Projetando uma Plataforma para Compartilhamento de Dados Científicos de Observação da Terra

Gabriel Sansigolo[1], **Gilberto R. de Queiroz**[1], **Karine R. Ferreira**[1]

[1]Instituto Nacional de Pesquisas Espaciais
Av. dos Astronautas, 1758
CEP 12227-010 - São José dos Campos - SP - Brazil

{gabriel.sansigolo, gilberto.queiroz, karine.ferreira}@inpe.br

Abstract. *The growing demand on scientific information sharing has motivated scientists and institutions to look for new computational tools for research data management and sharing. Today there are different platforms for publishing scientific data, such as Pangea or Zenodo. However, these platforms, due to their restricted characteristics, do not integrate data with tools used by Earth observation researchers. This paper presents ongoing work on defining a platform for Earth observation research data sharing, that integrates tools for storage, cataloging, management, processing and dissemination. Thus contemplating all the research activities.*

Resumo. *A crescente demanda por compartilhamento de informações científicas motivou cientistas e instituições a procurar novas ferramentas computacionais para gerenciamento e compartilhamento de dados de pesquisa. Hoje existem diferentes plataformas para publicação de dados científicos, como o Pangaea ou o Zenodo. Porém essas plataformas, devido a suas características fechadas, não possuem integração com ferramentas usadas por pesquisadores de observação da Terra. Este artigo apresenta um trabalho em andamento de projetar uma plataforma para compartilhamento de dados científicos de observação da Terra, integrando ferramentas de armazenamento, catalogação, gerenciamento, processamento e disseminação. Assim contemplando todas as atividades de uma pesquisa.*

1. Introdução

Ciência Aberta é o conjunto de práticas, ferramentas e políticas criadas para permitir a colaboração e compartilhamento de pesquisas. Isso inclui uma variedade de práticas como: Acesso Aberto, Dados de Pesquisa Abertos, *Softwares* de Código Aberto, entre outras [Woelfle et al. 2011, Bezjak et al. 2018]. Na Ciência Aberta dados, anotações e outros processos de uma pesquisa estão sobre termos que permitem o reúso, redistribuição e reprodução [Saez and Fuentes 2018].

Com o crescimento de popularidade de Dados Abertos, diferentes infraestruturas de dados e políticas nos âmbitos nacional, federal e institucional foram criadas. Em 2008, a Infraestrutura Nacional de Dados Espaciais (INDE[1]) foi instituída. A INDE é um conjunto de tecnologias, políticas e padrões criados para facilitar instituições do governo

[1]www.inde.gov.br

na geração e disseminação de seus dados geoespaciais. Em 2016, através da política do governo brasileiro para Dados Abertos[2], foi instituída a Infraestrutura Nacional de Dados Abertos (INDA). Composta por padrões, tecnologias e procedimentos, essa política busca tornar disponíveis para a sociedade dados governamentais.

Dados de Pesquisa Abertos, também chamados de dados científicos, são todos os dados que fazem parte do processo de uma pesquisa. Para a promoção de dados científico, a revista *Nature* lançou o *Scientific Data*[3], um periódico para descrições de conjuntos de dados, materiais e pesquisas com relevância científica. O periódico promove o compartilhamento e reutilização de dados científicos, princípio fundamental da Ciência Aberta. Essa demanda pode ser observada também na Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), que promovendo práticas de Ciência Aberta, criou um plano de gestão de dados[4], componente hoje obrigatório na fase de submissão de um projeto.

Nesse cenário, pesquisadores precisam de uma plataforma para compartilhar dados científicos de observação da Terra. Hoje existem diferentes plataformas para publicação de dados científicos, como o *Pangaea*[5], uma plataforma para publicação de conjuntos de dados geocientíficos [Diepenbroek et al. 2002], ou o Zenodo[6], um repositório aberto para resultados da pesquisa de uso geral. Porém, essas plataformas foram criadas para resolver problemas como armazenamento, compartilhamento e preservação. E devido a suas características fechadas, não possuem integração com ferramentas usadas por pesquisadores de observação da Terra durante as atividades de uma pesquisa. Nesse contexto existe uma demanda de uma plataforma que forneça, de maneira integrada, diferentes tecnologias para produção, processamento, gerenciamento e disseminação de dados de observação da Terra. Assim contemplando todas as atividades de uma pesquisa.

