sid.inpe.br/mtc-m21c/2020/10.01.13.13-TDI

TERRABRASILIS RESEARCH DATA: UMA PLATAFORMA PARA COMPARTILHAMENTO DE DADOS CIENTÍFICOS GEOESPACIAIS

Gabriel Sansigolo

Dissertação de Mestrado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Gilberto Ribeiro de Queiroz e Karine Reis Ferreira, aprovada em 09 de outubro de 2020.

URL do documento original:
<http://urlib.net/8JMKD3MGP3W34R/43BKP28>

INPE
São José dos Campos
2020

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GBDIR)
Serviço de Informação e Documentação (SESID)
CEP 12.227-010
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/7348
E-mail: pubtc@inpe.br

sid.inpe.br/mtc-m21c/2020/10.01.13.13-TDI

TERRABRASILIS RESEARCH DATA: UMA PLATAFORMA PARA COMPARTILHAMENTO DE DADOS CIENTÍFICOS GEOESPACIAIS

Gabriel Sansigolo

Dissertação de Mestrado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Gilberto Ribeiro de Queiroz e Karine Reis Ferreira, aprovada em 09 de outubro de 2020.

URL do documento original:
<http://urlib.net/8JMKD3MGP3W34R/43BKP28>

INPE
São José dos Campos
2020

Dados Internacionais de Catalogação na Publicação (CIP)

Sansigolo, Gabriel.

Sa58t TerraBrasilis Research Data: uma plataforma para compartilhamento de dados científicos geoespaciais / Gabriel Sansigolo. – São José dos Campos : INPE, 2020.

xviii + 72 p. ; (sid.inpe.br/mtc-m21c/2020/10.01.13.13-TDI)

Dissertação (Mestrado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2020.

Orientadores : Drs. Gilberto Ribeiro de Queiroz e Karine Reis Ferreira.

1. Ciência aberta. 2. Gerenciamento de dados. 3. Integração de dados. 4. Dados de observação da Terra. I.Título.

CDU 528:001.103

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÕES

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

DEFESA FINAL DE DISSERTAÇÃO DE GABRIEL SANSIGOLO

No dia 09 de outubro de 2020, às 09h, por videoconferência, o(a) aluno(a) mencionado(a) acima defendeu seu trabalho final (apresentação oral seguida de arguição) perante uma Banca Examinadora, cujos membros estão listados abaixo. O(A) aluno(a) foi APROVADO(A) pela Banca Examinadora por unanimidade, em cumprimento ao requisito exigido para obtenção do Título de Mestre em Computação Aplicada. O trabalho precisa da incorporação das correções sugeridas pela Banca Examinadora e revisão final pelo(s) orientador(es).

Novo Título: TerraBrasilis Research Data: uma plataforma para compartilhamento de dados científicos geoespaciais.

Eu, Rafael Duarte Coelho dos Santos, como Presidente da Banca Examinadora, assino esta ATA em nome de todos os membros.

Membros da Banca

Dr. Rafael Duarte Coelho dos Santos, Presidente INPE.
Dra. Karine Reis Ferreira, Orientador(a) INPE.
Dr**.** Gilberto Ribeiro de Queiroz, Orientador(a) INPE.
Dra. Silvana Amaral Kampel, Membro da Banca INPE.
Dr. Clodoveu Augusto Davis Junior, Convidado (a) UFMG.

Documento assinado eletronicamente por **Rafael Duarte Coelho dos Santos**, **Tecnologista**, em 15/10/2020, às 12:14 (horário oficial de Brasília), com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.mctic.gov.br/verifica.html, informando o código verificador **5989108** e o código CRC **1B16EAE7**.

Referência: Processo nº 01340.007102/2020-10 SEI nº 5989108

AGRADECIMENTOS

Agradecemos à Gabriel Crivellaro Gonçalves e Felipe Menino Carlos que gentilmente cederam parte do tempo para ajudar no desenvolvimento desse projeto. Agradecemos também aos pesquisadores dos laboratórios LiSS (Laboratório de Investigação Sistemas Socioambientais) e LabISA (Laboratório de Instrumentação de Sistemas Aquáticos) pela disponibilização dos dados para os estudos de caso. Agradecemos igualmente a todos que contribuíram para desenvolvimento de projeto.

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

RESUMO

Ciência Aberta é um 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 Abertos, Pesquisa Aberta Reprodutiva, entre outras. Com o crescimento de popularidade de Dados Abertos, diferentes infraestruturas de dados e políticas nos âmbitos nacional, federal e institucional foram criadas, como a Infraestrutura Nacional de Dados Abertos (INDA) e a Infraestrutura Nacional de Dados Espaciais (INDE). A demanda por uma abertura de dados também pode ser observada 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, componente hoje obrigatório para submissão de um projeto. Nesse cenário, pesquisadores precisam de uma plataforma para compartilhar dados científicos de observação da Terra. Atualmente, pesquisadores têm utilizado algumas plataformas existentes para armazenar dados científicos como o PANGAEA e o Zenodo. Porém, essas plataformas são fechadas e foram criadas para resolver problemas de armazenamento e preservação de dados, não possuindo integração com ferramentas usadas por pesquisadores durante suas atividades de pesquisa e análise. Nesse contexto existe uma demanda por uma plataforma para gerenciamento de dados que forneça, de maneira integrada e automatizada, diferentes tecnologias para produção, processamento, gerenciamento e disseminação de dados. Para atender essa demanda, esse trabalho busca projetar e desenvolver uma plataforma chamada TerraBrasilis Research Data, para o compartilhamento de dados científicos geoespaciais. O objetivo dessa plataforma é integrar ferramentas para armazenamento, catalogação, gerenciamento, processamento e disseminação de dados. Ela fornece facilidades para a comunicação entre os dados publicados pelos pesquisadores e infraestruturas de dados estabelecidas, como INDA e INDE, Sistemas de Informações Geográficas (SIG) e outras ferramentas. Para testar e validar a plataforma proposta, foram usados dados de dois laboratórios da Coordenação-Geral de Observação da Terra (CGOBT) do Instituto Nacional de Pesquisas Espaciais (INPE): o Laboratório de Investigação Sistemas Socioambientais (LiSS) e o Laboratório de Instrumentação de Sistemas Aquáticos (LabISA). Conclui-se que a abordagem modular em conjunto com as tecnologias e serviços abertos utilizados permitiu a criação de uma plataforma web capaz de suportar diferentes tipos de dados de observação da Terra. O TerraBrasilis Research Data é uma plataforma que resolve a demanda por automatização no processo de disseminação de dados de gerenciadores de dados, assim contemplando todas as atividades de uma pesquisa.

Palavras-chave: Ciência aberta. Gerenciamento de dados. Integração de dados. Dados de observação da Terra.

TERRABRASILIS RESEARCH DATA – PLATFORM FOR SHARING SCIENTIFIC GEOSPATIAL DATA

ABSTRACT

Open Science is a set of practices, tools and policies created to allow research collaboration and sharing. This includes a variety of practices such as: Open Access, Open Data, Open Reproductive Research, among others. With the growing popularity of Open Data, different data infrastructures and policies at the national, federal and institutional levels were created, such as the National Open Data Infrastructure (INDA) and the National Spatial Data Infrastructure (INDE). The demand for opening data can also be observed at the São Paulo State Research Support Foundation (FAPESP), which promotes Open Science practices, created a data management plan, a mandatory component for submitting a new project. In this scenario, researchers need a platform to share scientific data from Earth observation. Currently, researchers have used some existing platforms to store scientific data like PANGAEA and Zenodo. However, these platforms are closed and were created to solve problems of data storage and preservation, and have no integration with tools used by researchers during their research and analysis activities. In this context, there is a demand for a data management platform that provides, in an integrated and automated way, technologies for data production, processing, management and dissemination. To meet this demand, this work seeks to design and develop a platform called TerraBrasilis Research Data, for sharing geographic research data. This platform integrates tools for data storage, cataloging, management, processing and dissemination. It provides simple communication between data published by researchers and established data infrastructures, such as INDA and INDE, Geographic Information Systems (GIS) and other tools. To test and validate the proposed platform, data were used from two laboratories of the General Coordination for Earth Observation (CGOBT) of the National Institute for Space Research (INPE): The Socio-Environmental Systems Research Laboratory (LiSS) and the Laboratory of Instrumentation of Aquatic Systems (LabISA). We conclude that the modular approach, in conjunction with technologies and open services, allowed the creation of a web platform capable of supporting different types of Earth observation data. TerraBrasilis Research Data is a platform that solves the demand for automation in the data dissemination process of data managers, thus covering all research activities.

Keywords: Open science. Data management. Data integration. Earth observation data.

LISTA DE FIGURAS

Pág.

LISTA DE TABELAS

Pág.

LISTA DE ABREVIATURAS E SIGLAS

FAPESP	–	Fundação de Amparo a Pesquisa de São Paulo
CRUESP	–	Conselho de Reitores das Universidades Estaduais de São Paulo
USP	–	Universidade de São Paulo
UNICAMP	–	Universidade de Campinas
UNESP	–	Universidade Estadual Paulista
INPE	–	Instituto Nacional de Pesquisas Espaciais
OBT	–	Coordenação-Geral de Observação da Terra
RDA	–	Research Data Alliance
INDA	–	Infraestrutura Nacional de Dados Aberto
INDE	–	Infraestrutura Nacional de Dados Espaciais
PDA	–	Plano de Dados Abertos
FAIR	–	*Findability*, *Accessibility*, *Interoperability* e *Reusability*
SDI	–	*Spatial Data Infrastructure*
SIG	–	Sistemas de Informação Geográfica
ISO	–	*International Organization for Standardization*
OGC	–	*Open Geospatial Consortium*
WMS	–	*Web Map Server*
WFS	–	*Web Feature Service*
WCS	–	*Web Coverage Service*
CSW	–	*Catalogue Service Web*
SHP	–	*Shapefile*
MGB	–	Metadados Geoespaciais do Brasil
CSV	–	*Comma-separated values*
XLS	–	*Excel Spreadsheet*
TIFF	–	*Tagged Image File Format*
URN	–	*Uniform Resource Name*
DOI	–	*Digital Object Identifier*
PURL	–	*Persistent URL*
ARK	–	*Archive Resource Key*
JDDCP	–	*Joint Declaration of Data Citation Principles*
CKAN	–	*Comprehensive Knowledge Archive Network*
OKF	–	*Open Knowledge Foundation*
DETER	–	Sistema de Detecção do Desmatamento em Tempo Real
PRODES	–	Sistema de Monitoramento do Desmatamento
LabISA	–	Laboratório de Instrumentação de Sistemas Aquáticos
LiSS	–	Laboratório de Investigação Sistemas Socioambientais

SUMÁRIO

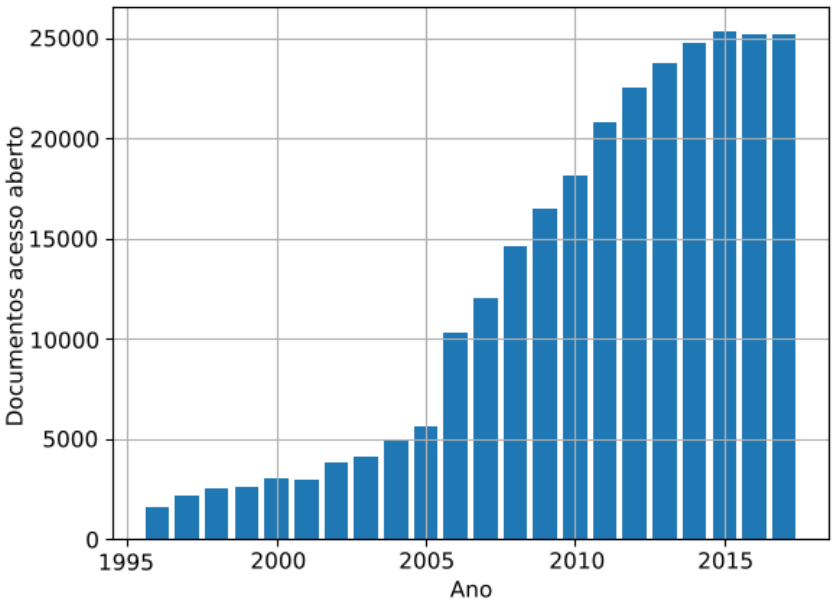
Pág.

1 INTRODUÇÃO

Ciência Aberta é um 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 demais artefatos de uma pesquisa são disponibilizados de maneira aberta com objetivo de facilitar o seu reúso, redistribuição e reprodução (SAEZ; FUENTES, 2018). Com o objetivo de entender o estado da Ciência Aberta no Brasil, é necessário antes entender o estado das práticas que a compõem.

O Brasil se destaca no cenário internacional em relação a políticas de acesso aberto. Em 2017 o Brasil publicou 25.211 artigos em revistas de acesso ouro, revistas que oferecem acesso *online* gratuito aos artigos (Scimago Institutions Rankings, 2019). Permitir o acesso a esse grande volume de informações é fundamental para o crescimento da ciência, pois pesquisadores, dessa forma, conseguem acompanhar as pesquisas que estão sendo feitas em suas respectivas áreas. Na Figura 1.1 é possível observar o crescimento ano a ano do número de artigos produzidos no Brasil. Através desses números e da crescente demanda por repositórios institucionais no mundo (LYNCH, 2003; WILKINSON et al., 2016), conclui-se que existe uma necessidade por repositórios institucionais no Brasil.

Figura 1.1 - Número de artigos publicados no Brasil por ano



Fonte: Scimago Institutions Rankings (2019).

Essa demanda pode ser confirmada através de políticas criadas nos últimos anos com propósito de incentivar o armazenamento de artigos científicos e dados resultantes de pesquisas. Um exemplo disso pode ser visto nas novas políticas da Fundação de Amparo a Pesquisa de São Paulo[1] (FAPESP). Em 2013 o conselho da FAPESP aprovou uma política de acesso aberto. A política estabelece que em artigos resultantes de pesquisas financiadas pela fundação e publicadas em revistas internacionais indexadas em banco de dados como *Web of Science* devem ser armazenados em repositórios de acesso aberto (PIERRO, 2019). Em 2013 a política resultou na criação do Repositório da Produção Científica do Conselho de Reitores das Universidades Estaduais de São Paulo (CRUESP[2]), um repositório de artigos, teses, dissertações e outros documentos científicos da Universidade de São Paulo (USP), Universidade de Campinas (UNICAMP) e Universidade Estadual Paulista (UNESP).

Em 2019, a FAPESP evoluiu a sua política de acesso aberto, criando uma política para acesso a dados resultados de pesquisas (Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP, 2019). A política determina que autores de projetos e bolsas financiados pela fundação devem divulgar seus resultados em periódicos que permitam o arquivamento de uma cópia do trabalho em um repositório público, que possa ser consultado na web por qualquer pessoa.

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 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 de pesquisa.

1.1 Motivação

Nesse cenário, pesquisadores precisam de uma plataforma para compartilhar dados científicos. Hoje existem diferentes plataformas para publicação de dados científicos, como o *PANGAEA*[5], uma plataforma para publicação de conjuntos de dados geo-

[1] www.fapesp.br
[2] www.cruesp.sp.gov.br
[3] www.nature.com/sdata
[4] www.fapesp.br/gestaodedados
[5] www.pangaea.de

científicos (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 de armazenamento e preservação de dados. Devido a suas características fechadas, não possuem integração com ferramentas usadas por pesquisadores durante as atividades de uma pesquisa. Nesse contexto, existe uma demanda por uma plataforma para gerenciamento de dados que forneça, de maneira integrada e automatizada (WEIGEL, 2015), diferentes tecnologias para produção, processamento, gerenciamento e disseminação de dados, assim contemplando todas as atividades de uma pesquisa.

No contexto da área de observação da Terra, pesquisadores do Instituto Nacional de Pesquisas Espaciais (INPE) criaram plataformas para disseminação de dados. Alguns exemplos são o TerraBrasilis[7], uma infraestrutura para organização, acesso e uso dos dados geográficos de monitoramento ambiental, o TerraMA2[8], uma plataforma computacional para atender uma demanda crescente de aplicações de monitoramento, análise e alerta em áreas como qualidade do ar, qualidade da água, incêndios florestais, enchentes entre outras, e o Pauliceia[9] uma plataforma computacional para disseminação de dados históricos colaborativos. Apesar desses esforços, os laboratórios da Coordenação-Geral de Observação da Terra (OBT) do INPE não possuem uma plataforma para disseminação de dados científicos, o que poderia ser útil para compartilhar os dados científicos produzidos.

1.2 Hipótese

A crescente adoção das práticas de Ciência Aberta por pesquisadores vem acompanhada da necessidade de plataformas computacionais voltadas para disseminação de dados científicos. Essas plataformas devem ser apoiadas nos conceitos de Dados Abertos, em protocolos de interoperabilidade entre sistemas, e devem facilitar a adoção de normas e padrões de dados estabelecidos, para possibilitar que os pesquisadores compartilhem o resultado de seus trabalhos de maneira efetiva.

1.3 Objetivo

Esse trabalho projeta e desenvolve uma plataforma para compartilhamento de dados científicos geoespaciais. O objetivo principal é criar uma plataforma para gerenciamento de dados que integre ferramentas para armazenamento, catalogação, gerenci-

[6] zenodo.org
[7] www.terrabrasilis.dpi.inpe.br
[8] www.terrama2.dpi.inpe.br
[9] www.pauliceia.dpi.inpe.br

amento, processamento e disseminação de dados. Chamada TerraBrasilis Research Data, a plataforma se baseia nos princípios de Ciência Aberta e foi validada com dados científicos produzidos por pesquisadores da OBT-INPE.

Assim, subdividimos nosso objetivo principal em duas partes:

- **Criação de um modelo de arquitetura para um gerenciador de dados geoespaciais científicos**

  Esse modelo deve integrar ferramentas para armazenamento, catalogação, gerenciamento, processamento e disseminação. Esse modelo tem como objetivo contemplar todas as atividades de uma pesquisa, não somente armazenamento.

  Para isso iremos estudar modelos de arquitetura usado em gerenciadores de dados estabelecidos. E formas de entregar tecnologias para cada uma das respectivas funcionalidades.

- **Implementação do modelo proposto através da plataforma TerraBrasilis Research Data**

  Um dos objetivos do trabalho é implementar o modelo proposto em uma plataforma computacional para compartilhamento de dados científicos geoespaciais. Essa implementação tem como objetivo entregar aos pesquisadores e usuário todas as funcionalidades providas pela plataforma.

  Portanto para esse segundo objetivo será implementado cada um dos componentes que compõe o modelo de arquitetura. E para alguns casos, implementaremos também funcionalidades para facilitar que pesquisadores publiquem seus dados seguindo infraestruturas estabelecidas.

- **Validação do modelo proposto e da plataforma implementada**

  Para testar e validar o modelo proposto e a plataforma implementada, um dos objetivo desse trabalho é fazer um experimentos com dados geoespaciais científicos produzidos por pesquisadores da OBT-INPE.

  Através desse objetivo determinaremos de forma analítica se a plataforma proposta é integradora e independente de modelos de dados. E determinaremos quais são as ameaças à validade da plataforma.

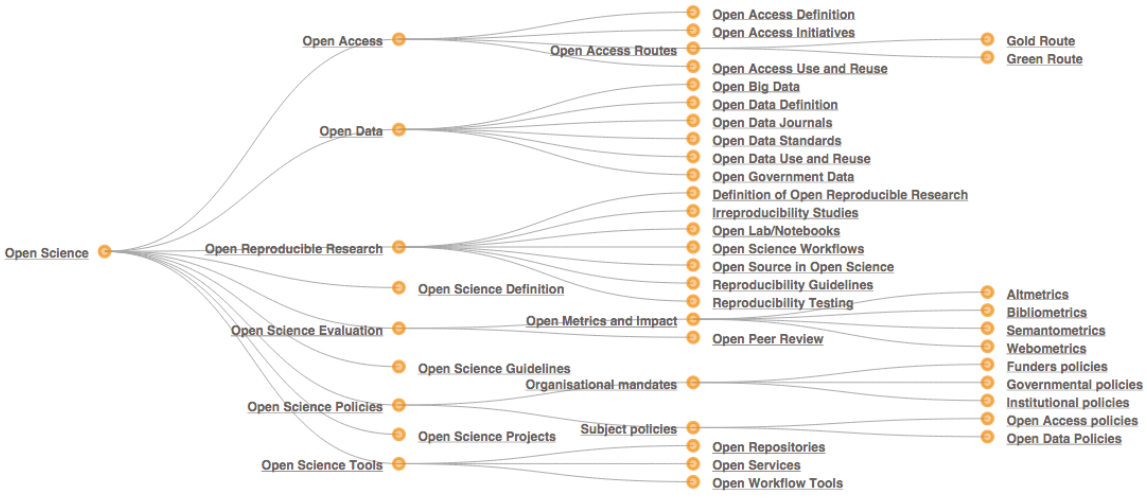
2 REVISÃO DA LITERATURA

Ciência Aberta é a prática da ciência que busca permitir a colaboração e contribuição em pesquisas pela sociedade, pois dados, anotações e outros processos de uma pesquisa estão abertos a reuso, redistribuição e reprodução (FOSTER, 2019; SAEZ; FUENTES, 2018). Esse é um importante mecanismo para a evolução do conhecimento.

A Ciência Aberta é composta por uma variedade de práticas, como: acesso aberto a publicações, acesso aberto a dados de pesquisa, softwares de código aberto e muitas outras (PONTIKA et al., 2015; WOELFLE et al., 2011; BEZJAK et al., 2018). Existem diversas definições do termo "aberto" no contexto da ciência. A *Open Knowledge Foundation* (OPEN KNOWLEDGE FOUNDATION, 2019) define dados abertos como quando o dado pode ser usado, editado e compartilhado gratuitamente, por qualquer um e para qualquer propósito. A mesma definição pode ser estendida para as demais práticas que compõem a Ciência Aberta.

A Figura 2.1 mostra as diversas ramificações da Ciência Aberta. Dados Abertos, como mostrado nessa figura, é um dos conjuntos de práticas que compõem o conceito de Ciência Aberta.

