostgreSQL	PostgreSQL	PostgreSQL	PostgreSQL
Comunicação de interfaces	OAI -PMH	API REST	OAI -PMH	Não habilitado
Processo de instalação	Fácil, documentação bem detalhada	Complicada, instalação incompleta	Fácil, documentação bem detalhada	Fácil, documentação bem detalhada
Formatos atendidos	SPSS, STATA, R, XLSX, CSV	IIIF Image API support	PDF, XML, txt, asc, MARC, JPEG, JPG, GIF, PNG, TIFF, AIFF, RTF, Postscript	XML and JSON CSV, Excel, XML, PDF, RDF
Formato do metadata	DDI Lite, DDI 2.5 Codebook, DataCite 3.1, and Dublin Core	MARC	Dublin Core	DCAT

CONCLUSÕES

Foi realizada uma avaliação, evidenciando-se os critérios relevantes, com base no grau de facilidade de uso de tecnologia da informação, à implementação de um repositório de dados científicos. Esse tipo de análise deve permitir nortear a escolha do software na implementação de um repositório de dados científicos.

A documentação do Dataverse é detalhada e redigida em termos compreensíveis, o que facilitou a implementação de um repositório baseado nessa tecnologia. Suas ferramentas atendem às necessidades de um repositório de dados científicos *out-of-the-box*.

A documentação do Invenio descreve de maneira completa o software e suas funções, porém se mostrou incompleta no procedimento de instalação. O software tem uma arquitetura que envolve ferramentas robustas como o ElasticSearch.

Já o DSpace é um software que possui ampla comunidade de usuários. Com isso, há manuais de instalação completos, o que facilita sua implementação e manutenção. Em contrapartida, o DSpace não é um software desenvolvido com a finalidade de implementação de um repositório de dados, e necessita, portanto, de customizações para que desempenhe a contento essa função.

O CKAN é, fundamentalmente, um software destinado à implementação de repositórios de documentos governamentais, e eficiente para esse tipo trabalho. No entanto, algumas funcionalidades conhecidas no meio de arquivos abertos não estão presentes por padrão na ferramenta. Por exemplo, o CKAN não carrega consigo originalmente uma interface OAI-PMH, o que dificulta sua aplicação a redes interoperáveis de repositórios de dados de pesquisa.

REFERÊNCIAS

AMORIM, R. C. *et al*. A comparison of research data management platforms: architecture, flexible metadata and interoperability. *Universal Access in the Information Society*, [*S.l.*], v. 16, n. 4, p. 851-862, 2017.

CKAN. *CKAN - The Open Source Data Portal Software*. [*S.l.*], 2019. Disponível em: https://ckan.org. Acesso em: 6 jan. 2019.

DATAVERSE. *The Dataverse Project*. [*S.l.*], 2019. Disponível em: https://dataverse.org. Acesso em: 3 jan. 2019.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA. *Sistema para Construção de Repositórios Institucionais Digitais (DSpace)*. Brasília, 2019. Disponível em: http://www.ibict.br/dspace. Acesso em: 10 jan. 2019.

INVENIO. [*S.l.*], 2019. Disponível em: https://invenio-software.org/. Acesso em: 10 fev. 2019.

MARTINS, D. L.; SILVA, M.F. Critérios de avaliação para sistemas de bibliotecas digitais: uma proposta de novas dimensões analíticas. *InCID*: Revista de Ciência da Informação e Documentação, Ribeirão Preto, v. 8, n. 1, p. 100-121, 20 abr. 2017.

PAVÃO, C. M. G. *et al*. *Acesso aberto a dados de pesquisa no Brasil*: repositórios brasileiros de dados de pesquisa: relatório 2018. Porto Alegre: UFRGS, 2018.

PRÍNCIPE, P. *et al*. Estratégia Institucional para a gestão de dados de investigação na UMINHO: o papel dos SDUM. *In*: CONGRESSO NACIONAL DE BIBLIOTECÁRIOS, ARQUIVISTAS E DOCUMENTALISTAS, 13., 2018, Portugal. *Actas* [...]. Portugal: [*s.n.*], 2018.

RE3DATA. *Home*. [*S.l.*], 2019. Disponivel em: https://www.re3data.org/. Acesso em: 26 set. 2019.

ROCHA, R. P. *et al*. *Acesso aberto a dados de pesquisa no Brasil*: Soluções Tecnológicas para Compartilhamento de Dados no Brasil repositórios brasileiros de dados de pesquisa. Porto Alegre: UFRGS, 2018.

RODRIGUES, E. *et al*. RepositóriUM: criação e desenvolvimento do Repositório Institucional da Universidade do Minho. *In*: CONGRESSO NACIONAL DE BIBLIOTECÁRIOS, ARQUIVISTAS E DOCUMENTALISTAS, 8., 2004, Estoril, Portugal. *Actas* [...]. Estoril, Portugal: [*s.n*], 2004.

SEMELER, A. R. *et al*. *Ciência da informação em contextos de e-science*: bibliotecários de dados em tempos de Data Science. Florianópolis: UFSC, 2017.169 p.