A Organização Internacional para Padronização (ISO) e o *Open Geospatial Consortium* (OGC), promovendo a interoperabilidade entre sistemas, propuseram padrões para representação, intercâmbio e disseminação de dados espaciais [OGC 2017]. Alguns desses padrões são os serviços *Web Map Server* (WMS), o *Web Feature Service* (WFS), o *Web Coverage Service* (WCS) e o *Catalogue Service Web* (CSW). As especificações da INDE são baseadas nos padrões OGC.

Esse artigo apresenta um trabalho em andamento em projetar uma plataforma para compartilhamento de dados científicos de observação da Terra. O objetivo é projetar uma plataforma que integre ferramentas para armazenamento, catalogação, gerenciamento, processamento e disseminação de dados de observação da Terra. Essa plataforma visa facilitar a integração com infraestruturas como INDA e INDE, e Sistemas de Informações Geográficas (SIG) através de facilidades para exportação usando os padrões OGC.

2. Frameworks para criação de bibliotecas digitais

Bibliotecas digitais são ferramentas projetadas para apoiar a disseminação de produtos de conhecimento [Amorim et al. 2017]. Para isso busca-se preservar, além de dados, artigos científicos, repositórios, materiais, entre outros produtos [Bezjak et al. 2018]. No con-

[2] www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm
[3] www.nature.com/sdata
[4] www.fapesp.br/gestaodedados
[5] www.pangaea.de
[6] zenodo.org

texto de criação de bibliotecas digitais, três *frameworks* se destacam: Invenio, Dataverse e CKAN.

Invenio é um *framework* aberto para construção de bibliotecas digitais de grande escala. A galeria de instâncias do *framework* é principalmente composta por plataformas relacionadas ao CERN (*Conseil Européen pour la Recherche Nucléaire*). Criado com o *framework* Invenio, o Zenodo[7] é um repositório aberto para resultados de pesquisa de uso geral . Ele foi especificamente projetado para ajudar pesquisadores e instituições menores a compartilhar os resultados de suas pesquisas [Sicilia et al. 2017].

Dataverse é um *framework* de Código Aberto para criação de plataformas *web* para compartilhar, preservar, citar, explorar e analisar dados de pesquisa [King 2007]. Instalado em dezenas de instituições ao redor do mundo, o sistema gera uma citação formal, para cada depósito. Proposto em 2007, o Dataverse foi responsável pela padronização de infraestruturas de compartilhamento de dados.

Comprehensive Knowledge Archive Network (CKAN) é um *framework* de Código Aberto para criação de *hub* de dados. Desenvolvido e promovido pela *Open Knowledge Foundation* (OKF) ele visa editores de dados, governos, empresas e organizações que querem tornar seus dados abertos [Wainwright 2012]. Um exemplo de instância do *framework* CKAN é o Portal Brasileiro de Dados Abertos[8], recomendado pela INDA para disseminação de dados públicos do Brasil.

A Tabela 1 sintetiza a análise de *frameworks* para criação de bibliotecas digitais, foram usadas funcionalidades recomendadas pela literatura de análise de *framework* [Amorim et al. 2017], e funcionalidades interessantes a observação da Terra. Essas funcionalidades foram selecionadas pois vão de encontro com práticas de Ciência Aberta e com práticas para boa navegabilidade.

Tabela 1. Comparação de funcionalidades dos *frameworks*

Funcionalidade	CKAN	Invenio	Dataverse
Código aberto	X	X	X
Versionamento de conteúdo	X		X
Pré-reserva de DOI		X	X
Esquema de dados flexível	Flexível	Flexível	Fixo
Visualização de conteúdo	X	X	X
Suporte a dados espaciais	X		X
Busca espacial	X		

Na Tabela 1 as funcionalidades apontadas foram: (a) Código aberto: quando o código fonte é disponibilizado pelo autor através de mecanismos que permitem estudo, edição e distribuição; (b) Versionamento de conteúdo: permitir o acompanhamento de todas as alterações feitas em um conteúdo; (c) Pré-reserva de DOI: gerar de forma automatizada um Identificador de Objeto Digital (DOI) para cada depósito; (d) Esquema de dados flexível: permitir adição de diferentes tipos de dados, sem a necessidade de redefinir a estrutura de dados principal; (f) Visualização de conteúdo: permitir que usuários do por-

[7] `zenodo.org`
[8] `dados.gov.br`

tal vejam os dados sem a necessidade de baixá-los; (g) Suporte a dados espaciais: possuir características geoespaciais como consulta e visualização de dados de cobertura da Terra; (h) Busca espacial: permitir que usuários encontrem dados através da sua localização espacial.