Figura 2.1 - Taxonomia da Ciência Aberta



Fonte: Pontika et al. (2015).

2.1 Dados abertos

Dados Abertos são o produto da ideia de que alguns tipos de dados devem estar disponíveis livremente para todos os tipos de usuários, para uso e reprodução sem restrições de cópia, patente ou mecanismo de controle (AUER et al., 2007). Esta ideia cresceu em popularidade na última década.

A demanda por dados abertos na pesquisa vem crescendo exponencialmente na última década. Em 2013 foi fundado o Research Data Alliance (RDA), uma iniciativa global com objetivo de promover a interoperabilidade de dados através de países, produtores e consumidores de dados (WEIGEL, 2015). Para isso, a iniciativa busca permitir que pesquisadores compartilhem abertamente dados, tecnologias e disciplinas entre países. Em janeiro de 2020 a RDA contava com 9.600 membros de 137 países (RESEARCH DATA ALLIANCE, 2020).

Em 2014, a *Nature*, uma famosa editora científica multidisciplinar, criou a *Scientific Data*[1], uma revista científica de acesso aberto cujo o foco é a descrição de conjuntos de dados relevantes para as ciências naturais. Através da *Scientific Data*, a *Nature* começou a exigir a disponibilidade de materiais e dados, formas de replicar a pesquisa e comprovar as afirmações publicadas pelos autores. Para conjuntos de dados de ciências da Terra, a *Nature* recomenda o PANGAEA, uma plataforma para publicação de dados geocientíficos e ambientais. O PANGAEA[2] é uma plataforma de acesso aberto para arquivar, publicar e distribuir dados georreferenciados de pesquisas de sistemas da Terra (DIEPENBROEK et al., 2002).

A plataforma PANGAEA é um compartilhador de dados de sistemas da Terra estabelecido. Hoje a comunidade científica de sistemas da Terra tem a usado como principal ferramenta de compartilhamento de conjunto de dados. Um motivo para este uso reside na possibilidade de exportação de citações, para uso científico.

Com o crescimento da demanda por dados de pesquisa abertos, novas iniciativas e diretrizes surgiram. Uma delas é os princípios FAIR, descrito como os quatro princípios fundamentais do bom gerenciamento de dados (WILKINSON et al., 2016). FAIR é um acrônimo das palavras: (a) *Findability*, Encontrabilidade, facilidade para se encontrar os dados; (b) *Accessibility*, Acessibilidade, facilidade para acesso aos dados (c) *Interoperability*, Interoperabilidade, facilitar a troca/utilização de dados; (d) *Reusability*, Reusabilidade, tornar os dados e subprodutos do ciclo de vida de desen-

[1] www.nature.com/sdata
[2] www.pangaea.de

volvimento fáceis de usar. Os princípios FAIR surgiram de uma necessidade urgente para melhorar a infraestrutura para suporte a reutilização de dados escolares. O procedimento para implementar o guia de princípios é composto por escolhas, não por sugestões de tecnologias, padrões ou implementações. Hoje esses princípios são usados como requisitos para desenvolvimento de plataformas de Dados Abertos para comunidade científica.

2.2 Infraestrutura de dados espaciais

Infraestrutura de Dados Espaciais, em inglês *Spatial Data Infrastructure* (SDI), é uma infraestrutura digital para promoção de compartilhamento e consumo de dados que implementa suporte para dados geográficos, metadados e ferramentas para uso dos dados espaciais. Desta forma uma SDI facilita a interação entre pessoas e dados (RAJABIFARD et al., 2004). Uma SDI como plataforma visa facilitar o acesso e a integração de dados espaciais com múltiplas fontes dentro de uma estrutura com vários componentes tecnológicos.

Desde o início da década de 2000, a comunidade de Sistemas de Informação Geográfica (SIG) fez esforços para promover a interoperabilidade de dados espaciais. A *International Organization for Standardization* (ISO) e o *Open Geospatial Consortium* (OGC) propuseram padrões para representar e armazenar informações espaciais em arquivos de dados e sistemas de bancos de dados, bem como para servir dados espaciais, metadados e processos via *Web Services* (OGC, 2017). Os padrões podem ser agrupados em três grupos: dados, metadados e processamento. As especificações de serviço para dados espaciais incluem: o *Web Map Server* (WMS) fornece uma interface para solicitar imagens de mapas registrados geograficamente, a partir de um ou mais bancos de dados; o *Web Feature Service* (WFS) fornece acesso à informação geográfica contidas nas *features*; o *Web Coverage Service* (WCS) fornece acesso multidimensional à cobertura de dados e o *Catalogue Service Web* (CSW), um padrão para publicar e pesquisar coleções de metadados para dados espaciais, serviços e objetos relacionados. A especificação Infraestrutura Nacional de Dados Espaciais (INDE) de metadados geográficos, estabelecida através da ISO 19115:2003[3] e revisada via ISO 19115-1:2014[4], é baseada nos padrões OGC *Web Services*.

Com o crescimento da popularidade de dados abertos e a fim de incentivar a demanda pelos mesmos, diferentes infraestruturas de dados e políticas nos âmbitos nacional,

[3] www.iso.org/standard/26020.html
[4] www.iso.org/standard/53798.html

federal e institucional foram criadas. Em 2016 através do Decreto No. 8777/2016[5] uma política de dados abertos foi instaurada no Brasil. A mesma é composta por uma série de documentos, de planejamento e orientação, que definem, no âmbito executivo federal, ações para fazer com que dados governamentais fiquem disponíveis para a sociedade. Com objetivo de incentivar a demanda por dados governamentais abertos, a Infraestrutura Nacional de Dados Abertos (INDA) foi criada. A INDA define padrões, tecnologias e procedimentos para satisfazer as condições necessárias para disseminação de dados e informações públicas seguindo o modelo de Dados Abertos. Um componente chave é o portal de dados abertos do Brasil[6], que tem o objetivo de ser um ponto central para uso, busca e acesso a dados governamentais do Brasil.

Apesar do plano de dados abertos ter sido protocolado em 2016, os conceitos por trás dele se baseiam em iniciativas passadas. Uma delas, no âmbito nacional, foi o Decreto No. 6.666/08[7], que exige que todas as instituições governamentais do poder executivo federal compartilhem e divulguem seus dados geoespaciais e seus metadados na Infraestrutura Nacional de Dados Espaciais (INDE). A INDE[8] é um portal de compartilhamento e disseminação de dados geoespaciais criado com o propósito de catalogar e integrar os dados geoespaciais produzidos pelas instituições governamentais do Brasil.

Em 2018 por meio da Portaria No. 307[9], o Plano de Dados Abertos (PDA) do INPE foi instituído. Composto por um documento orientador, o plano, através de ações de implementação e promoção, busca promover a abertura de dados. Um dos objetivos é garantir que conjuntos de dados produzidos pelo instituto sejam apropriadamente catalogados e publicados nos portais INDA, INDE e em portais próprios do instituto. Iniciativas de Dados Abertos também podem ser encontradas no âmbito da ciência. A FAPESP, reconhecendo a importância de uma gestão adequada dos dados de pesquisa como parte essencial das boas práticas, criou um Plano de Gestão de Dados[10]. O mesmo vem se tornando componente obrigatório na fase de submissão de um projeto.

[5] www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm
[6] dados.gov.br
[7] www.planalto.gov.br/ccivil_03/_Ato2007-2010/2008/Decreto/D6666.htm
[8] www.inde.gov.br
[9] www.inpe.br/dados_abertos/arquivos/port307v2018.pdf
[10] www.fapesp.br/gestaodedados

2.3 Identificadores persistentes

Com crescimento em volume, variedade e número de objetos os gerenciadores de dados precisam automatizar os seus processos de publicação e acesso. Para conseguir informações confiáveis de forma automatizada, usa-se identificadores persistentes para identificar objetos digitais (WEIGEL, 2015). Uma característica de identificadores é a capacidade de referenciar e acessar objetos digitais independentes do armazenamento local original, assim permitindo acesso automatizado ou humano a informações. Identificadores persistentes são componentes importantes de infraestruturas científicas digitais.

Um dos objetivos do uso de identificadores persistentes é a preservação a longo prazo, deixando a informação digital útil, passível de leitura e entendimento, independente do autor original. Alguns identificadores persistentes são: *Uniform Resource Name*[11] (URN), o *Persistent URL*[12] (PURL), *Digital Object Identifier*[13] (DOI) e *Archive Resource Key*[14] (ARK).

Digital Object Identifier

Identificador de Objeto Digital, em inglês *Digital Object Identifier* (DOI) é um identificador persistente para objetos únicos que atribui um número único e exclusivo a todo material publicado (textos, imagens, etc). Para geração de identificadores DOI é necessária a assinatura de um plano nos serviços do *DOI Registration Agency*.

As características do sistema DOI são: (a) unicidade, um identificador referência uma única entidade; (b) resolução, um serviço que receba o identificador e retorne informações sobre a entidade; (c) persistência, uma vez criado, um identificador permanece apontando para a mesma entidade para sempre; e (d) interoperabilidade, permite o uso do identificador fora do controle do responsável pelo identificador.

Persistent URL

URLs Persistentes (PURL) são endereços da Web ou URLs que atuam como identificadores permanentes em espaço de infraestrutura. Em vez de resolver diretamente os recursos da Web, as PURLs fornecem um nível de indireção que permite que os endereços da Web subjacentes sejam alterados ao longo do tempo. PURL é uma

[11] tools.ietf.org/html/rfc8141
[12] purlz.org
[13] www.doi.org
[14] n2t.net/e/ark_ids.html

solução de código aberto semelhante ao DOI.

2.4 Frameworks para bibliotecas digitais

Bibliotecas digitais, também chamadas de repositórios institucionais, são ferramentas criadas para prover suporte para disseminação de produtos de conhecimento (LYNCH, 2003; AMORIM et al., 2017). O objetivo principal de uma biblioteca digital é preservar o conhecimento produzido pela humanidade, para isso buscam-se preservar, além de dados, artigos científicos relacionados aos dados, repositórios com os materiais necessários para a reprodução do experimento, entre outros produtos, criando assim um ambiente de conhecimento (BEZJAK et al., 2018).

Uma biblioteca digital é um conjunto de serviços que uma universidade ou instituições de pesquisas oferecem a membros de sua comunidade para a gestão e disseminação de materiais digitais criados pela instituição e por membros da comunidade (LYNCH, 2003). Para a criação de um ambiente para gestão e disseminação de materiais digitais é necessário um *framework*. No âmbito de criação de bibliotecas digitais, dois *frameworks* se destacam: o Invenio e o Dataverse. Com funcionalidades semelhantes, um terceiro *framework* que se destaca é o CKAN, utilizada para a criação de portais de dados abertos.

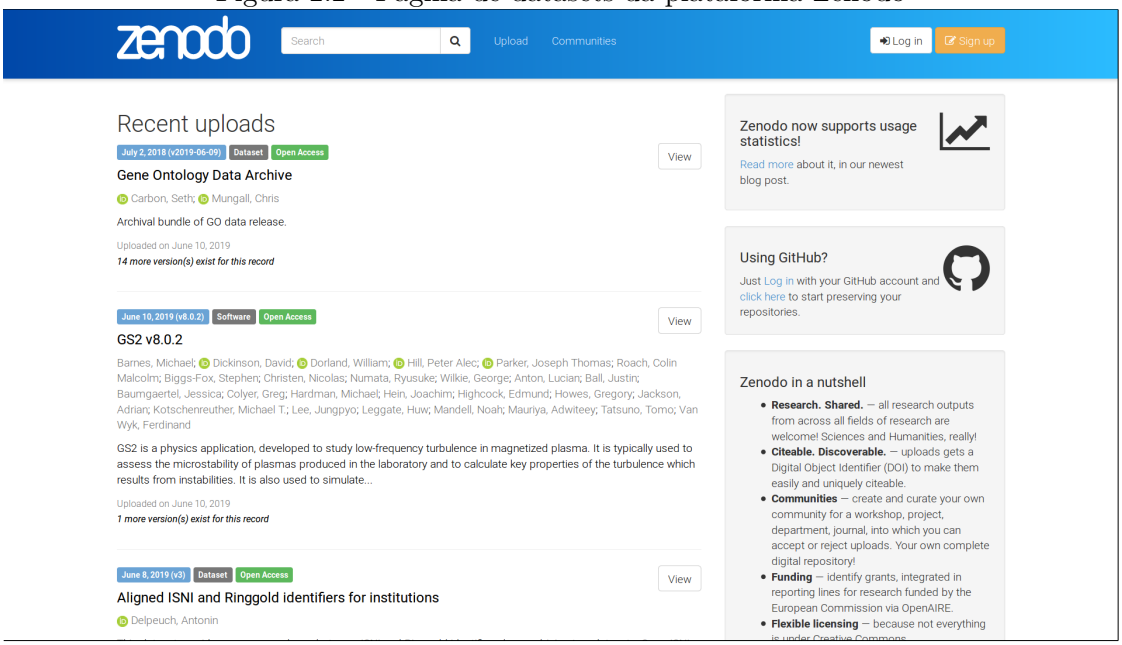
2.4.1 Invenio

Invenio é um *framework* aberto para construção de bibliotecas digitais de grande escala. Características chave do *framework* são: um modelo de dados flexível, um mecanismo de pesquisa integrado e um sistema de gerenciamento de arquivos digitais (Invenio, 2019). A galeria de instâncias do *framework* é principalmente composta por plataformas relacionadas à Organização Europeia para a Pesquisa Nuclear, conhecida como CERN[15].

O Zenodo[16] é um repositório online hospedado no CERN, uma instância do *framework* Invenio. Lançado em maio de 2013, o repositório Zenodo foi especificamente projetado para ajudar pesquisadores a compartilharem resultados em uma ampla variedade de formatos em todos os campos da ciência (SICILIA et al., 2017). Algumas comunidades usam o Zenodo em seus fluxos de trabalho de arquivamento, aproveitando também a integração com a plataforma Github (HERTERICH; DALLMEIER-TIESSEN, 2016). Na Figura 2.2 é mostrada a página de datasets da plataforma Zenodo.

[15] `home.cern`
[16] `zenodo.org`

Figura 2.2 - Página de datasets da plataforma Zenodo



Fonte: Invenio (2019).

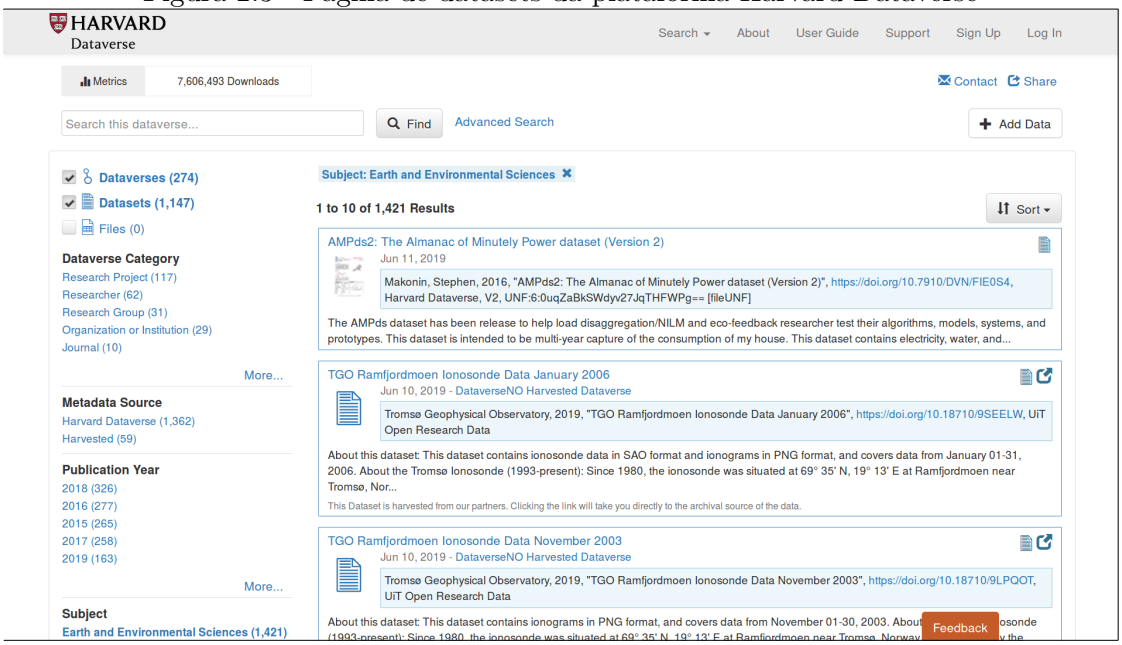
Na Figura 2.2 é possível observar etiquetas como datasets e softwares. Isso é possível através do modelo de dados flexível da plataforma que, usando um JSON Schema, uma estrutura que permite a descrição de vários tipos de dados, permite a serialização de *softwares*, datasets, teses ou outros, para a produção de citações em BibTeX, DataCite e outros. O Invenio usa Elasticsearch nativo, um mecanismo de busca distribuído nativo JSON que suporta quaisquer dados textuais, assim permitindo uma busca não só através de metadados mas também do conteúdo de dados textuais. Em categoria de preservação, o *framework* usa Identificadores de Objetos Digitais (DOI) para os ativos do repositório.

Além de ser um repositório online, o Zenodo também é um serviço para gerenciamento de dados de pesquisa. Com o pagamento de uma taxa mensal, pesquisadores podem usufruir de todas as funcionalidades do *framework* através da instância Zenodo. Usando o Zenodo como serviço contratado, o armazenamento deixa de ser local e se torna totalmente remoto, e o modelo de dados deixa de ser flexível e se torna fixo.

2.4.2 Dataverse

Dataverse é um *framework* código aberto para criação de plataformas *web* para compartilhamento, preservação, citação, exploração e análise de dados de pesquisa (KING, 2007). Com o foco em ser um serviço pronto para uso por instituições, o Dataverse possui instâncias ao redor do mundo. Uma de suas características chave é a possibilidade de gerar uma citação formal para cada depósito (Harvard University, 2019). Proposto em 2007, o Dataverse foi responsável pela padronização de arquiteturas de infraestruturas de compartilhadores de dados de pesquisa. Na Figura 2.3 é mostrada a página de datasets do Harvard Dataverse[17], uma instância do *framework* dataverse. Atualmente o repositório conta com mais de 60,000 datasets.

Figura 2.3 - Página de datasets da plataforma Harvard Dataverse



Fonte: Harvard University (2019).

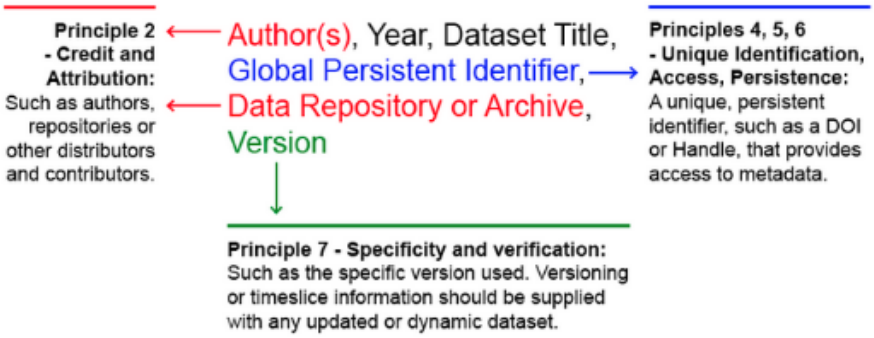
Na Figura 2.3 observa-se uma estrutura organizada, com diferentes formas de filtragem e métricas. Isso, somado à robusta infraestrutura para compartilhamento de dados do *framework*, tornam o Dataverse uma das soluções mais procuradas. No artigo de introdução à infraestrutura do Dataverse (KING, 2007) são apresentados os oito principais requisitos da plataforma. Requisitos para melhorar infraestruturas de compartilhamento de dados, são eles: (a) reconhecimento, ajudando autores a

[17] `dataverse.harvard.edu`

disponibilizar dados associados aos seus artigos publicados; (b) distribuição pública, deixando dados publicados por autores disponíveis para comunidade; (c) autorização, permitir que dados não livres ou abertos sejam acessíveis através de contato com autor; (d) validação, permitir a verificação dos conjuntos específicos de dados; (e) verificação, garantia que a associação entre dados e artigos publicados seja mantida fixa; (f) persistência, garantir que mesmo depois de décadas seja possível de encontrar os dados e as associações; (g) facilidade de uso, tornar fácil o mantimento dos *softwares* e *hardwares* necessários para que profissionais arquivem seus dados; e (h) proteção legal, garantir que os dados não serão editados;

Os requisitos citados anteriormente foram responsáveis por tornar o *framework* a solução estabelecida que ela é hoje. Um exemplo de conceitos padronizados pelo Dataverse é a citação de conjuntos de dados, seguindo a Declaração Conjunta de Princípios de Citação de Dados, do inglês *Joint Declaration of Data Citation Principles* (JDDCP) (MARTONE, 2014). Baseado em iniciativas anteriores, o Dataverse adicionou padrões para citação de dados, como mostra a Figura 2.4.

Figura 2.4 - Citação de dados no Dataverse baseada no JDDCP



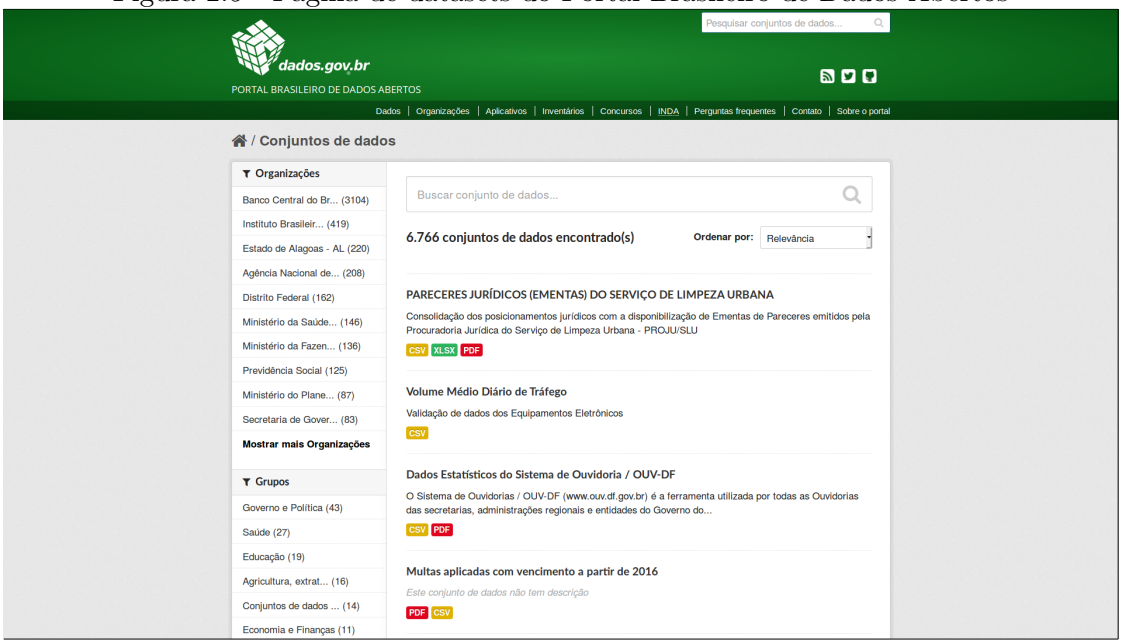
Fonte: Martone (2014).

Outro conceito padronizado pelo Dataverse foi a arquitetura tecnológica de repositórios de dados de pesquisas: (a) uma *API* para orquestração do sistema e exportação das funcionalidades para usuários de baixo nível; (b) um banco de dados relacional para armazenamento dos dados e dos metadados e (c) um motor de busca para realização de busca textual através dos metadados e do conteúdo textual dos datasets. Essa arquitetura evoluiu para outros *frameworks* incluindo diferentes tecnologias, como estruturas de gerenciamento de dados na memória e *software* de mensagem.

2.4.3 CKAN

A Rede Abrangente de Arquivos de Conhecimento, do inglês *Comprehensive Knowledge Archive Network* (CKAN), é um pacote de código aberto para criação de *hubs* 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 e disponíveis (WAINWRIGHT, 2012). Diferente dos *frameworks* citados anteriormente, o CKAN tem o foco em ser um *hub* de dados e não uma biblioteca de dados científicos. Apesar disso, instituições de pesquisa governamentais usam instâncias CKAN para armazenamento de seus dados. Exemplos são o Reino Unido[18], Estados Unidos[19], a União Europeia[20], o Brasil[21] entre outros. Há exemplos também de instâncias da comunidade, como o DataHub[22]. O Portal Brasileiro de Dados Abertos, o portal recomendado pela INDA para disseminação de dados e informações públicas do Brasil, é uma instância CKAN. Na Figura 2.5 é mostrada a página de datasets do portal.

Figura 2.5 - Página de datasets do Portal Brasileiro de Dados Abertos



Fonte: Portal Brasileiro de Dados Abertos (2019).

[18] data.gov.uk
[19] data.gov
[20] europeandataportal.eu
[21] dados.gov.br
[22] datahub.io

Na Figura 2.5 observa-se uma organização semelhante às dos *frameworks* citadas anteriormente, isso se deve ao fato do CKAN possuir recursos como metadados configuráveis, armazenamento de dados com visualização, histórico de versionamento, grupos de conjunto de dados, busca textual, uma interface web entre outros (WAINWRIGHT, 2012). Um artigo pode ser catalogado como um conjunto de dados com recursos, tais como diferentes versões do impresso; links para a página do site do periódico; planilhas de resultados experimentais; o código-fonte para processar os resultados; entre outras, criando assim um ambiente de conhecimento, característica chave da Ciência Aberta.

2.4.4 Análise dos frameworks

Com objetivo de selecionar o *framework* que dará suporte à plataforma proposta, uma comparação entre as mesmas foi realizada. Após revisões práticas e teóricas dos *framework* pode-se concluir o objetivo de cada uma. O Dataverse, com o objetivo em promover Dados de Pesquisa Abertos, o Invenio, com o objetivo de promover Ciência Aberta e o CKAN, com o objetivo de promover Dados Abertos. Logo, para destacar a melhor ferramenta para ser usada para gestão e publicação dos dados, uma comparação de características foi feita, usando características recomendadas pela literatura de análise de *frameworks* (AMORIM et al., 2017) e funcionalidades interessantes à observação da Terra. Essas funcionalidades foram selecionadas pois vão ao encontro das práticas de Ciência Aberta e de boa navegabilidade. O resultado da comparação das plataformas é mostrado na Tabela 2.1.

Tabela 2.1 - Comparação entre funcionalidades de *framework*

Características	CKAN	Invenio	Dataverse
Código aberto	X	X	X
Customização	X	X	
Internacionalmente amigável	X		
Versionamento de conteúdo	X		X
Geração de DOI		X	X
Esquema de dados	Flexível	Flexível	Fixo
Visualização de conteúdo publicado	X	X	X
Suporte a dados espaciais	X		X
Busca espacial	X		

Fonte: Produção do autor

Na Tabela 2.1 as características apontadas foram: (a) código aberto: quando o código

fonte é disponibilizado pelo detentor de a forma permitir estudo, edição e distribuição; (b) customização: permitir a customização da aparência do portal; (c) internacionalmente amigável: tradução para múltiplos idiomas de forma automatizada dos componentes de interface; (d) versionamento de conteúdo: permitir o acompanhamento de todas as alterações feitas a um conteúdo publicado; (e) geração de DOI: gerar de forma automatizada um identificador DOI para cada depósito; (f) esquema de dados: permite adicionar diferentes tipos de dados, sem a necessidade de redefinir a estrutura de dados principal; (g) visualização do conteúdo publicado: permitir que o usuário do portal veja os dados sem a necessidade de baixá-lo; (h) suporte a dados espaciais: possui características geoespaciais para armazenamento, visualização e consulta de dados nos principais formatos; e (i) busca espacial: permitir que usuários encontrem dados através da sua localização. A Tabela 2.2 mostra as características principais de cada plataforma.

Tabela 2.2 - Características chave de bibliotecas digitais

Plataforma	**Características chave**
CKAN	- Personalização avançada, com um sistema de extensões. - Suporte a dados espaciais, com pesquisa e consultas. - Pronto para uso, não requerendo conhecimento avançado.
Invenio	- Sistema de geração e exportação de citações. - Personalização do esquemas de dados. - Sistema com pré-reserva de identificadores DOI, pronto para uso.
Dataverse	- Sistema de geração e exportação de citações. - Pronto para uso, não requerendo conhecimento avançado. - Sistema com pré-reserva de identificadores DOI, pronto para uso.

Fonte: Produção do autor

Com isso é possível concluir que, para o propósito do projeto, os dois *frameworks* que se destacam são CKAN e Invenio. Os fatores decisivos para a escolha foram a característica customizável do *framework*, a complexidade para instalação e acesso à documentação. Após a análise foi possível concluir que CKAN possui a melhor customização na questão de adaptabilidade e operabilidade, características importantes para o objetivo da plataforma proposta. Sendo assim, CKAN será usada em conjunto com a plataforma proposta para dar suporte à disseminação dos dados. De acordo com a literatura de análise de *framework* para criação de bibliotecas digitais

(AMORIM et al., 2017), o CKAN possui vantagens em relação às demais plataformas.

2.5 Dados abertos da OBT/INPE

O INPE é o principal instituto federal de pesquisa nas áreas de Observação da Terra e Ciências Espaciais. Desde 1988, o INPE lidera projetos que produzem informações oficiais sobre uso da Terra e cobertura, utilizados pelo governo brasileiro para planejar políticas públicas na área ambiental. Esses conjuntos de dados são abertos e alguns deles produzidos pelos projetos PRODES[23], DETER[24] e TerraClass[25] (INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS, 2019; DINIZ et al., 2015).

Desde 2004, o projeto de desmatamento em tempo quase-real (DETER) mapeia os alertas de desmatamento para a Amazônia brasileira usando imagens de baixa resolução (MODIS 250 metros e CBERS4-WFI 64 metros). Embora apenas áreas desmatadas com mais de 25 hectares possam ser detectadas, a alta resolução temporal dessas imagens permite detectar áreas no mesmo dia da eliminação da floresta, em condições atmosféricas livres de nuvens.

O sistema de monitoramento da floresta amazônica brasileira por satélite (PRODES) é um projeto internacionalmente aclamado que monitora o desmatamento por corte raso na Amazônia brasileira desde 1988 e no Cerrado desde 2016. Diferente do DETER, que identifica áreas de alertas de desmatamento, o PRODES fornece taxas oficiais precisas de desmatamento a cada ano. O PRODES utiliza imagens de média resolução de diferentes satélites, entre eles, Landsat e CBERS. Reduzindo assim o intervalo de tempo entre as observações, atenuando a perda de informações devido à cobertura de nuvens. Desde 2016 o PRODES fornece taxas oficiais de desmatamento para o bioma Cerrado brasileiro. Semelhante à Amazônia, o governo brasileiro usa essas informações para estabelecer políticas públicas, estratégias de vigilância e ações de planejamento.

O projeto TerraClass mapeia a cobertura do solo das áreas anteriormente desmatadas por corte raso, detectadas pelo PRODES. Possibilita assim a compreensão dos processos de mudanças de uso e cobertura da Terra na Amazônia brasileira e no bioma Cerrado. Com base em técnicas de análise de dados de sensoriamento remoto e geoinformação, especialistas classificam as áreas desmatadas em diferentes classes de uso e cobertura da Terra.

[23] www.obt.inpe.br/OBT/assuntos/programas/amazonia/prodes
[24] www.obt.inpe.br/OBT/assuntos/programas/amazonia/deter
[25] www.dpi.inpe.br/tccerrado

Os conjuntos de dados produzidos por esses projetos são organizados em infraestruturas de dados espaciais com os serviços web padrão de dados geográficos WMS, WFS e WCS. O TerraBrasilis[26] é uma plataforma web desenvolvida pelo INPE para organização, acesso e uso dos dados geográficos de monitoramento ambiental. O objetivo da plataforma é disseminar informações de uso e cobertura da Terra produzidas pelos projetos PRODES e DETER (ASSIS et al., 2019). A Figura 2.6 mostra o mapa de desmatamento apresentado pela plataforma. Nela é possível observar, além dos dados de desmatamento, o gerenciamento de camadas, como camadas com a região da Amazônia Legal, o desmatamento anual de 2013 a 2018, áreas de floresta entre outras.

Figura 2.6 - Página inicial do portal TerraBrasilis

Fonte: Assis et al. (2019).

2.6 Considerações finais

A adoção de práticas de Dados Abertos e Ciência Aberta, indica que grupos de pesquisa vão aumentar a visibilidade dos produtos de dados derivados de suas pesquisa. Para que isso seja possível se torna necessário o uso de uma biblioteca digital, após revisões conclui-se que a *framework* CKAN se encaixa perfeitamente com o objetivo proposto. Porém com a adoção da *framework* não resolve problemas como

[26] www.terrabrasilis.dpi.inpe.br

dificuldade para disseminação de dados através dos protocolos OGC. Para isso é necessário que a plataforma ajude pesquisadores oferecendo facilidades para disseminação, para adoção de políticas/normas estabelecidas no âmbito de toda a esfera governamental.

3 PROJETO DA PLATAFORMA

Esse capítulo apresenta o projeto de uma plataforma para compartilhamento de dados científicos geoespaciais que chamamos de TerraBrasilis Research Data, a principal contribuição desta pesquisa. Com o objetivo de criar uma plataforma para gerenciamento de dados e metadados científicos geoespaciais automatizada foram usados princípios de Dados Abertos e Ciência Aberta. Para validação da plataforma proposta, usamos dados científicos produzidos por pesquisadores da OBT-INPE.

O objetivo é que a plataforma proposta ofereça facilidades para a comunicação entre os dados e metadados publicados na plataforma e infraestruturas estabelecidas, sistemas de informações geográficas (SIG) e outros meios para acesso a dados.

Para que a solução ajude pesquisadores a compartilharem os seus dados e metadados é necessário que a mesma forneça comunicação entre os dados da plataforma e infraestruturas estabelecidas através da adoção de políticas e normas para publicação de dados espaciais. Para que os objetivos sejam alcançados, uma série de requisitos foi levantada.

3.1 Requisitos

A partir de reuniões e consultas a diferentes grupos de pesquisadores da OBT-INPE, foram estabelecidos um conjunto de requisitos. Através dessas consultas foi possível compreender informações como: (a) o domínio da aplicação; (b) as partes interessadas; (c) os problemas a serem resolvidos; e (d) os componentes que vão compor a plataforma.

Baseado nessas informações, os requisitos (R1 a R12) foram estabelecido. Os requisitos do TerraBrasilis Research Data são:

- **R1:** Gerenciamento de usuários;
  - **Pesquisadores** da OBT-INPE, atuam como produtores de dados;
  - **Usuários externos**, alunos, professores e pesquisadores de outras instituições, atuam como consumidores de dados;
  - **Administradores**, atuam como gerenciadores da infraestrutura da plataforma, monitoram os recursos computacionais;

- **R2:** Prover uma interface capaz de criar e gerenciar os repositórios de dados;
  - Esses repositórios de dados compreendem um ambiente isolado com soluções para armazenamento de dados, serviços para disseminação, catalogação e processamento;
- **R3:** Cada repositório de dados deve prover serviços para as seguintes funcionalidades:
  - Armazenamento de dados vetoriais, raster e tabulares;
  - Visualização de dados;
  - Suporte a dados em formatos convencionais (GeoTIFF; Shapefile; CSV; GeoJSON; GeoPackage);
  - Processamento de dados;
- **R4:** Prover um ponto centralizado para descoberta e acesso de dados;
  - Isso compreende permitir que usuários busquem os dados disponíveis na plataforma pelos seus metadados, de forma intuitiva;
- **R5:** Prover facilidades para usuários publicarem dados e metadados na plataforma;
- **R6:** Prover formas para criar e gerenciar grupos de pesquisa e laboratórios;
- **R7:** Prover formas de pesquisar e fazer download dos dados e metadados da plataforma;
  - Isso compreende uso de ferramentas para indexação dos metadados, para os mesmos serem encontrados pelo buscador da plataforma e buscadores externos;

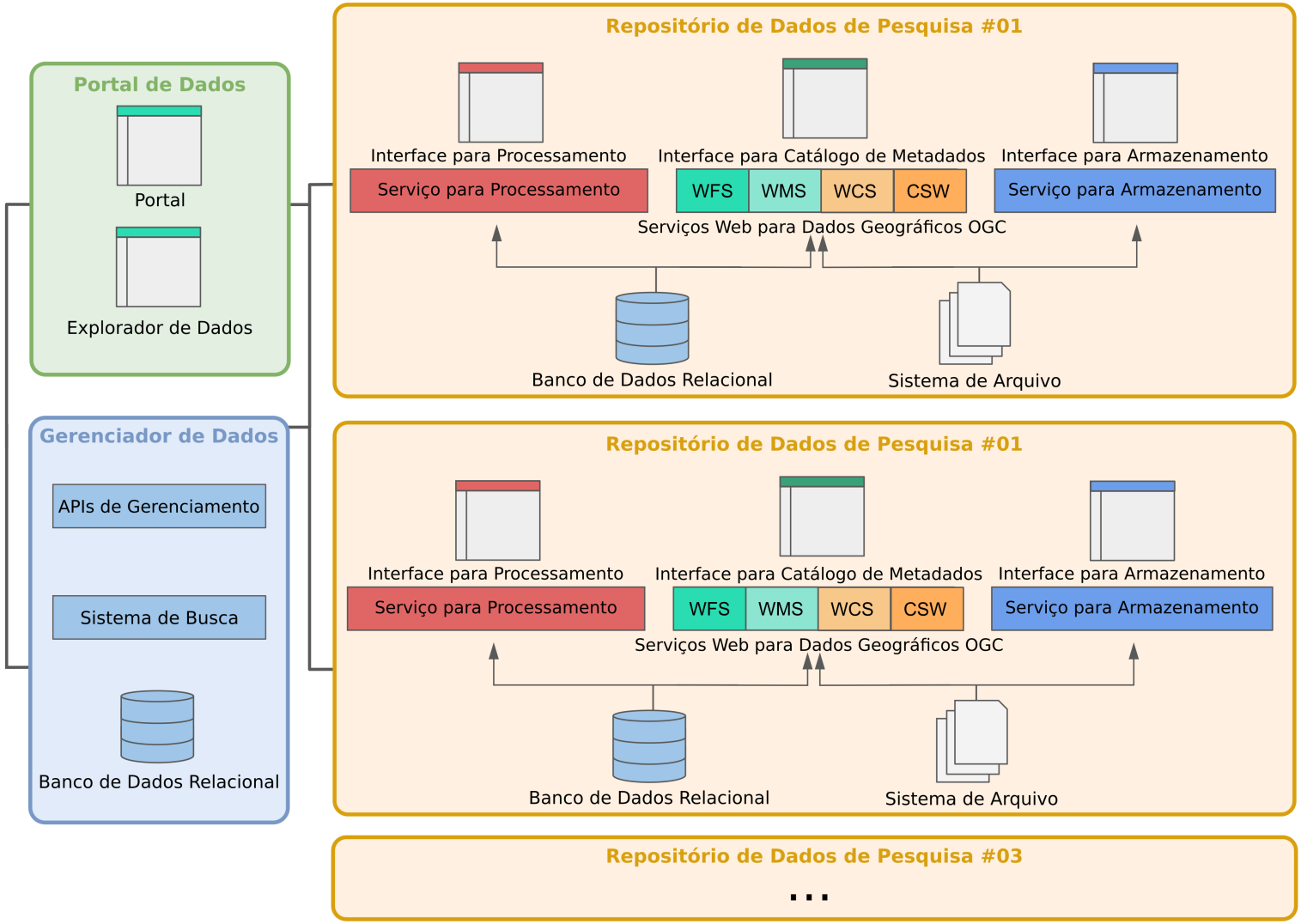
- **R8:** Permitir que os usuários exportem citações dos dados na plataforma;
  - Isso compreende exportar citações seguindo padrões estabelecidos para citações;
- **R9:** Fornecer maneiras de se associar dados e metadados da plataforma a recursos externos;
  - Isso compreende armazenamento e serviços de disseminação abertos, que permitam que sistemas de informação geográficos, serviços de processamento e outras formas de se acessar os dados através de padrões;
- **R10:** Fornecer uma interface capaz de monitorar a plataforma, através de estatísticas;
  - Isso compreende o monitoramento de acessos o número de citações exportadas, o número de download dos dados e outras estatísticas;
- **R11:** Fornecer recursos para visualização dos dados e dos metadados;
- **R12:** Fornecer facilidades para integração dos dados e metadados a infraestruturas de dados estabelecidas, como INDE e INDA;

Com os requisitos estabelecidos, foi possível estabelecer: (a) o domínio da aplicação, um portal para dados de pesquisa; (b) os problemas a serem resolvidos, dificuldades dos usuários para disseminarem seus dados e metadados seguindo padrões estabelecidos; e (c) os componentes que vão compor a plataforma, um portal de dados, um gerenciador e repositórios de dados, com os serviços.

3.2 Arquitetura da plataforma

Com os requisitos estabelecidos, foi possível esboçar uma arquitetura para representar os componentes que compõem a plataforma, como mostrado na Figura 3.1.

Figura 3.1 - Arquitetura da plataforma TerraBrasilis Research Data



Fonte: Produção do autor

3.2.1 Componentes

Nessa seção são introduzidos os componentes da plataforma. Como citado nos requisitos, para prover um ambiente em que pesquisadores possam gerenciar dados científicos geoespaciais se faz necessário uma abordagem com múltiplos componentes. A arquitetura da plataforma TerraBrasilis Research Data, mostrada na Figura 3.1, é composta por três componentes: os Repositórios de Dados de Pesquisa, o Gerenciador de Dados e o Portal de Dados.

Repositório de Dados de Pesquisa

Um Repositório de Dados de Pesquisa fornece a infraestrutura computacional subjacente para armazenamento, gerenciamento, catalogação, compartilhamento e visualização dos dados e metadados científicos. Este componente adota protocolos e interfaces baseados em padrões abertos, em especial os do consórcio OGC, para facilitar o compartilhamento e interoperabilidade dos metadados. Assim, os pesquisadores têm a garantia de que os produtos de dados disponibilizados nesta plataforma possam ser diretamente utilizados em Sistemas de Informação Geográficos (SIG) e demais ferramentas computacionais, além de se adequarem às normas nacionais de metadados, como a INDE e INDA. Para criação de ambientes isolados que forneçam instâncias da infraestrutura computacional citada anteriormente, foi adotada uma abordagem usando contêineres, tecnologia para abstração e virtualização de ambientes. Essa abordagem, em conjunto com uma tecnologia para orquestração de contêineres, permitirá a entrega de um Repositório de Dados de Pesquisa pronto para uso.

Gerenciador de Dados

O Gerenciador de Dados é responsável pela gestão de todas as entidades da plataforma, usuários, grupos e repositórios de dados pesquisa; pela gestão das publicações; e pelas as funções de infraestrutura, como a criação dos repositórios. As funcionalidades do Gerenciador de Dados são: (a) controle de usuários da plataforma; (b) gerência dos grupos de pesquisa (c) publicação dos conjuntos de dados e seus metadados; (d) busca textual pelos metadados; além de outras funções de gerenciamento. O componente também é responsável por permitir a comunicação entre os Repositórios e o Portal de Dados.