Após a análise foi possível concluir que CKAN possui vantagens em relação aos demais *frameworks*. Sendo assim, CKAN será usada em conjunto com a plataforma proposta para dar suporte a disseminação dos dados.

3. Plataforma para dados de pesquisa de observação da Terra

A arquitetura da plataforma proposta é composta por três componentes: o Portal de Dados, o Gerenciador de Dados e os Repositórios de Dados de Pesquisa, como mostrado na Figura 1.

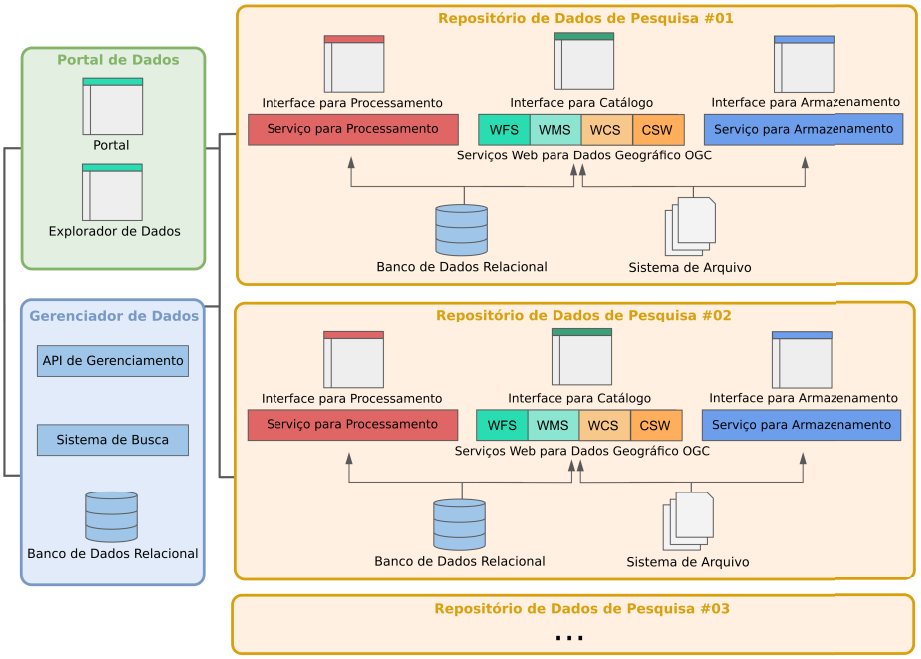

Figura 1. Arquitetura da plataforma proposta

Um Repositório de Dados de Pesquisa (RDP) é responsável por prover, para pesquisadores, ferramentas para gerenciamento, catálogo e disseminação de seus dados científicos. Isso é feito através de duas formas diferentes de armazenamento: um banco de dados relacional e um sistema de arquivos. Para disseminação um RDP conta com: um serviço e interface para análise e processamento de dados, um serviço e interface de gerenciamento e sincronia de arquivos, um grupo de *Web Services* para dados geográficos, seguindo os padrões OGC e uma interface para catalogo. Cada RDP manterá um catálogo local de metadados, acessível através do padrão CSW. Ao usar os padrões OGC pesquisadores seguirão a INDE e a INDA.

O Gerenciador de Dados é responsável pela criação, gerenciamento e entrega dos RDPs, dessa forma cada pesquisador ao criar um repositório recebe um ambiente vir-

tual com ferramentas para armazenamento, catalogação, gerenciamento, processamento e disseminação, todas prontas para uso. Esse componente também é responsável pela composição de um catálogo global de metadados, integrando todos os catálogos locais, assim permitindo buscas textuais pelos metadados de toda a plataforma.

O Portal de Dados, inspirado em bibliotecas digitais, é a interface que entrega as funcionalidades da plataforma aos pesquisadores e aos usuários, em forma de *website*. Outra interface desse mesmo componente é o Explorador de Dados, ele proverá visualização de dados georreferenciados da plataforma, com navegação baseada em mapas, gerenciamento de camadas, e outras funções de *Web* GIS.