Portal de Dados

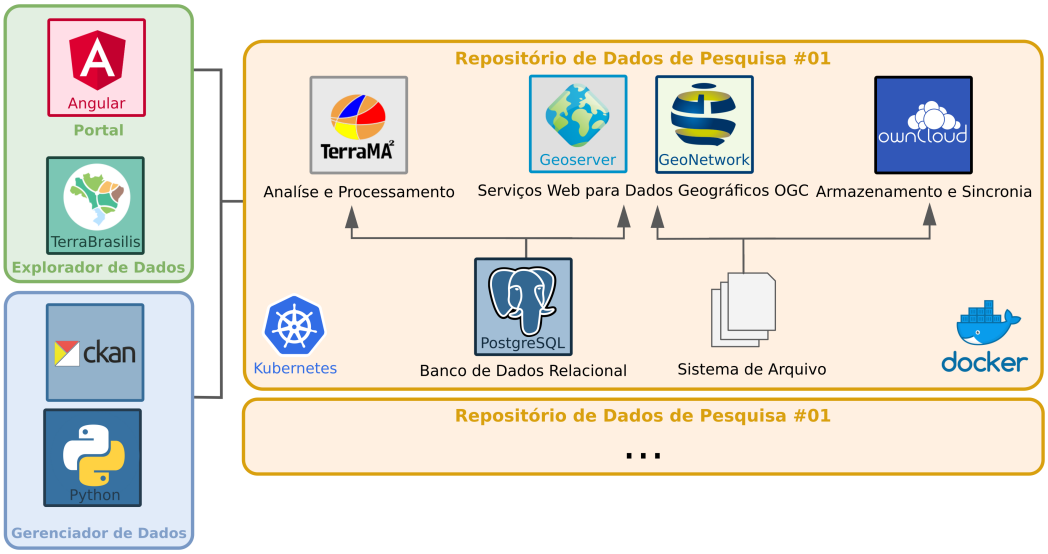
O Portal de Dados é uma interface web que permite a qualquer usuário pesquisar,

visualizar, citar e fazer download de todos os conjuntos de dados geoespaciais publicados. Para usuários autenticados, permite a criação e gestão de publicações, grupos e repositórios de dados. Inspirado em bibliotecas digitais, o componente possibilita a navegação pelos metadados, por todos os conjuntos de dados publicados, visualização dos grupos de pesquisa, geração de citação, download, entre outras funções. Outra interface desse mesmo componente é o Explorador de Dados, responsável por prover visualização dos dados georreferenciados da plataforma, com navegação baseada em mapas, gerenciamento de camadas, e outras funções de Web GIS.

3.3 Tecnologias

Essa seção descreve as tecnologias usadas para implementar o projeto descrito na seção anterior. Para os componentes Portal de Dados e Gerenciador de Dados não existe uma única tecnologia que resolva o problema proposto, logo foi necessária uma abordagem técnica específica. Para o caso do Gerenciador de Dados foi usada uma solução híbrida com a framework CKAN, expandindo-a para suportar os conceitos da plataforma, e uma API com isso será possível usar todas as vantagens da framework sem comprometer o objetivo e os requisitos. Para o sistema de orquestração dos repositórios e os todos componentes que o compõem serão usados apenas *softwares* de Código Aberto. Com os *softwares* estabelecidos, foi possível esboçar o projeto da plataforma, uma evolução da arquitetura, apresentando as tecnologias e suas relações com os componentes, mostrada na Figura 3.2.

Figura 3.2 - Tecnologias da plataforma TerraBrasilis Research Data



Fonte: Produção do autor

Na Figura 3.2 cada uma das tecnologias foi associada a um ou mais componentes da arquitetura. É importante pontuar que o Kubernetes e o Docker não estão ligados a nenhum componente, porém ambos são responsáveis pelo gerenciamento dos repositórios e pela entrega dos componentes que a compõem. Como cada repositório se materializa em um nó do Kubernetes, cada tecnologia será entregue em um ambiente conteinerizado via Docker.

O fator principal para seleção das tecnologias foi a característica aberta. Como o projeto da plataforma é código aberto, qualquer indivíduo ou instituição poderá replicar tudo o que foi feito.

3.3.1 Kubernetes e Docker

Para garantir que cada repositório entregue para pesquisadores um ambiente isolado pronto para uso é necessário um sistema de orquestração de containers. Para prover ambientes isolados com as ferramentas desejadas será usado Docker[1], um software baseado em contêineres que fornece uma camada adicional de abstração permitindo a virtualização de ambientes. Para que a plataforma possa criar esses ambientes será usado o Kubernetes[2], uma solução código aberto para automatização da implantação, dimensionamento e gerenciamento de aplicações via contêineres.

O Kubernetes possui um sistema de *nós* e *Pods*. *Pods* são equivalentes a máquinas virtuais, podem rodar um ou mais contêineres e compartilham armazenamento e rede internamente. Os *nós* são estruturas virtuais para compartilhamento de recursos entre *Pods*, o sistema operacional, hardware e a interface de rede. Um ecossistema real pode conter inúmeros *nós* e múltiplos mestres, estrutura gerenciadora de trabalho e recursos. Ambos as estruturas são criadas a partir de modelos, arquivos YAML que descrevem todas as características do componente.

3.3.2 TerraMA2

Como objetivo de prover a pesquisadores uma ferramenta dentro dos repositórios para a análise e o processamento de dados geoespacias a plataforma TerraMA2[3] será usada. Desenvolvida pelo INPE, a plataforma computacional, com uso da biblioteca geográfica TerraLib, atende a uma demanda de aplicações de monitoramento e análise. A plataforma tem capacidade de interagir com serviços e modelos geográficos, seguindo os padrões OGC contidos na plataforma e ao usar a solução, pesquisado-

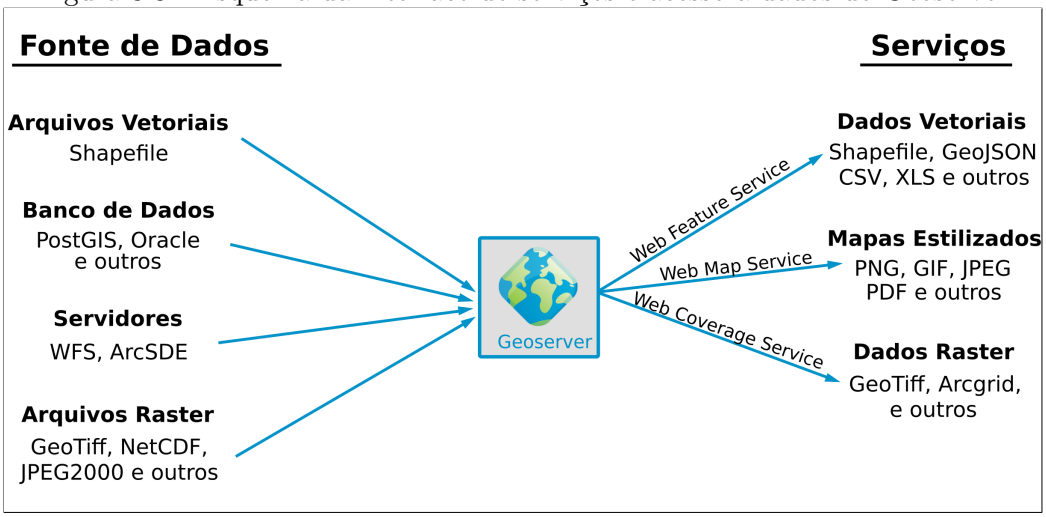
[1] www.docker.com
[2] kubernetes.io
[3] www.terrama2.dpi.inpe.br

res poderão visualizar e processar seus dados dentro das suas próprias repositórios, através do serviço de análise do TerraMA2 e de sua interface.

3.3.3 Geoserver

Os padrões para representação de dados espaciais e armazenamento propostos pela OGC, como WMS, WFS, WCS e CSW demandam serviços web específicos para sua geração. Para criação desses serviços web foi usado o GeoServer[4], um *software* para compartilhamento, processamento e edição de dados geoespaciais. Esse *software* foi desenvolvido a partir da demanda por publicação de dados. Ele permite o uso de diversas fontes de armazenamento de dados e, é uma referência no que diz respeito à implementação dos padrões de serviço OGC. Como mostrado na Figura 3.3, o GeoServer é capaz de oferecer uma interface de serviços para os vários tipos de fonte de dados.

Figura 3.3 - Esquema da interface de serviços e acesso a dados do Geoserver



Fonte: Adaptado de eAtlas (2019).

Os dados exportados pelo GeoServer seguem os padrões de serviço web OGC, padrões esses que compõem a INDE. Com isso, através da plataforma, pesquisadores podem exportar seus metadados, via serviço, seguindo os padrões de metadados requeridos pela INDE.

[4] geoserver.org

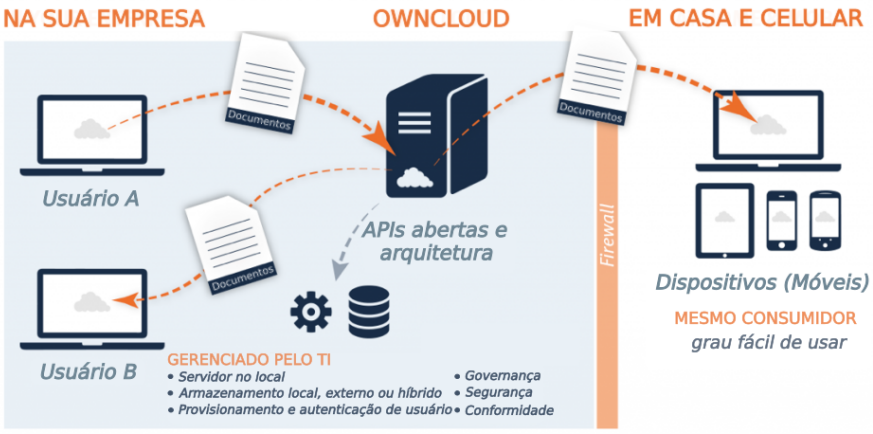
3.3.4 GeoNetwork

Com o objetivo de prover uma interface para visualização e edição de metadado dos catálogos dos repositórios será usado o GeoNetwork[5], uma aplicação para gerenciamento de catálogos de metadados de recursos georreferenciados. A aplicação provê ferramentas para edição de metadados, sistema de busca e visualização através de um mapa interativo. Ao usar o GeoNetwork pesquisadores receberão uma interface com diversos recursos para navegação e edição nos seus respectivos catálogos.

3.3.5 OwnCloud

Além do banco de dados, outra forma de armazenamento dos repositórios é o sistema de arquivos. Para prover usabilidade é necessário um sistema para gerenciamento e sincronização de dados, para que pesquisadores possam armazenar dados fora de um banco de dados relacional com facilidade. Para essa função será usado o sistema OwnCloud[6], um serviço código aberto para armazenamento e sincronia de dados. O serviço provê funcionalidades interessantes para pesquisadores como armazenamento e acesso a dados de qualquer dispositivo, como mostrado na Figura 3.4.

Figura 3.4 - Esquema de funcionamento do OwnCloud



Fonte: Adaptado de OwnCloud (2019).

Como a aplicação OwnCloud compõe um ambiente Kubernetes, o espaço de armazenamento de arquivos dos repositórios de dados de pesquisa é o mesmo acessado pelo OwnCloud, que por sua vez é o mesmo do restante das aplicações. Isso só é

[5] geonetwork-opensource.org
[6] owncloud.org

possível graças ao compartilhamento de arquivos entre *Pods* do Kubernetes.

3.4 Considerações finais

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 é capaz de contemplar o armazenamento, catalogação, gerenciamento, processamento e disseminação de dados científicos.

Dessa forma, a arquitetura da plataforma proposta, evolui do conceito estalecido de bibliotecas digitais e permite que pesquisadores possam além de armazenar, catalogar, gerenciar, processar e disseminar seus dados. Com o projeto da plataforma estabelecido e validado, o próximo passo é a implementação.

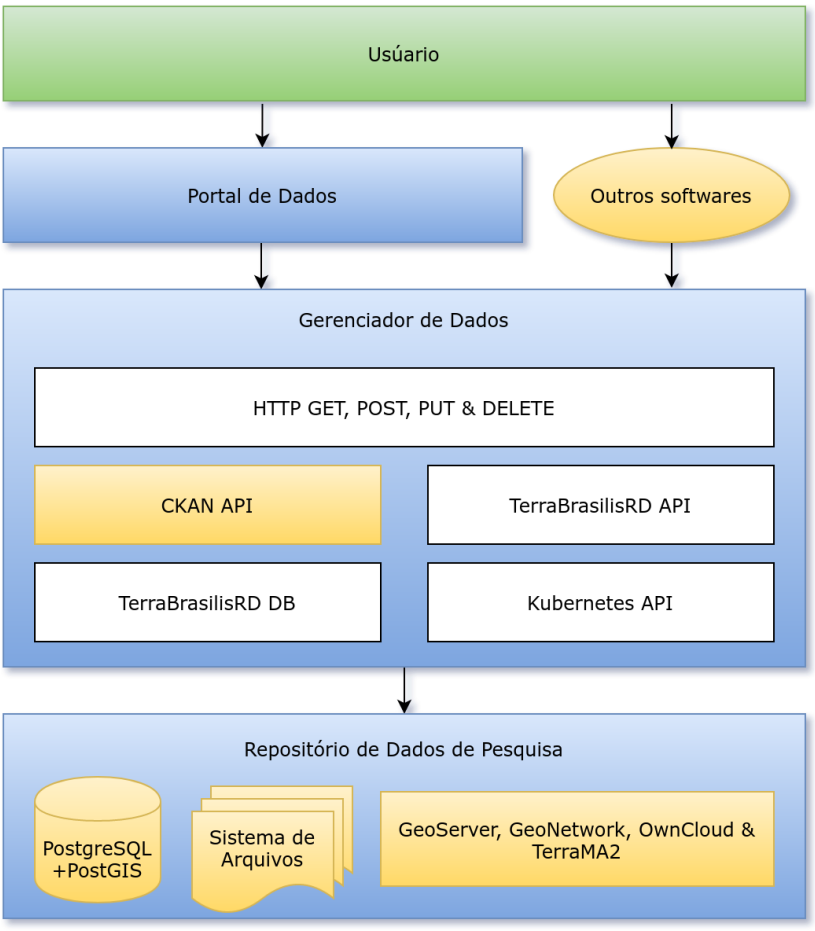
4 TERRABRASILIS RESEARCH DATA

Esse capítulo descreve como a plataforma TerraBrasilis Research Data foi implementada seguindo os princípios FAIR e a arquitetura proposta no Capítulo 3.

4.1 Visão geral

A estrutura da plataforma proposta é moldada pelos seus requisitos e materializada em um projeto em componentes. A visão geral de como os componentes da plataforma se conectam é ilustrada na Figura 4.1. Em azul observa-se os componentes desenvolvidos para a plataforma e em amarelo as tecnologias existentes utilizadas.

Figura 4.1 - Visão geral TerraBrasilis Research Data



Fonte: Produção do autor

4.2 Portal de dados

O Portal de Dados é o componente que permite a navegação pelos metadados, por todos os conjuntos de dados publicados, visualização dos grupos de pesquisa, geração de citação, download, entre outras funções. Esse componente se materializou como uma interface web após uma análise de alternativas para entrega de dados científicos. Dessa forma foi possível garantir que o Portal de Dados entregue todas as informações essenciais aos pesquisadores e usuários.

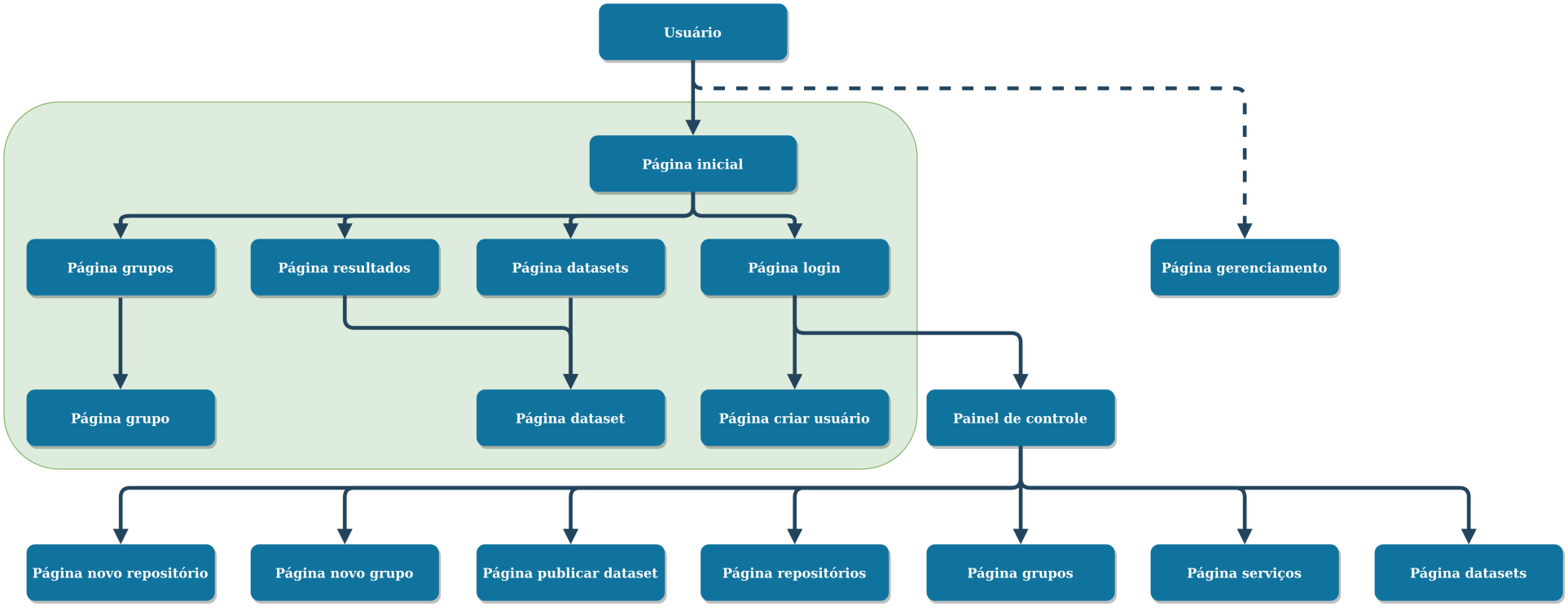
Para que todas informações referentes aos conjuntos de dados publicados na plataforma sejam entregues, foram usados como referências portais estabelecidos para disseminação de dados científicos, em especial Zenodo, PANGAEA e Dataverse.

4.2.1 Estrutura

Com os portais de referência estabelecidos, foram projetadas páginas de forma a permitir que todas as funcionalidades da plataforma sejam acessíveis através do portal. A primeira decisão estrutural foi a da divisão das páginas em duas áreas, são elas: (a) área aberta, composta pelas páginas para acesso a informações publicadas, grupos e datasets, e (b) área de conteúdo com acesso restrito, composta pelas páginas para criação, gerenciamento de repositórios, grupos e serviços, e publicação dos conjuntos de dados. Foi projetada uma terceira parte composta por páginas para administração de recursos computacionais da plataforma. Para a criação do portal foi adotada uma estrutura modular, de forma a permitir a evolução de componentes de interface, páginas e serviços utilizados. Essa estrutura se reflete na adoção da *framework* Angular, para o desenvolvimento do portal.

A arquitetura do Portal de Dados pode ser observada na Figura 4.2. Em verde, são observadas as páginas que fazem parte da área aberta da plataforma. O restante das páginas fazem parte da área fechada, logo requerem autenticação para acesso.

Figura 4.2 - Arquitetura do Portal de Dados



Fonte: Produção do autor

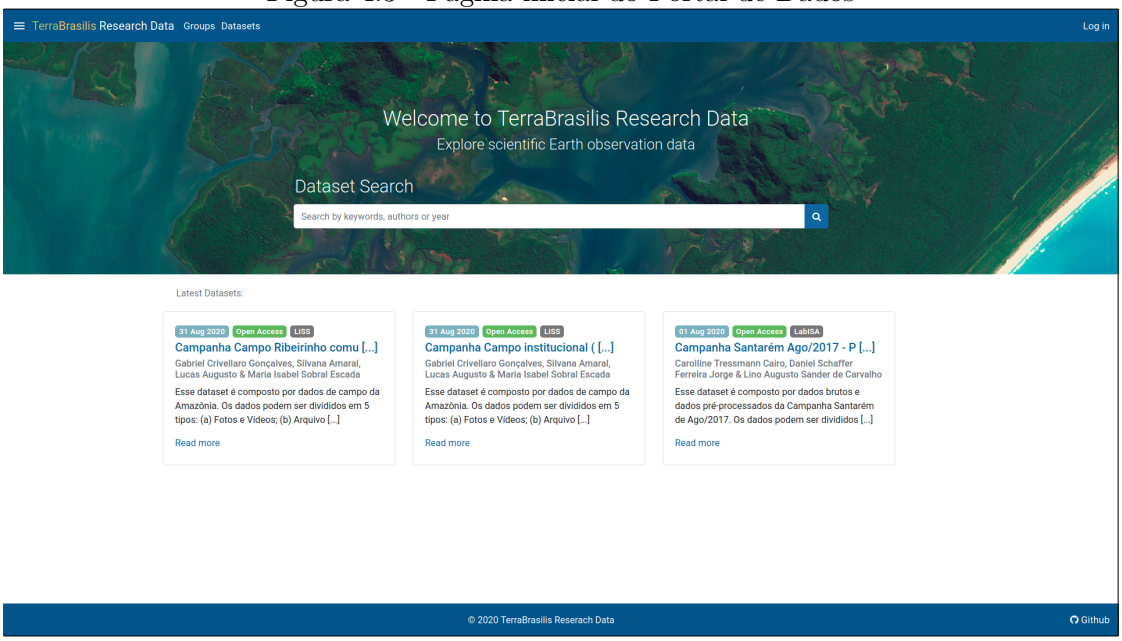
4.2.2 Páginas

Esta subseção apresenta o protótipo do Portal de Dados implementado seguindo a especificação e arquitetura apresentadas ao longo deste trabalho.

Página inicial

A Figura 4.3 apresenta a página inicial da plataforma, por ela é possível buscar pelos conjuntos de dados publicados e visualizar os conjuntos destaque, os últimos três publicados. O buscador dessa página se conecta com o motor de busca do CKAN, o Apache Solr, permitindo uma busca através do título, o nome dos autores, as palavras-chave ou conteúdo textual dos dados.

Figura 4.3 - Página inicial do Portal de Dados



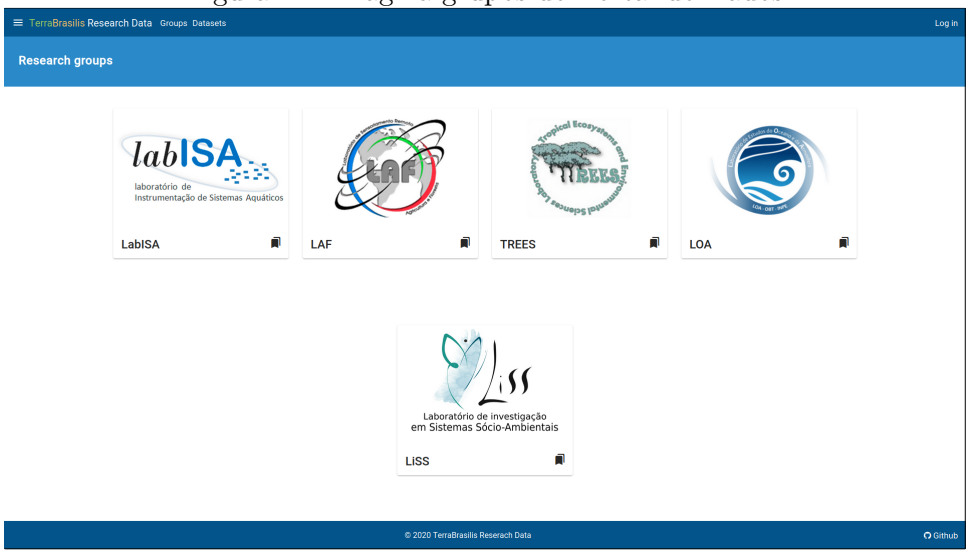
Fonte: Produção do autor

No canto superior esquerdo da página observa-se o menu institucional do portal, por ele é possível chegar às páginas grupos, datasets e login. Nessa página, no lado esquerdo se localiza a barra lateral, um menu de acesso rápido as páginas da área fechada da plataforma. Esse menu de acesso pode ser aberto ou fechado clicando no ícone ao lado do título do site.

Página grupos

A Figura 4.4 apresenta a página "grupos da plataforma". Nesta página observa-se em forma de cartões (capa, título e *abstract*), cada um dos grupos de pesquisa da plataforma. Essa página permite que usuários da plataforma possam visualizar, em uma tela, os responsáveis pelos conjuntos de dados científicos, sejam laboratórios, projetos, ou outro conjunto de autores de dados científicos. Para essa página foi usado como referência o portal Science Learn[1].

Figura 4.4 - Página grupos do Portal de Dados



Fonte: Produção do autor

O *abstract* de cada um dos grupos se torna visível ao passar o cursor sobre o respectivo cartão do grupo, como mostrado Figura 4.5.

Figura 4.5 - Cartão grupo LabISA

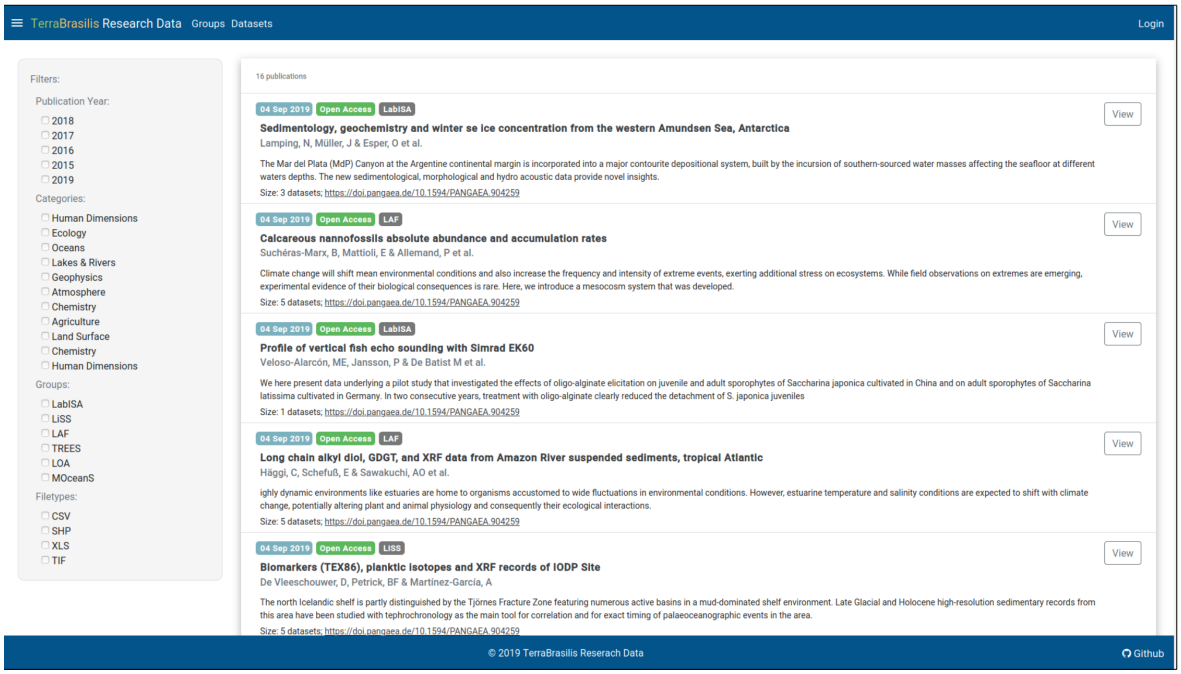
Fonte: Produção do autor

[1] www.sciencelearn.org.nz

Página datasets

A Figura 4.6 apresenta a página datasets da plataforma. Por meio desta página é possível visualizar todos os conjuntos de dados publicados. Isso inclui, para cada conjunto de dados, o título, autores, *abstract*, tamanho do dataset, endereço digital, data de publicação, grupo de pesquisa e tipo de visibilidade. Para essa página foi usado como referência o portal Zenodo.

Figura 4.6 - Página datasets do Portal de Dados



Fonte: Produção do autor

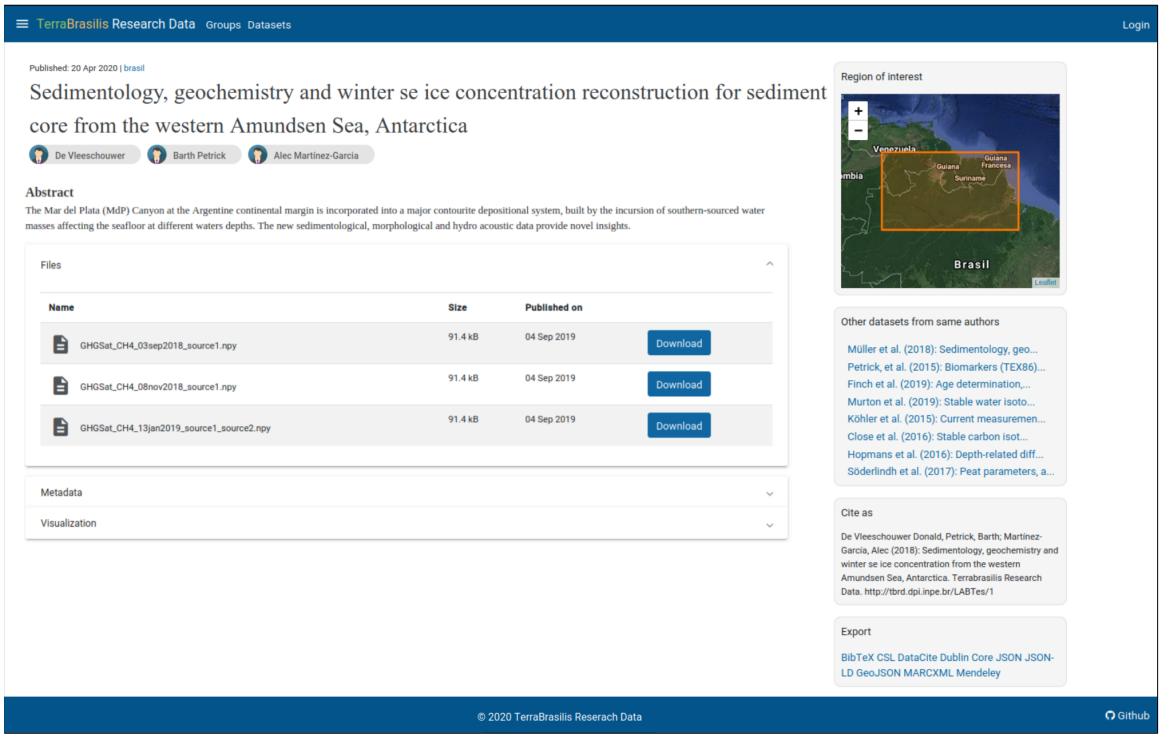
No canto esquerdo da página observa-se a janela filtros. Nela é possível discriminar os conjuntos de dados mostrados no centro da página. Os filtros disponíveis na janela são: (a) Ano de Publicação; (b) Palavras-Chave; (c) Grupo; e (d) Tipos de Arquivo. Através desses filtros, pesquisadores e usuários podem restringir os conjuntos de dados a escopos específicos.

Portanto, esta página apresenta o resultado das buscas na plataforma a partir de consultas textuais, seja por título, ano de publicação autores, palavras-chave entre outros. Esta funcionalidade é baseada nos recursos fornecidos pelo CKAN, que por sua vez utiliza o Apache Solr para busca textual.

Página dataset

A Figura 4.7 apresenta a página dataset da plataforma, nesta página é possível observar todas as principais informações referentes a um conjunto de dado selecionado. Isso inclui título, *abstract*, os autores da publicação, a área de cobertura do dado entre outras informações. Para essa página foram usados como referências os portais PANGAEA, Zenodo e Dataverse.

Figura 4.7 - Página dataset do Portal de Dados



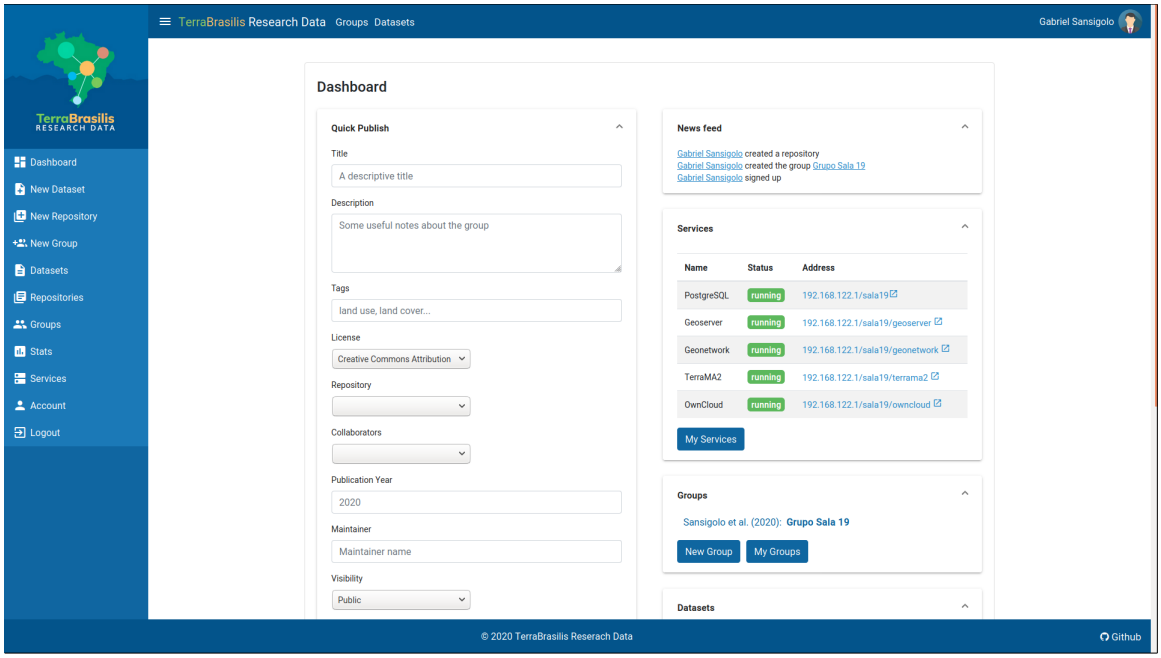
Fonte: Produção do autor

No centro da página observam-se três abas: dados, metadados e visualização. A primeira aba contém informações referentes aos dados que compõem o conjunto de dados selecionado, como: nome do arquivo, formato e o botão de download. A segunda aba é composta por todos os metadados referentes ao conjunto de dados, tanto os obrigatórios da plataforma quanto os criados pelo pesquisador responsável. E, por último, a aba visualização que provê a visualização em um mapa dos dados espaciais do conjunto de dados. No canto direto observa-se os componentes referentes à citação do conjunto de dados com o texto para citação e botões para exportar as citações em diferentes padrões, como BibTeX, DataCite, entre outros.

Painel de controle

A Figura 4.8 apresenta o painel de controle da plataforma. Através dessa página é possível publicar novos conjuntos de dados, ver últimas atualizações, acessar os seus serviços e navegar pelos grupos, repositórios e datasets do usuário logado. Para essa página foram usados como referência a tecnologia de gestão de serviços Portainer.

Figura 4.8 - Página painel de controle do Portal de Dados



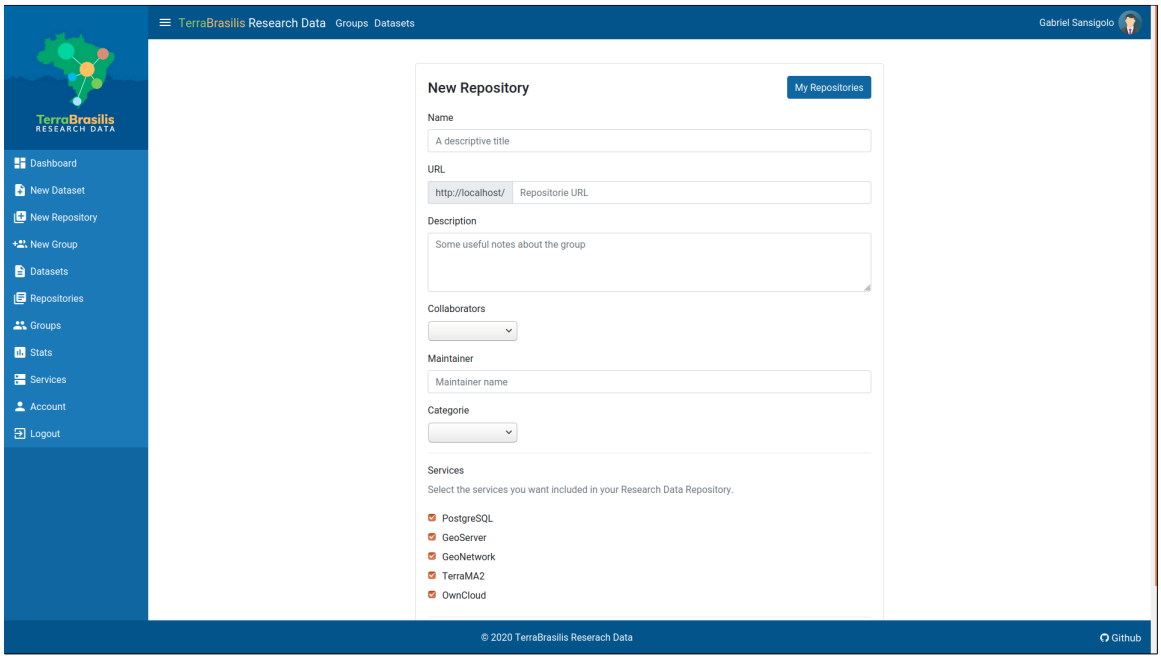
Fonte: Produção do autor

No lado esquerdo da página observa-se o menu lateral da plataforma, por ele é possível acessar todas as páginas da área fechada da plataforma. No centro ao lado esquerdo observa-se a janela de publicação rápida, essa janela compartilha todas as funções da página para publicação de um novo dataset. O centro da página é composto por cinco janelas, são elas: o *feed* de notícias, com as atualizações relacionadas ao usuário; a janela serviços com as informações essenciais relacionadas aos serviços do usuário logado; a janela grupo com o nome e endereço para os grupos no qual o usuário pertence; a janela repositórios com o nome e endereço para os repositórios no qual o usuário compartilha recursos; e a janela datasets com o nome e o endereço dos conjuntos de dados que o usuário é autor. Com exceção do *feed* de notícias, todas as janelas têm um botão para suas respectivas páginas.

Página novo repositório

A Figura 4.9 apresenta a página novo repositório. Nela pesquisadores poderão criar os seus Repositórios de Dados de Pesquisa, ambientes com a infraestrutura computacional subjacente para armazenamento, gerenciamento, catalogação, compartilhamento e visualização dos dados científicos. Nessa página se encontram formulários com os metadados essenciais para criação de um repositório no sistema.

Figura 4.9 - Página nova repositório



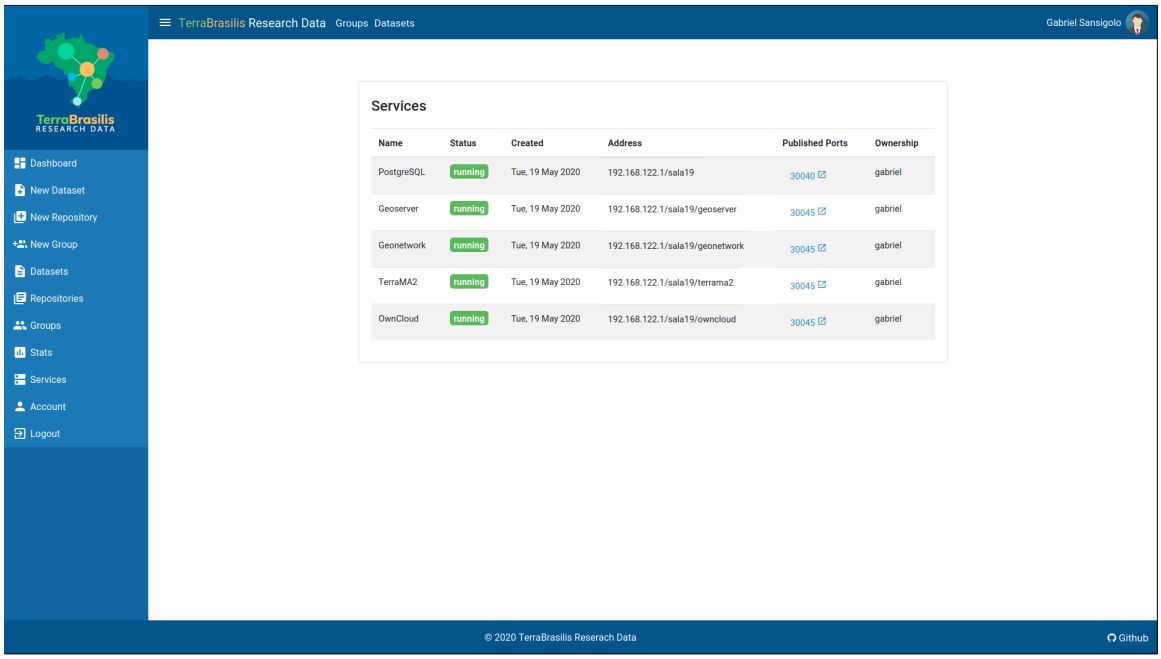
Fonte: Produção do autor

Na parte inferior da página observa-se um conjunto de caixas de seleção, cada um com uma tecnologia, são elas: (a) PostgreSQL; (b) GeoServer; (c) GeoNetwork; (d) TerraMA2; e (e) OwnCloud. Através da seleção pesquisadores poderão criar Repositórios de Dados de Pesquisa com as tecnologias que eles desejam que componham seus ambientes. Essa página se comunica diretamente com o TerraBrasilis RD API, como todas as outras, porém ela também se comunica com a Kubernetes API, para a realização da entrega do ambiente selecionado.

Página serviços

A Figura 4.10 apresenta a página serviços da plataforma, nesta página é possível observar todas as informações relacionadas aos serviços que o usuário logado tem acesso. Isso inclui nome do serviço, seu status, a data de criação, o endereço, a porta e quem está registrado como dono. Para essa página foram usados como referência a tecnologia de gestão de serviços Portainer.

Figura 4.10 - Página serviços



Fonte: Produção do autor

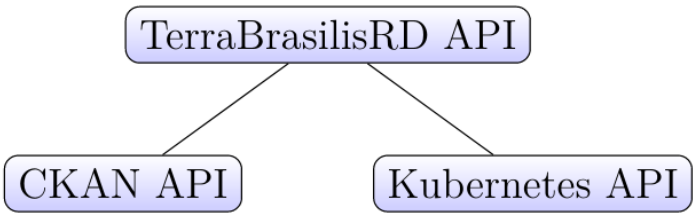
Essa página, assim como a anterior (novo repositório), além de se comunicar com TerraBrasilisRD API, também se comunica com a Kubernetes API. Com isso é possível acessar externamente, identificar endereço e visualizar em tempo real o status de execução (rodando, parado e em criação) de cada um dos serviço. Com as informações contidas nessa página, os pesquisadores podem visualizar o status de todos os serviços associados a ele.

4.3 Gerenciador de dados

O Gerenciador de Dados é responsável pela gestão de todas as entidades da plataforma. A decisão de como se materializaram esse componente se deu dividindo as funcionalidades em três APIs, uma para gestão de grupos, repositórios e usuários, outra sendo a *framework* CKAN, para a gestão dos conjuntos de dados e uma terceira com foco em infraestrutura, exclusivamente para a comunicação com o Kubernetes.

A primeira decisão estrutural foi a divisão em APIs. O Gerenciador de Dados é composto por três APIs: (a) TerraBrasilisRD API, responsável por todas as funções relacionadas à gestão de usuários, grupos, repositórios e serviços; (b) CKAN API, responsável por todos as funções relacionadas aos conjuntos de dados; e (c) Kubernetes API, responsável no contexto de infraestrutura por criar os Repositórios de Dados de Pesquisa. A relação de dependência entre a implementação da TerraBrasilisRD API, CKAN API e Kubernetes API é ilustrada na Figura 4.11.