A arquitetura com tecnologias da plataforma proposta é mostrada na Figura 2. Para a implementação serão usados apenas *Softwares* de Código Aberto.

Figura 2. Arquitetura com tecnologias da plataforma proposta

Para o Portal de Dados será construído um *website* usando Angular, uma plataforma para criação de aplicações web. Para o Explorador de Dados, será usado o TerraBrasilis[9], uma infraestrutura para disseminação de dados de desmatamento do Brasil. Para o Gerenciador de Dados, será usado CKAN, em conjunto com um *Web Service* construído em Python. Para o sistema de virtualização dos RDPs, será usado o Kubernetes, um sistema de Código Aberto para automação e implantação de contêineres, tecnologia para abstração e virtualização de ambientes.

O Kubernetes é composto por um gerenciador principal e nós, instâncias de ambiente criadas a partir de uma galeria de imagens. Para compor a galeria de imagens serão usados: (a) GeoServer, um serviço que permite compartilhar e processar dados geoespaciais seguindo os padrões OGC; (b) GeoNetwork, um aplicativo para gerenciar catálogos de recursos geográficos e edição de metadados, usando o padrão CSW; (c) OwnCloud, um serviço de Código Aberto de armazenamento e sincronização de arquivos; (d) TerraMA2[10], uma plataforma computacional para processamento e análise de dados; (e) PostgreSQL, um sistema de gerenciamento de banco de dados relacional.

[9] www.terrabrasilis.dpi.inpe.br
[10] www.terrama2.dpi.inpe.br

4. Conclusão

A crescente adoção de práticas de Ciência Aberta, por pesquisadores e instituições, sugere que repositórios de dados de pesquisa devem acompanhar os pesquisadores durante todas as suas atividades, não somente no final do processo de uma pesquisa. Nesse contexto, a arquitetura da plataforma proposta nesse trabalho é capaz de contemplar o armazenamento, catalogação, gerenciamento, processamento e disseminação de dados científicos.

Esse artigo apresenta um trabalho em andamento no desenvolvimento de uma plataforma para compartilhamento de dados científicos de observação da Terra. Com o projeto da plataforma estabelecido, o próximo passo é a implementação da plataforma proposta. A implementação dessa plataforma utilizará apenas *Softwares* de Código Aberto.

Atualmente o projeto está sendo desenvolvido em parceria com diferentes laboratórios da Coordenação-Geral de Observação da Terra (CGOBT) do Instituto Nacional de Pesquisas Espaciais (INPE).

Para a avaliação os conceitos da plataforma proposta serão feitos dois casos de estudo, o Laboratório de Investigação de Sistemas Sócio-Ambientais (Liss) e o Laboratório de Instrumentação de Sistemas Aquáticos (LabISA), dois laboratórios da CGOBT que possuem atividades em andamento no âmbito de tornar dados de pesquisa abertos.

5. Agradecimento

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Referências

Amorim, R. C., Castro, J. A., et al. (2017). A comparison of research data management platforms: architecture, flexible metadata and interoperability. *Universal Access in the Information Society*, 16(4):851–862.

Bezjak, S., Clyburne-Sherin, A., et al. (2018). *Open Science Training Handbook*. Zenodo.

Diepenbroek, M., Grobe, H., et al. (2002). Pangaea—an information system for environmental sciences. *Computers Geosciences*, 28(10):1201 – 1210.

King, G. (2007). An introduction to the dataverse network as an infrastructure for data sharing. *Sociological Methods Research*, 36(2):173–199.

OGC (2017). Ogc standards and supporting documents. `http://www.opengeospatial.org/standards`.

Saez, R. V. and Fuentes, C. M. (2018). Open science now: A systematic literature review for an integrated definition. *Journal of Business Research*, 88:428 – 436.

Sicilia, M.-A., García-Barriocanal, E., et al. (2017). Community curation in open dataset repositories: Insights from zenodo. *Procedia Computer Science*, 106:54 – 60.

Wainwright, M. (2012). Using ckan: storing data for re-use. `https://ckan.org/files/2012/08/OKF-OR12-poster.pdf`. Accessed: 2019-03-21.

Woelfle, M., Olliaro, P., and Todd, M. H. (2011). Open science is a research accelerator. *Nature Chemistry*, 3:745 EP –.