Figura 4.11 - Relação de dependência entre a implementação do TerraBrasilisRD API, CKAN API e Kubernetes API



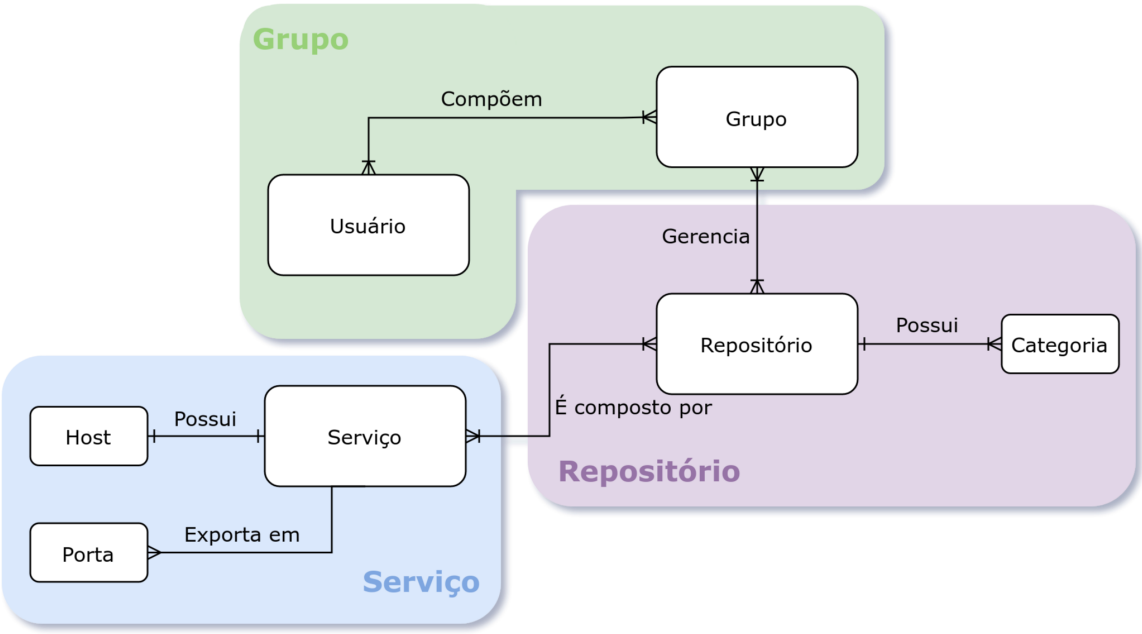
Fonte: Produção do autor

Para a materialização do TerraBrasilisRD API e do Kubernetes API a abordagem adotada foi a da criação de serviços web usando a linguagem Python.

4.3.1 TerraBrasilisRD API

O TerraBrasilisRD API é o principal serviço *web* uma vez que gerencia quase todas as entidades da plataforma. Ele é responsável pelo trabalho de criação, leitura, atualização e remoção de todas entidades relacionadas aos grupos, serviços e repositórios. A Figura 4.12 mostra as relações entre Usuário, Grupo, Repositório, Categoria, Serviço, Host e Porta.

Figura 4.12 - Diagrama de entidades do TerraBrasilisRD API



Fonte: Produção do autor

Todas as funções do TerraBrasilisRD API são materializadas através de rotas. Isso é, para cada uma das entidades e das relações ilustradas na Figura 4.12, existe uma rota HTTP para a realização da função.

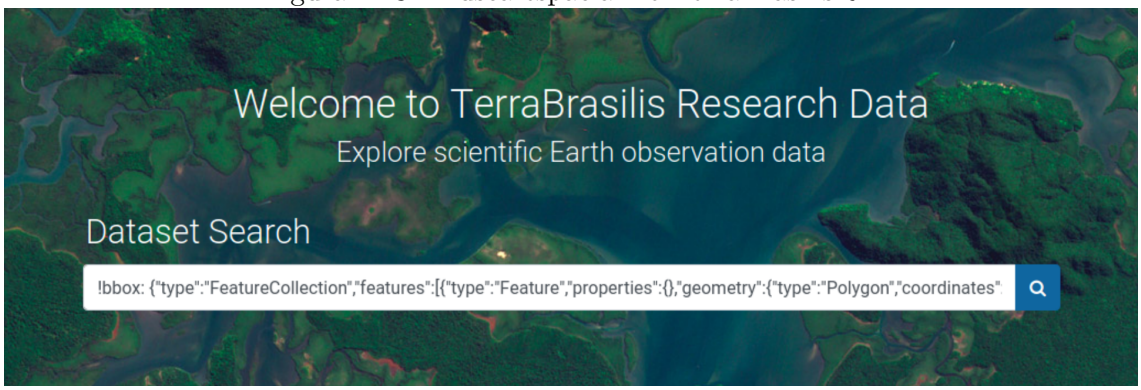
Para o entender melhor o funcionamento do componente, é necessário antes entender todas as entidades ligadas a ele. Como mostrado na visão geral da plataforma, Figura 4.1, o Gerenciador de Dados é composto por cinco entidades, sendo três delas serviços *web*. O outro é a interface de serviço do componente, que recebe chamadas HTTP GET, POST, PUT e DELETE, e direciona essas chamadas aos três serviços. E por último o banco PostgreSQL, que é o local onde todas as informações relacionadas às entidades e as relações são armazenadas (Figura 4.12).

Rota busca espacial

Uma rota importante do TerraBrasilisRD API é a rota de busca espacial. Ela é responsável por realizar a filtragem através da caixa delimitadora mínima, em inglês, *minimum bounding box*, de todos os datasets publicados na plataforma. Com a principal entrada sendo o buscador, localizado na página inicial, essa rota foi desenvolvida de forma que a entrada da área buscada pelo usuário siga um padrão.

Funcionando em paralelo com o CKAN, essa rota foi projetada com objetivo de permitir que pesquisadores, através do buscador textual da plataforma, visualizem todos os conjuntos de dados publicados que se encontrem dentro da caixa delimitadora mínima buscada. Para isso, foi adotado o padrão para representação de recursos geográficos GeoJSON. O uso da busca espacial é feito pelo buscador de datasets da plataforma através comando: `!bbox:` seguido pelo GeoJSON escrito. Um exemplo de uso da busca espacial na plataforma é mostrada na Figura 4.13.

Figura 4.13 - Busca espacial no TerraBrasilisRD



Fonte: Produção do autor

A busca espacial é responsável por listar todos os conjuntos de dados publicados com metadado de localização espacial, executando a operação `ST_Contains`, em cada um dos casos. Essa operação, de forma resumida, consiste em verificar se a geometria A contém a geometria B. Para execução dessa operação de forma eficiente foi usada o PostGIS, extensão espacial do sistema de banco de dados PostgreSQL. Após isso, a rota retorna os conjuntos de dados contidos dentro da área buscada, permitindo assim que pesquisadores, através da plataforma, possam discriminar datasets espacialmente.

Rota upload

Compondo a série de rotas do TerraBrasilisRD API, a rota *upload* é responsável por tratar todos os arquivos vindos do Portal de Dados. As principais entradas são: a página de configurações de usuário, com o envio de foto de perfil, a página de criação de grupos, com o envio de capa do grupo, e a página novo conjunto de dados, com o envio dos dados para publicação.

Conectado diretamente com o CKAN, essa rota foi projetada com o objetivo de prover facilidades para usuários publicarem seus dados. Para isso foi decidido que a plataforma fornecerá suporte para alguns tipos dados. Esse suporte se resume em facilidades relacionadas ao banco de dados PostgreSQL e ao *software* para compartilhamento GeoServer. Isso é feito através de uma filtragem baseada nos tipos de arquivo suportados, como mostrado na Tabela 4.1.

Tabela 4.1 - Relação entre tipos de dado e suporte

Tipo do Dado/Suporte	**PostgreSQL**	**GeoServer**
Shapefile (SHP)	X	X
Comma-separated values (CSV)	X	
Excel Spreadsheet (XLS)	X	
Tagged Image File Format (TIFF)		X

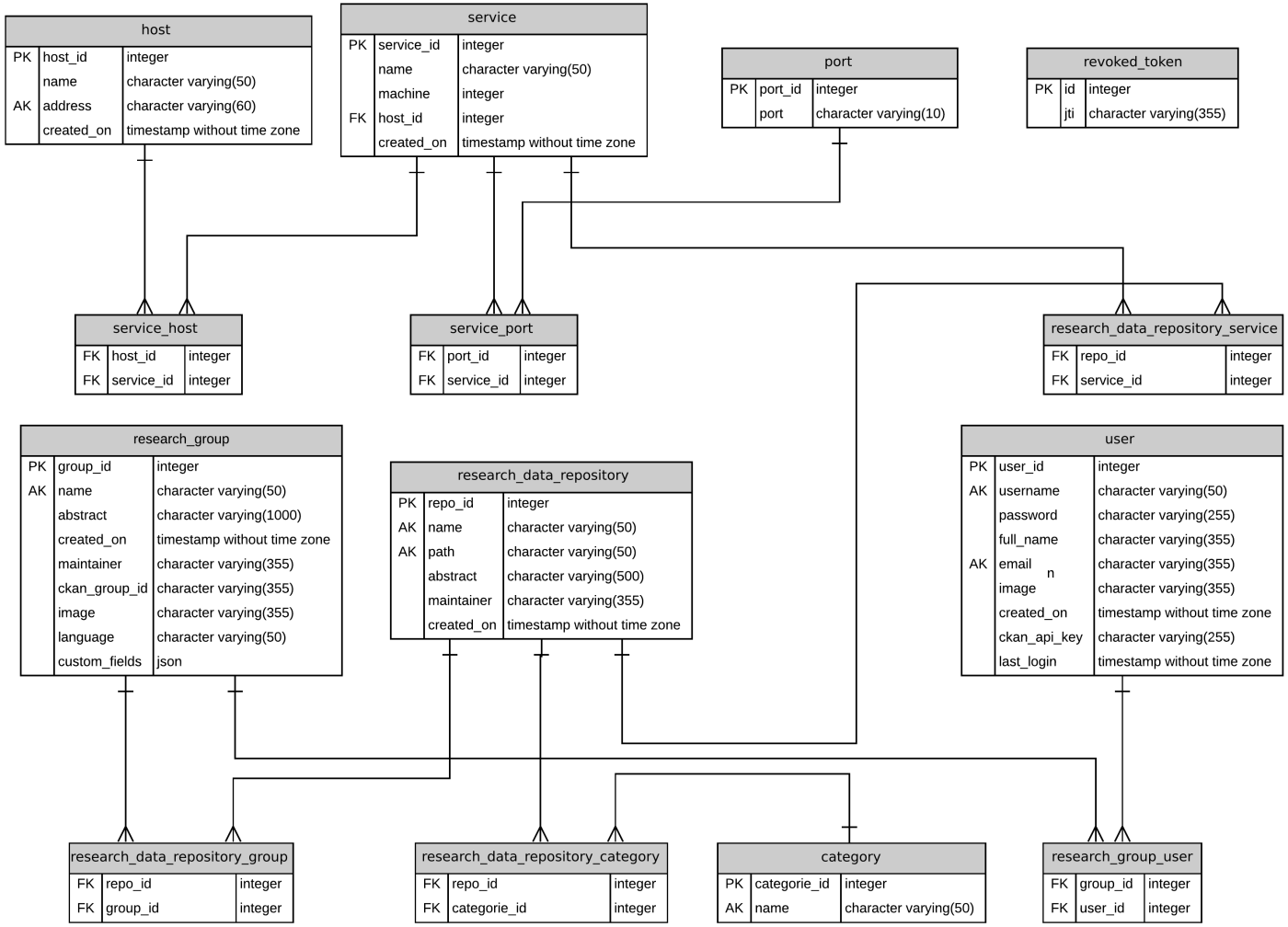
Fonte: Produção do autor

Quando feito o *upload* para todos os tipos de dados, a rota faz o armazenamento no sistema de arquivo do repositório do usuário. Caso seja um dos dados com maior suporte, de forma automatizada, é feita a ingestão do dado no banco PostgreSQL do repositório do usuário e, em alguns casos, a publicação do dado no serviço GeoServer.

TerraBrasilisRD DB

O TerraBrasilisRD DB é o banco de dados responsável pelo armazenamento dos dados e metadados da plataforma. O banco foi projetado em torno de sete entidades, são elas: (a) Usuário; (b) Grupo; (c) Repositório; (d) Serviço; (e) *Host*; (f) Porta; e (g) Categoria. Para a composição dos campos das tabelas foram usadas informações dos requisitos levantados no Capítulo 3. Com essas entidades estabelecidas foi possível projetar o modelo entidade-relacionamento do banco, ilustrado na Figura 4.14.

Figura 4.14 - Diagrama entidade-relacionamento TerraBrasilisRD DB

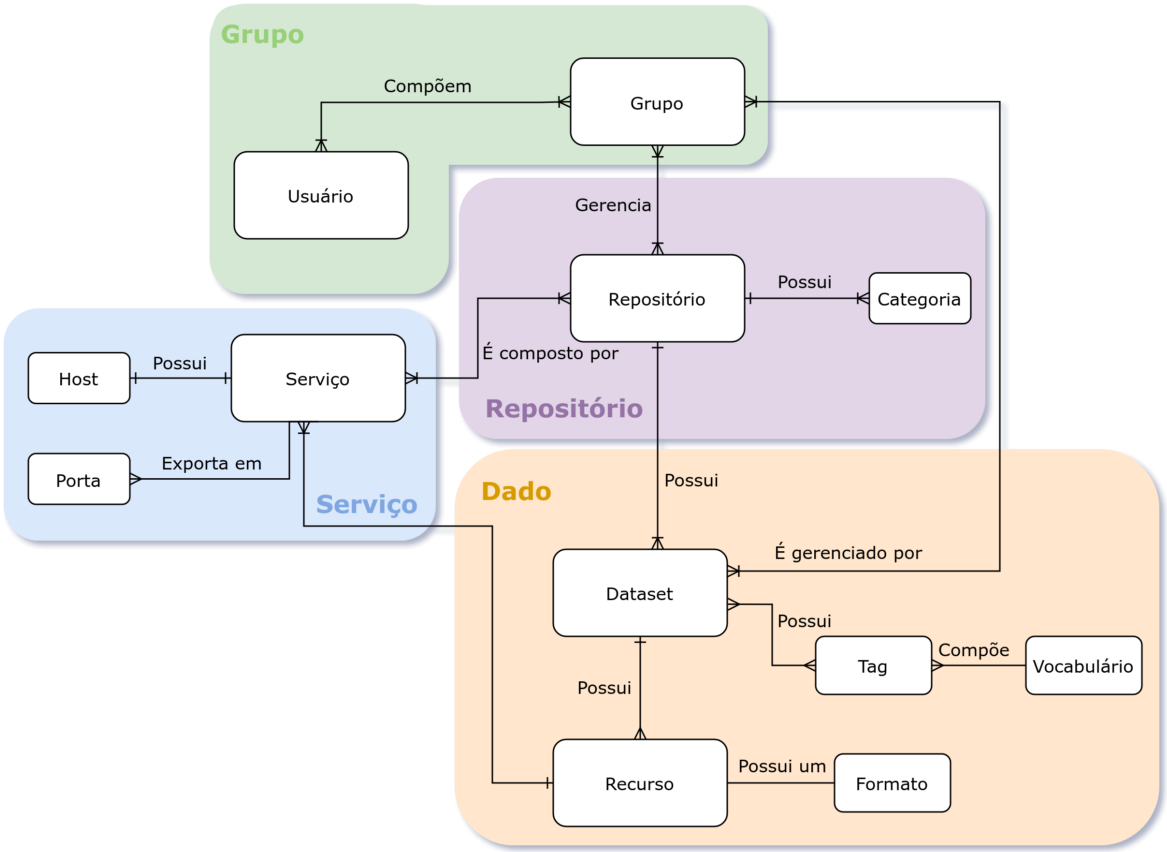


Fonte: Produção do autor

4.3.2 CKAN API

O CKAN API é o serviço *web* para gestão dos conjuntos de dados da plataforma. Ele é responsável pelo trabalho de criação, leitura, atualização e remoção de todas entidades relacionadas aos usuários e conjuntos de dados. De forma a exemplificar todas as relações entre essas entidades já apresentadas e as geridas pelo serviço a Figura 4.15 é apresentada.

Figura 4.15 - Diagrama de entidades do TerraBrasilisRD API e CKAN API



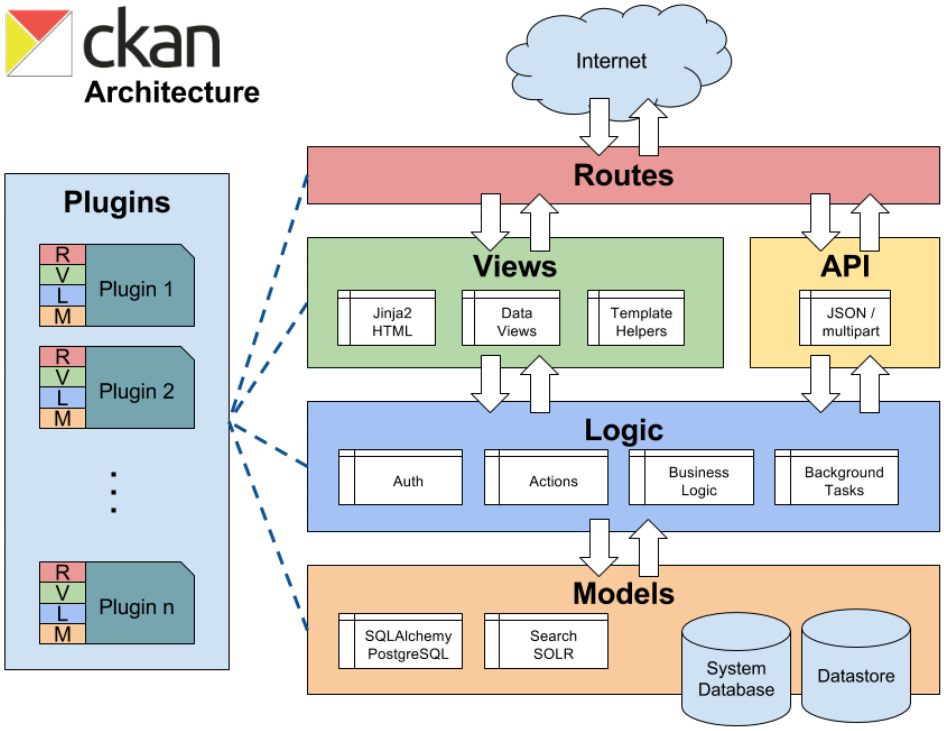
Fonte: Produção do autor

Como ilustrado na Figura 4.15, o CKAN API é responsável pelas entidades delimitadas em laranja, são elas: (a) Dataset; (b) Recurso; (c) Formato; (d) *Tag*; e (e) Vocabulário. Esse modelo permite que os conjuntos de dados atuem como conjuntos de recursos, cada um com seu respectivo formato, e possuam palavras-chaves, conjuntos de vocabulário, para a discriminação. Essa estrutura foi projetada com objetivo de ser genérica, permitindo adaptação para diferentes tipos de dados.

CKAN

A CKAN API no contexto do TerraBrasilis Research Data é composta por uma instância CKAN, isso inclui todas as interfaces de acesso, visualizações, controle e modelos. Com uma arquitetura multinível, o CKAN funciona com uma plataforma para gerenciamento de dados, fornecendo acesso à lógica da plataforma pela API através de rotas, como ilustrada na Figura 4.16.

Figura 4.16 - Arquitetura interna do CKAN



Fonte: CKAN (2020).

A arquitetura interna do CKAN (Figura 4.16), é materializada em cinco componentes: o portal de acesso, a API, o banco de dados PostgreSQL, o gerenciador de memória Redis e o motor de busca Solr. Para a plataforma proposta serão usados todos os seus componentes com exceção do portal, permitindo assim um maior aproveitamento dos recursos fornecidos pelo CKAN. Essa abordagem permite também o aproveitamento de *plugins*, extensões de funcionalidades a arquitetura interna do CKAN. Uma das extensões usadas na plataforma é a *extensão geoespacial*, que possui recursos como visualização, pesquisa e descoberta de dados.

4.3.3 Kubernetes API

O Kubernetes API é o serviço *web* para criação dos Repositórios de Dados de Pesquisa da plataforma. Ele é responsável pelo trabalho de comunicação entre o TerraBrasilisRD API e o Kubernetes. Essa funcionalidade foi projetada inicialmente como parte do TerraBrasilisRD API, porém com maiores estudos ficou claro que para se levantar um ambiente como os Repositórios de Dados de Pesquisa seria uma API própria para um melhor tráfico.

No contexto do componente, um repositório é um ambiente composto por um grupo de *pods*, um para cada uma das tecnologias e um para o armazenamento dos dados. A principal função do serviço é a de criação de repositórios, porém o serviço conta com outras funções, cada uma acessível através de uma rota da API, como mostrada na Tabela 4.2.

Tabela 4.2 - Rotas do Kubernetes API

Nome da rota	**URL**	**Descrição**
Criar um repositório	`/create`	Cria os *pods* solicitados.
Parar um repositório	`/stop`	Para todos os *pods* rodando em um repositório.
Reiniciar um repositório	`/restart`	Reinicia todos os *pods* de um repositório.
Remover um repositório	`/delete`	Remove todos os *pods* de um repositório.

Fonte: Produção do autor

O Kubernetes API possui quatro rotas mostradas na Tabela 4.2. Das rotas apresentadas na tabela, apenas a rota `/create` se conecta com o TerraBrasilisRD API. Ao criar um repositório no portal, é criado um repositório no TerraBrasilisRD DB e é acionada a criação dos *Pods*. As demais rotas são acessadas através da página serviços do portal, não passando pelo TerraBrasilisRD API. Esse serviço foi implementado na linguagem Python, usando a biblioteca oficial do Kubernetes nessa linguagem.

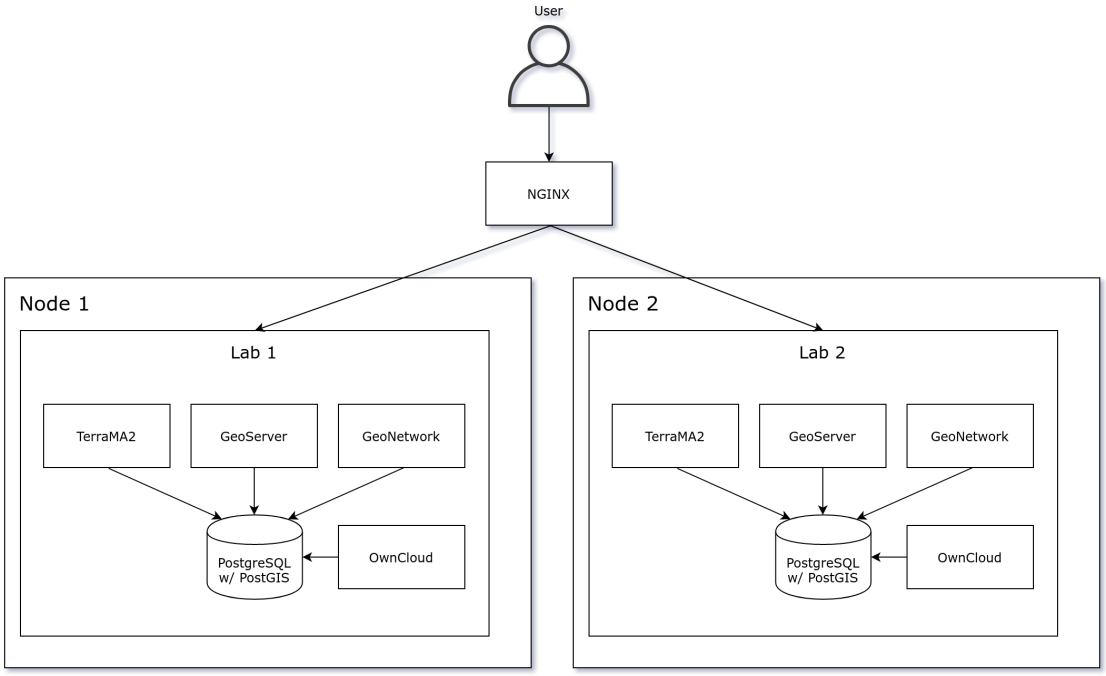
4.4 Repositório de dados de pesquisa

Um Repositório de Dados de Pesquisa é responsável por fornecer a infraestrutura computacional subjacente para armazenamento, gerenciamento, catalogação, compartilhamento e visualização dos dados científicos. A decisão de implementação desse componente considerou duas alternativas: o Docker Swarm, com vários *hosts* executados no modo enxame e atuando como gerentes e trabalhadores, e o Kubernetes. Optou-se pelo Kubernetes devida a similaridade da estrutura organizacional dessa tecnologia com o projeto dos repositórios.

4.4.1 Arquitetura

Como mostrado no Capítulo 3, foram estudadas tecnologias que integraram os repositórios, são elas: (a) PostgreSQL/PostGIS; (b) GeoServer; (c) GeoNetwork; (d) TerraMA2; e (e) OwnCloud. Essas tecnologias no ambiente Kubernetes se materializaram em *pods* e os repositórios em *namespaces*, dessa forma permitindo o compartilhamento de espaço de armazenamento e rede, característica essencial para um Repositório de Dados de Pesquisa. Cada *namespace* delimita um repositório, como ilustrado na Figura 4.17.

Figura 4.17 - Arquitetura dos Repositórios de Dados de Pesquisa



Fonte: Produção do autor

Na Figura 4.17 observa-se uma arquitetura com dois *namespaces* e dois *nodes*, porém com o uso de Kubernetes é possível escalar essa arquitetura aumentando o número de *namespaces* e *nodes*. Essa escalabilidade na arquitetura somada à possibilidade de escalar e replicar *Pods*, sem necessidade de interrompê-los como no caso Docker, representam vantagens do uso do emprego do ambiente Kubernetes.

4.4.2 Serviços

Cada uma das tecnologias que compõem um repositório é criada em espaço de infraestrutura de serviço por dois objetos serializados, o objeto referente a implantação do *Pod* e o objeto referente ao serviço, que aponta para o *Pod* permitindo acesso externo. Esses objetos são descritos através de YAMLs (formato de serialização de dados). O Kubernetes lê os YAMLs e executa a ação de criação do respectivo objeto. No Kubernetes todos os objetos podem ser descritos através de YAMLs, incluindo os *clusters*, maior entidade.

Para implementação das tecnologias foi necessária a criação de YAMLs para cada uma das tecnologias que integrariam os repositórios. Para isso foram usados como referência imagens Docker das respectivas tecnologias. Outro YAML extremamente importante para constituição dos repositórios é o Nginx, um servidor HTTP, proxy reverso e *Load Balancer*. Com ele foi possível acessar de fora do Kubernetes os serviços, seguindo uma estrutura padronizada de URLs.

`http://endereço_do_portal/nome_do_repositório/nome_do_serviço`

Um exemplo seguindo essa estrutura é

`http://terrabrasilisrd.dpi.inpe.br/labisa/geoserver`

Através desses endereços é possível acessar cada um dos serviços externamente.

4.5 Considerações finais

Neste capítulo foi apresentada a principal contribuição dessa pesquisa, a plataforma TerraBrasilis Research Data, destacando sua estrutura e os componentes que a compõem. Bibliotecas digitais, apresentadas Capítulo 2, foram projetadas com objetivo de resolver problemas de armazenamento e preservação de dados. Ambos os problemas são importantes, porém com o crescimento da demanda por compartilhamento de informações científicas, novos problemas surgiram, entre eles a integração de dados. Apesar de bastante eficientes, bibliotecas digitais estabelecidas não foram

projetadas para suportar integração de dados.

A principal vantagem da abordagem adotada pelo TerraBrasilis Research Data é a integração e o fornecimento de ferramentas para armazenamento, catalogação, gerenciamento, processamento e disseminação de dados. Essa abordagem permite que todas as atividades de uma pesquisa envolvendo observação da Terra possam ser contempladas pela plataforma. A fim de validar e demonstrar a plataforma, dois estudos de caso foram conduzidos, e são apresentados no Capitulo 5.

5 TESTE E VALIDAÇÃO DA PLATAFORMA

Para testar e validar a plataforma projetada e implementada neste trabalho dois estudos de caso foram realizados. O objetivo dos estudos foi compartilhar dados científicos geoespaciais gerados por laboratórios da OBT: o (a) LiSS e o (b) LabISA. Ambos possuem atividades em andamento no âmbito de tornar seus dados de pesquisa abertos. Foram escolhidos esses laboratórios pela diversidade dos conjuntos de dados produzidos, buscando assim garantir que todas as funcionalidades da plataforma fossem representadas.

5.1 Laboratórios

Antes de explicar como a plataforma foi validada, é importante conhecer mais sobre os laboratórios estudados.

LabISA

O Laboratório de Instrumentação de Sistemas Aquáticos (LabISA) é um dos laboratórios que compõem a Coordenação-Geral de Observação da Terra (CGOBT) do INPE. O laboratório foi criado motivado pelo aumento no número de estudos voltados a aplicações de sensoriamento remoto para estimativa de propriedades físicas, biológicas e químicas de águas continentais. A principal atividade do laboratório é coleta de dados sobre propriedades óticas e limnológicas de águas interiores e costeiras. A principal área de estudo do LabISA são sistemas de águas interiores de diferentes biomas do Brasil, como o bioma Amazônia, reservatórios em cascata do rio Tietê, Funil no Rio de Janeiro, entre outros. Pesquisadores do LabISA estudam: caracterização bio-ótica, validação de dados de sensores orbitais adquiridos sobre ambientes aquáticos, correção atmosférica de imagens dos sensores orbitais a bordo das plataformas Landsat, Sentinel, EOS-Terra e Aqua, entre outros temas. O laboratório realiza coletas de campo, processa os dados coletados e desenvolve algoritmos para a caracterização bio-ótica de ambientes aquáticos, com o objetivo de entender a dinâmica óticas. Dessa forma, é possível monitorar sistematicamente a qualidade da água por sensoriamento remoto.

Atualmente LabISA conta com um banco de dados relacional PostgreSQL com extensão PostGIS para o armazenamento dos dados tabulares e arquivos vetoriais. Além do banco, o laboratório conta com uma instância do GeoServer, para disseminação dos dados através de serviços padrão e uma instância do GeoNetwork, para visualização e edição dos metadados.

LiSS

O Laboratório de Investigação Sistemas Socioambientais (LiSS) é um dos laboratórios que compõem a Coordenação-Geral de Observação da Terra (OBT) do INPE. Ele tem como objetivo estudar a influência das atividades antrópicas nas mudanças de uso e cobertura da Terra. A principal área de estudo do LiSS é a Amazônia Legal, porém pesquisas também vêm sendo feitas na região do Vale do Paraíba do Sul (SP) e no bioma do Pantanal. Pesquisadores do LiSS estudam: sistemas urbanos, padrões e processos de mudança de uso e cobertura da Terra, dinâmicas populacionais e assentamentos humanos, medidas de desigualdades e segregação territorial, entre outros temas. Pesquisadores do laboratório produzem dados para compreensão de uso e cobertura da Terra, como por exemplo, pesquisas que estudam os processos de extrativismo nas florestas. Essas informações são essenciais para compreensão de dados de uso da Terra. Dados como esse são produzidos através de aplicações de questionários in-situ entrevistando pessoas que residem a região.

Atualmente LiSS conta com um banco de dados relacional PostgreSQL para o armazenamento dos dados tabulares e arquivos vetoriais, um espaço em servidor para armazenamento de imagens de satélite, fotos, vídeos e documentos e um sistema próprio para o gerenciamento e ingestão de dados ao banco de dados.

5.2 Organização dos estudos

Com o objetivo de avaliar todo o processo de disseminação da plataforma. Foi adotada uma abordagem em três fases: (a) Preparo dos ambientes, composto pela criação de repositórios de dados de pesquisa, com os seus respectivos grupos de usuários; (b) Ingestão dos conjuntos de dados, usando dados reais produzidos pelos laboratórios; e (c) Análise de navegabilidade e acesso: para essa fase será analisada como se comportou a plataforma com os dados e como se deu o acesso externo;

As duas primeiras fases foram replicadas para cada um dos estudos de caso e, para as fases (a) Preparação dos ambientes e (b) Ingestão dos conjuntos de dados, todo o processo de criação dos usuários, grupos, repositórios e publicação dos datasets foi feita pelo portal, para assim testar o comportamento do portal de forma rotineira.

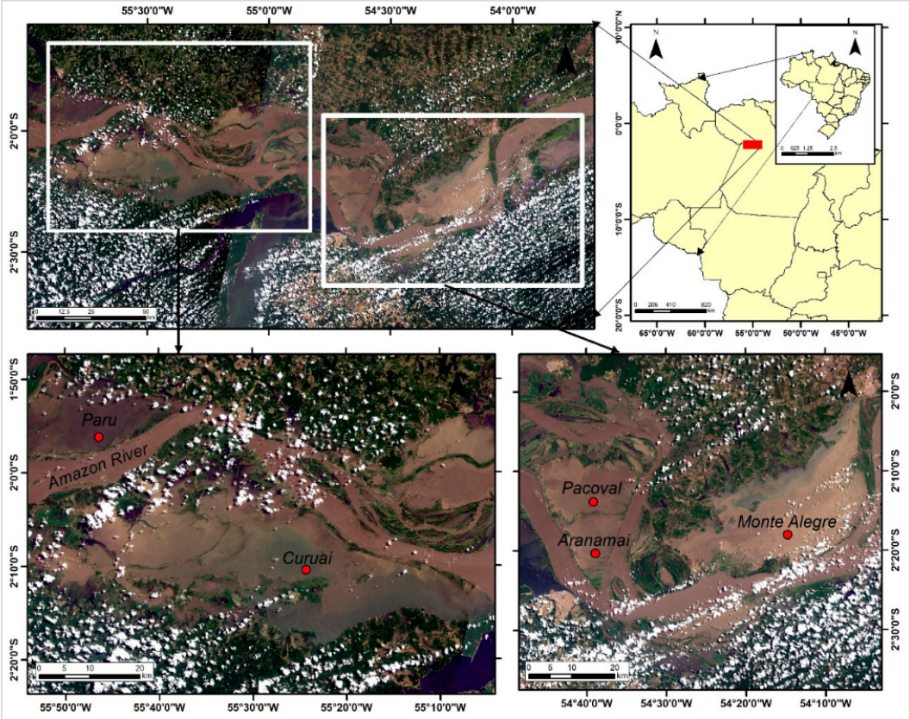
5.3 Estudo de caso 1: LabISA

O primeiro estudo de caso refere-se ao LabISA. Nesse estudo de caso foi feita a adição de tecnologias para as funções de: processamento, armazenamento e sincronia de arquivos, e visualização dos dados publicados ao processo de disseminação do laboratório.

Dados

Para esse estudo de caso foram disponibilizados dados brutos e dados pré-processados da Campanha Santarém de Agosto de 2017. Os dados do laboratório podem ser divididos pela natureza: (a) dados radiométricos: medições de radiação eletromagnética; (b) dados limnológicos: dados de águas interiores; e (c) dados IOP: dados de absorção. As áreas de estudo da campanha podem ser vistas na Figura 5.1. Os pontos em vermelho referem-se aos lagos onde as captações foram conduzidas.

Figura 5.1 - Áreas de estudo da campanha Santarém Agosto 2017 do LabISA



Fonte: Maciel et al. (2019).

No contexto das medições radiométricas os dois tipos de dados são: (a) dados brutos (extraídos/convertidos) e (b) dados processados. Ambos são representados em

vetores e matrizes, onde cada dado está relacionado a uma estação.

Processo de disseminação

A primeira etapa do processo se consistiu na organização e na projeção de uma estrutura para as publicações do laboratório, o objetivo foi organizar os conjuntos de dados em forma de publicações discrimináveis.

Os arquivos recebidos estavam organizados em uma estrutura de pastas, seguindo a hierarquia: (a) campanha; (b) tipo do dado; (c) sensor; e (d) ponto. Através de cadernetas de campo foi possível, para cada ponto, determinar seu respectivo local entre outros metadados referentes ao ponto. Com objetivo de criar um título para cada uma das publicações do laboratório foi projetada uma estrutura que agregasse todas as informações referentes ao conjunto de dados:

`Nome da Campanha` - **`Tipo do Dado`** - **`Sensor`** - **`Nome do Local`** (**`Código`**)

Um exemplo de publicação seguindo a estrutura é:

`Campanha Santarém Ago/2017 - Dados Brutos - ACS - Tapajos 04 (Tap_04)`

Com a estrutura estabelecida, para cada um dos conjuntos de dados foi possível atribuir um título. No que se refere a autores, foram usadas informações complementares do laboratório sobre a campanha, assim podendo atribuir corretamente a autoria dos conjuntos de dados. No que se refere ao *abstract*, foi redigida uma descrição para a campanha e somada a ela informações referentes aos pontos encontrados nas cadernetas de campo. Os metadados da publicação também foram gerados usando informações contidas nas cadernetas de campo.

Para a elaboração da lista de palavras-chaves foram usados: (a) nome da campanha, (b) tipo do dado, (c) sensor e (d) nome do local, dessa forma auxiliando a discriminação durante a busca de conjuntos de dados na plataforma. Para metadados espaciais como região de interesse e visualização prévia foram usados dados de arquivos vetoriais de pontos, em formato ESRI *Shapefile* provido pelo laboratório com informações referentes a todos os pontos de captação.

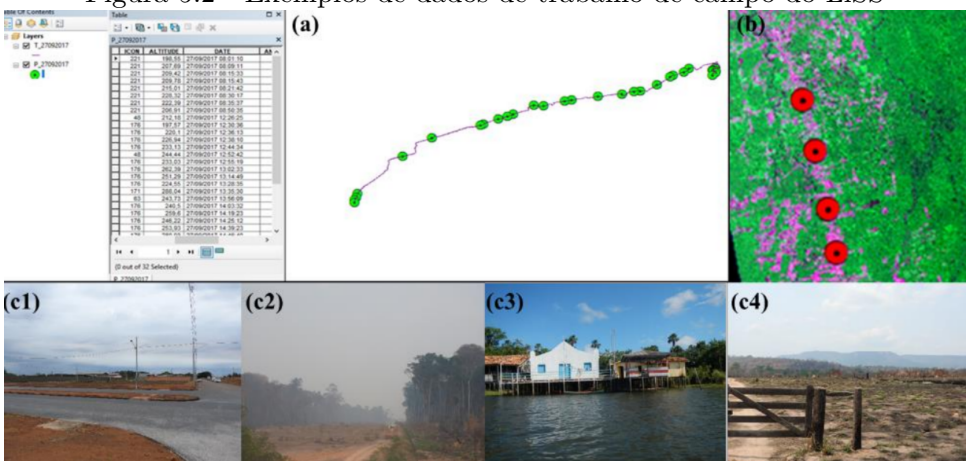
5.4 Estudo de caso 2: LiSS

O segundo estudo de caso refere-se ao LiSS. Nesse estudo de caso foi feita a adição de tecnologias para as funções de: catalogação, sincronia de arquivos e visualização dos dados publicados ao processo de disseminação do laboratório. Uma característica desse estudo de caso foi a existência da ferramenta para ingestão de dados tabulares ao banco de dados usada pelo laboratório.

Dados

Para esse estudo de caso foram disponibilizados dados de campo da Amazônia. Os dados do laboratório podem ser divididos em 5 tipos: (a) fotos e vídeos; (b) arquivos PDF, DOC, XLS, CSV e PPT; (c) imagens de satélite, (d) arquivos vetoriais e (e) gravações de áudio. A distribuição temporal dos dados é de 2008 a 2018. Uma representação dos tipos de dados do laboratório pode ser vista na Figura 5.2.

Figura 5.2 - Exemplos de dados de trabalho de campo do LiSS



Fonte: Crivellaro et al. (2018).

Exemplos de dados de trabalho de campo são: (a) planilha com localidades (pontos); (b) pontos a serem verificados em uma imagem de referência do Landsat; (c) fotos de uso e cobertura - área urbana (c1), desmatamento (c2), comunidade ribeirinha (c3) e pastagem (c4).

Processo de disseminação

A primeira etapa do processo se consistiu na organização e na projeção de uma estrutura para as publicações do laboratório. O objetivo foi organizar os conjuntos

de dados em forma de publicações discrimináveis.

Os arquivos recebidos estavam organizados em tabelas dentro de um banco de dados relacional. Algumas das entidades trabalhadas são: (a) campanha; (b) questionário; (c) pergunta; (d) resposta; (e) comunidade; e (f) pessoa. A primeira decisão estrutural foi a de como se daria a entrega dos dados aos pesquisadores e usuários da plataforma.

Com objetivo de usar o banco de dados do laboratório foi estabelecido o seguinte processo: (1) criação de uma *view* campanha relacionando as entidades do banco; (2) publicação de instâncias da *view* campanha no serviço GeoServer; (3) geração de um link para *download* e (4) publicação das campanhas como *shapefile* na plataforma. Usando o serviço GeoServer foi possível gerar links para *download* para cada uma das campanhas.

Essa *view* foi projetada com objetivo de entregar, através de uma única *query* SQL todas as informações referentes a uma campanha. Essa *view* levou em conta as relações entre as entidades: `pergunta`, `resposta`, `questionário`, `comunidade` e `pesquisa_campo`. Com objetivo de criar um título para cada uma das publicações do laboratório foi projetada uma estrutura que agregasse todas as informações referentes a conjunto de dados:

`Campanha` + `Nome da pesquisa de campo` - `Ano da pesquisa de campo`

Um exemplo de publicação seguindo a estrutura é:

`Campanha Campo Ribeirinho comunidades Tapajós - 2015`

Com a estrutura de títulos estabelecida, para atribuir corretamente a autoria dos conjuntos de dados, foram usadas informações complementares do laboratório sobre as campanhas. No que refere ao *abstract*, foi usada um resumo das campanhas somadas a descrições das respectivas campanhas. Para os metadados da publicação foram usadas informações contidas na tabela `pesquisa_campo` e no caso dos metadados espaciais, como região de interesse e visualização prévia, foram usados os metadados espaciais no formato ESRI *Shapefile*.

5.5 Ameaças a plataforma

O processo de validação da plataforma TerraBrasilis Research Data permitiu o destaque de dois pontos suscetíveis a problemas. Um deles é a demanda por organização por parte dos publicadores. Como a plataforma consiste em uma ferramenta para publicação de dados, os dados e os metadados se mostraram extremamente importantes, não só pelas suas respectivas funções para discriminação de publicações mas também pela aprestação da plataforma. Esse destaque se estende para também ao título e *abstract* das publicações, para manter um ambiente organizado é necessário que essas informações dos conjuntos de dados sejam pensadas para serem apresentaveis, como o título de um artigo. Isso se torna mais importante uma vez que os publicadores de dados, após terem suas conta ativadas, terem a permissão de criar publicações sem a necessidade de revisões ou curadorias.

Outra ameaça é possibilidade de publicação de dados externos na plataforma. Como a plataforma permite a publicação de conjuntos de dados através de *upload* e *link*, como foi no segundo estudo de caso com a publicação dos *shapefiles* através do serviço GeoServer, é importante ter em mente que todo recurso externo está suscetível ao chamado *Link rot*. *Link rot* é o fenômeno no qual *hiperlinks* deixam de apontar para o arquivo ou servidor original, devido à realocação do recurso ou indisponibilidade permanente. Essa ameaça à persistência de dados a longo prazo é resolvida através da adoção de identificadores persistentes a publicações através de *links*, além de uma curadoria da plataforma.

5.6 Resultados obtidos

Nesta subseção serão apresentados os resultados obtidos dos estudos de caso. Como o objetivo foi analisar o processo de disseminação da plataforma proposta, após a execução dos cenários foi possível concluir que todo o processo de disseminação foi contemplado pela plataforma, desde o armazenamento até a entrega dos dados. Através desses estudos foi possível também testar se a arquitetura proposta suportaria dados geoespaciais de diferentes estruturas e organizações, e como os serviços propostos se comportaram trabalhando com dados reais.

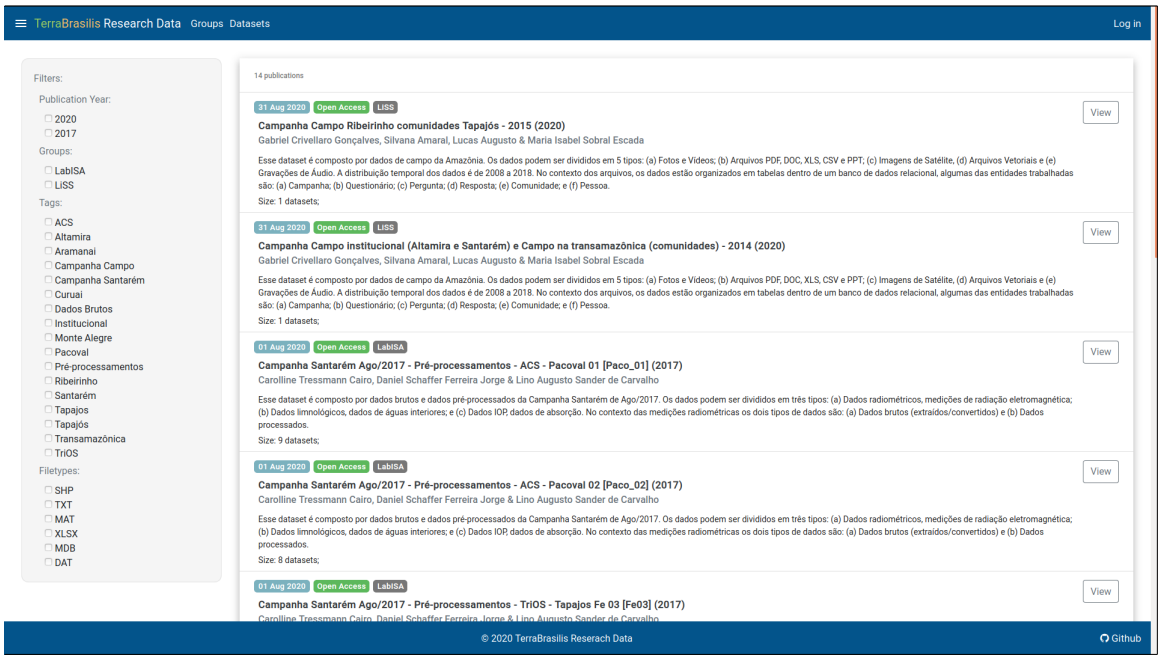
Em ambos os estudos de caso, a diversidade dos dados não foi um problema durante o processo de disseminação, inclusive um dos estudos sequer tinha seus principais dados em arquivo. Para cada um dos estudos foi adotada uma abordagem de disseminação, escolhida em conjunto com representantes desses laboratórios para garantir que todas as informações essenciais fossem entregues aos usuários em forma de pu-

blicações. Para os dois estudos, apenas parte dos dados foi adicionada à plataforma. Apesar disso, para todos os dados recebidos foram submetidos os respectivos processos de organização e estruturação propostos nessa seção.

Durante os experimentos de alimentação da plataforma, o ato de preenchimento dos metadados de uma publicação se mostrou bastante importante. Pois para cada publicação ser apresentável e discriminável na plataforma é necessário que seu título, seu *abstract* e suas palavras-chave sejam previamente pensados. Para ambos os estudos as estruturas para composição de títulos foram projetadas usando como referência as próprias estruturas de organização dos dados.

Na Figura 5.3 é mostrado a página dataset após ser alimentado com publicações do LabISA e LiSS.

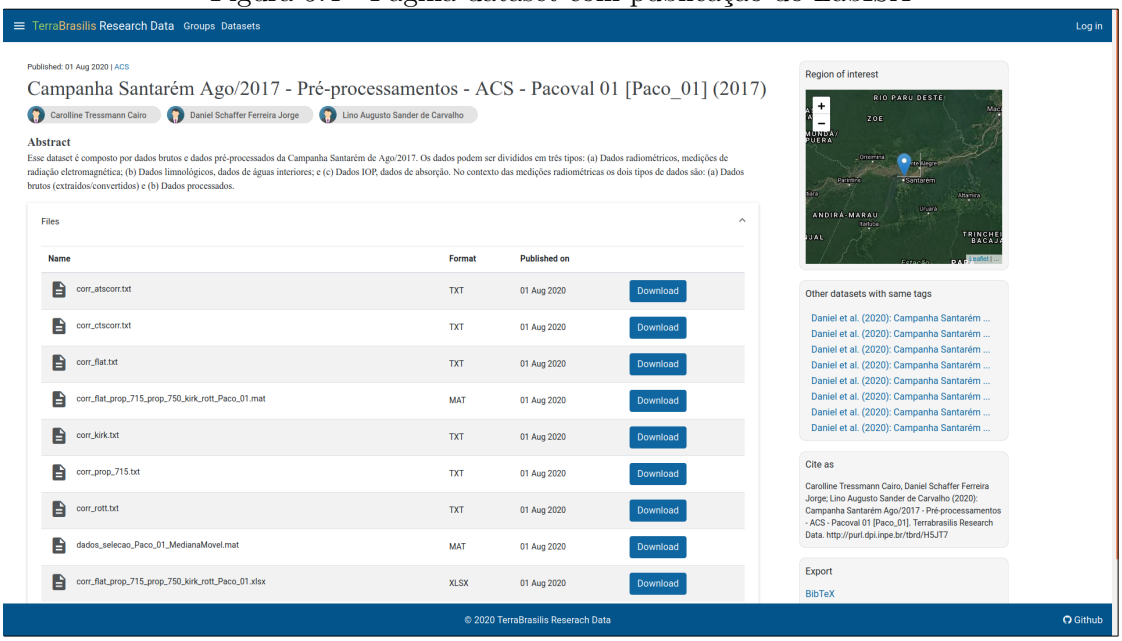
Figura 5.3 - Página datasets com publicações do LabISA e LiSS



Fonte: Produção do autor

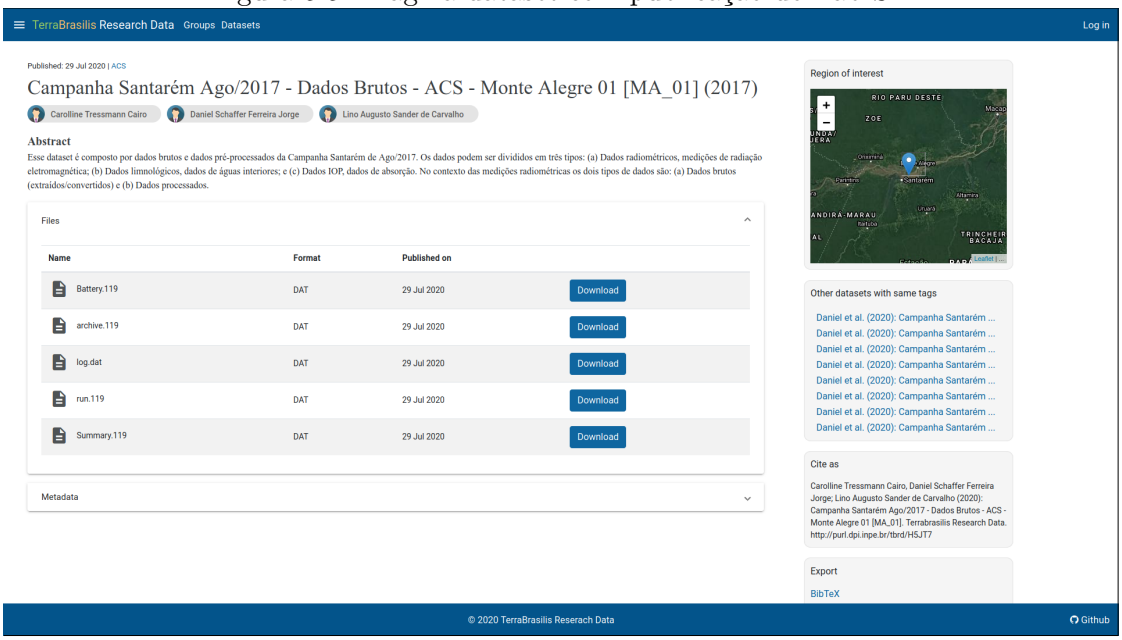
Nas Figuras 5.4 e 5.5 são apresentadas páginas de publicações do LabISA.

Figura 5.4 - Página dataset com publicação do LabISA



Fonte: Produção do autor

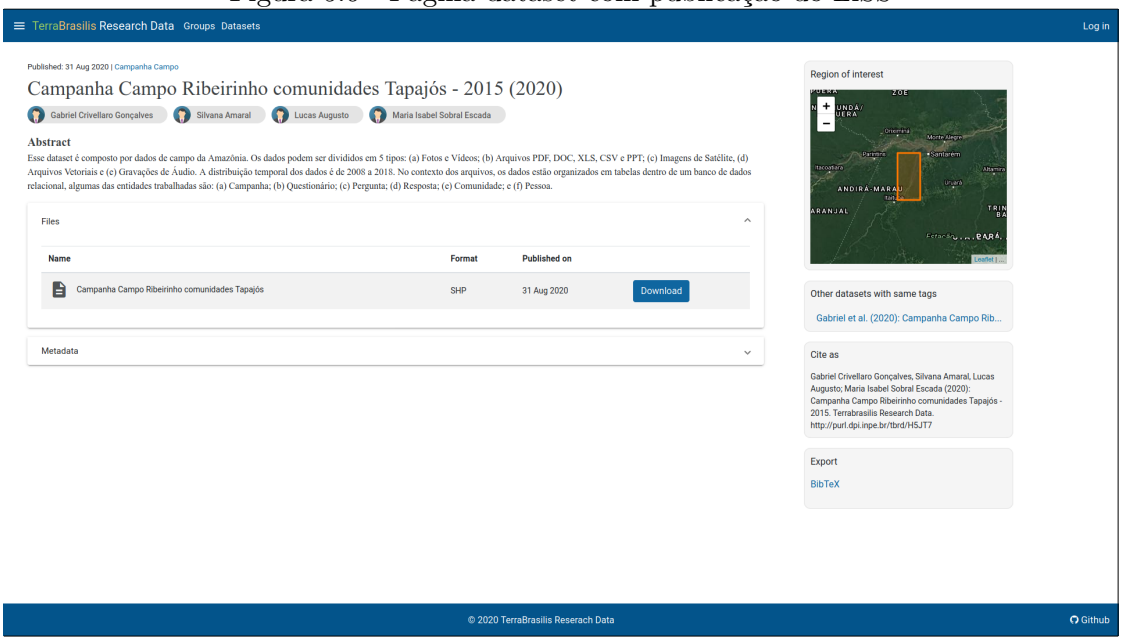
Figura 5.5 - Página dataset com publicação do LabISA



Fonte: Produção do autor

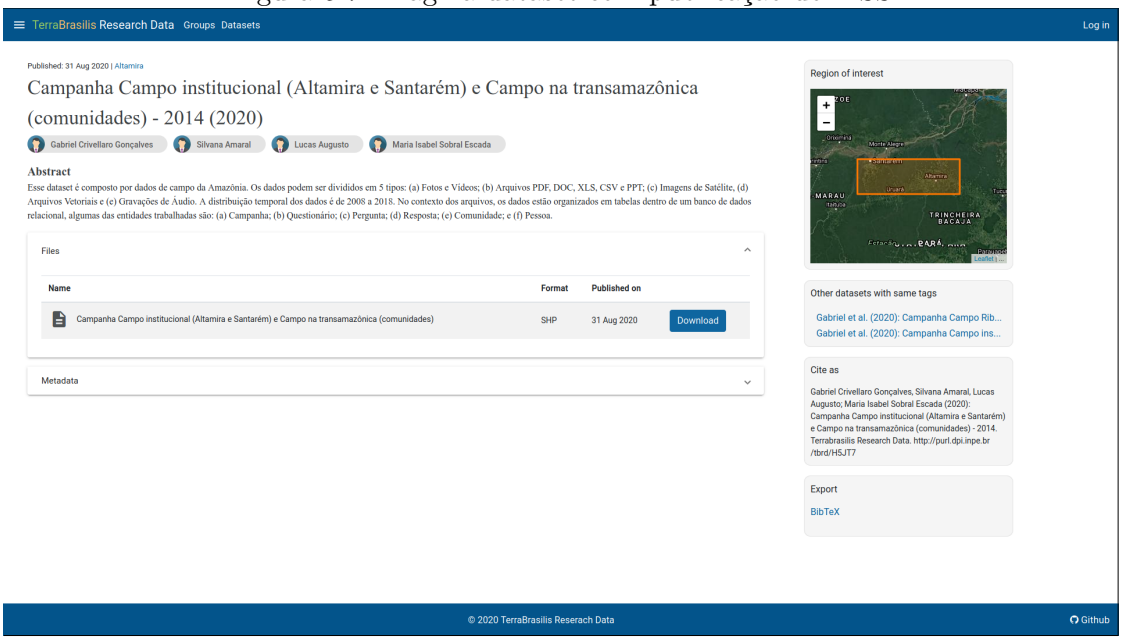
Nas Figuras 5.6 e 5.7 são apresentadas páginas de publicações do LiSS.

Figura 5.6 - Página dataset com publicação do LiSS



Fonte: Produção do autor

Figura 5.7 - Página dataset com publicação do LiSS



Fonte: Produção do autor

Nos dois estudos de caso apresentados neste capítulo, mostramos como o TerraBrasilis Research Data pode contemplar diferentes tipos de dados de pesquisa de observação da Terra. Esse suporte permitirá que pesquisadores possam publicar seus conjuntos de dados em um ambiente centralizado, expondo a comunidade científica seus resultados, e assim contribuindo com o avanço da ciência. Como o estudo de caso do LiSS foi possível destacar que grupos de pesquisa com dados armazenados em um banco de dados puderam ter seus dados publicados na plataforma.

Os resultados dos estudos de caso mostraram que a abordagem proposta, aliada a adoção de boas práticas da Ciência Aberta, foi capaz de integrar ferramentas para armazenamento, catalogação, gerenciamento, processamento e disseminação de dados. Essa abordagem, por ser aberta e integradora, permite que novas tecnologias sejam agregadas à infraestrutura computacional dos repositórios de dados de pesquisa.

O TerraBrasilis Research Data se mostrou uma ferramenta útil para realizar o compartilhamento de dados científicos de observação da Terra.

6 CONCLUSÃO

Essa dissertação apresenta o projeto e implementação de uma plataforma para gerenciamento de dados automatizada chamada TerraBrasilis Research Data, estendendo a estabelecido *framework* CKAN. A abordagem adotada foi modular, permitindo a evolução dos componentes de interface, páginas e serviços utilizados. A plataforma foi testada e avaliada através de dois estudos. Através desses estudos foi possível concluir que a abordagem modular, em conjunto com as tecnologias e serviços abertos utilizados, permitiu a criação de uma plataforma *web* capaz de suportar diferentes tipos de dados.

O TerraBrasilis Research Data é uma plataforma que resolve a demanda por automatização no processo de disseminação de dados de gerenciadores de dados (WEIGEL, 2015). Essa plataforma contempla todas as atividades de uma pesquisa, e de forma automatizada, integra diferentes tecnologias para armazenamento, catalogação, gerenciamento, processamento e disseminação de dados, no contexto de observação da Terra.

Embora a abordagem proposta seja integradora e independente de modelos de dados, ela não resolve problemas como a qualidade dos dados, por não propor uma solução automatizada para revisões ou curadorias dos conjuntos de dados publicados. Outro problema que essa abordagem não resolve é o de recursos externos suscetíveis a *Link rot*, por não exigir identificadores persistentes para publicações através de *links*.

No contexto das limitações da plataforma, foram observados diferentes casos, cada qual com seus respectivos problemas. Para caso geral a plataforma se comportou como planejado, porém implementações de funcionalidades adicionais devem ser feitas a fim de elevar a plataforma proposta ao nível de comparação com gerenciadores de dados estabelecidos.

O TerraBrasilis Research Data, diferente das demais soluções para gerenciamento de dados, entrega a pesquisadores e usuários, além de funcionalidades de gestão de dados, funcionalidades para gestão de infraestrutura. Graças aos Repositórios de Dados de Pesquisa, a plataforma fornece infraestruturas computacionais para armazenamento, gerenciamento, catalogação, compartilhamento e visualização dos dados científicos para cada grupo de pesquisa.

Todo código criado neste trabalho é livre e aberto, e está disponível no github em <`https://github.com/terrabrasilis-research-data`>. Para usar basta seguir

as respectivas documentações, encontradas nos `README.md` de cada um dos repositórios.

Para o modelo de arquitetura proposto, um artigo foi publicado no 20º Simpósio Brasileiro de Geoinformática com o título "Projetando uma plataforma para compartilhamento de dados científicos de observação da Terra", disponível em: <http://urlib.net/rep/8JMKD3MGPDW34R/3UFEDJB>.

6.1 Trabalhos futuros

Baseado nos estudos de caso e na conclusão dessa dissertação, descrevemos pontos para trabalhos futuros.

- No Portal de Dados, desenvolvimento das páginas edição de conjuntos de dados publicados, edição dos grupos de pesquisa, visualização de estatísticas referentes às publicações e visualização dos recursos computacionais da plataforma.

- No Gerenciador de Dados, implementar ferramentas de geração de identificadores persistentes para garantir que os conjuntos de dados publicados recebam uma URL persistente assim que publicados.

- Para a infraestrutura da plataforma, criação de *cluster* Kubernetes altamente disponíveis com `kubeadm`, em conformidade com as práticas recomendadas, para assim gerenciar melhor disponibilidade dos recursos computacionais alocados para os grupos de pesquisa.

- Publicar todos os dados de pesquisa do LabISA e LiSS, após realizar uma avaliação de viabilidade da plataforma com representantes desses laboratórios.

- Realizar apresentações sobre o TerraBrasilis Research Data e sobre a importância de dados de pesquisa abertos para laboratórios e grupos de pesquisa da OBT-INPE, para incentivar a publicação de dados.

- Retomar o desenvolvimento do Explorador de Dados, com adição de uma maior integração com os dados publicados da plataforma. Permitir que, ao clicar em um dado, abra-se uma janela com as principais informações referentes ao conjunto de dados.

- No contexto de implementação, levantar os requisitos computacionais para implementação da plataforma fora do ambiente local. Permitir que a plataforma e os serviços fiquem prontas para uso e abertos para acesso externo.

- No contexto de catalogação, adoção de normas estabelecidas, como a ISO 19115 ou o Perfil de Metadados Geoespaciais do Brasil (MGB). para os metadados das publicações.

- No contexto de reprodutibilidade, tornar mais fácil a replicação da plataforma proposta, com aprimoramento dos *containers* Docker e preparação de *scripts* para instalação. Além da criação de manuais para instalação e uso da plataforma.

REFERÊNCIAS BIBLIOGRÁFICAS

AMORIM, R. C.; CASTRO, J. A.; SILVA, J. R. da; RIBEIRO, C. A comparison of research data management platforms: architecture, flexible metadata and interoperability. **Universal Access in the Information Society**, v. 16, n. 4, p. 851–862, Nov 2017. ISSN 1615-5297. Disponível em: <https://doi.org/10.1007/s10209-016-0475-y>. 10, 15, 17

ASSIS, L. F.; FERREIRA, K.; VINHAS, L.; MAURANO, L.; CAMARGO, C. de; MACIEL, A.; ALMEIDA, C. A. D.; NASCIMENTO, J. R.; CARVALHO, A. Terrabrasilis: a spatial data infrastructure for disseminating deforestation data from brazil. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO. **Anais...** São José dos Campos: INPE, 2019. 18

AUER, S.; BIZER, C.; KOBILAROV, G.; LEHMANN, J.; CYGANIAK, R.; IVES, Z. Dbpedia: a nucleus for a web of open data. **Lecture Notes in Computer Science**, v. 6, p. 722–735, 01 2007. 6

BEZJAK, S.; CLYBURNE-SHERIN, A.; CONZETT, P.; FERNANDES, P.; GöRöGH, E.; HELBIG, K.; KRAMER, B.; LABASTIDA, I.; NIEMEYER, K.; PSOMOPOULOS, F.; ROSS-HELLAUER, T.; SCHNEIDER, R.; TENNANT, J.; VERBAKEL, E.; BRINKEN, H.; HELLER, L. **Open Science Training Handbook**. Zenodo, 2018. Disponível em: <https://doi.org/10.5281/zenodo.1212496>. 1, 5, 10

CKAN. **CKAN code architecture**. 2020. Disponível em: <https://docs.ckan.org/en/latest/contributing/architecture.html>. Acesso em: 26 abr. 2019. 47

CRIVELLARO, G. G.; SOUZA, L. A. d.; ESCADA, M. I. S.; KAMPEL, S. A. Spatial database to store years of earth observation information obtained from field expeditions in the amazon. In: SIMPÓSIO BRASILEIRO DE GEOINFORMáTICA. **Anais...** São José dos Campos: INPE, 2018. p. 134–139. Disponível em: <http://urlib.net/rep/8JMKD3MGPDW34P/3SG4DG8>. 57

DIEPENBROEK, M.; GROBE, H.; REINKE, M.; SCHINDLER, U.; SCHLITZER, R.; SIEGER, R.; WEFER, G. Pangaea—an information system for environmental sciences. **Computers & Geosciences**, v. 28, n. 10, p. 1201 – 1210, 2002. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0098300402000390>. 3, 6

DINIZ, C. G.; SOUZA, A. A. d. A.; SANTOS, D. C.; DIAS, M. C.; LUZ, N. C. d.; MORAES, D. R. V. d.; MAIA, J. S.; GOMES, A. R.; NARVAES, I. d. S.; VALERIANO, D. M.; MAURANO, L. E. P.; ADAMI, M. Deter-b: The new amazon near real-time deforestation detection system. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v. 8, n. 7, p. 3619–3628, 2015. 17

EATLAS. **What is GeoServer? why would I use it?** 2019. Disponível em: <https://eatlas.org.au/node/300>. Acesso em: 07 jun. 2019. 28

FOSTER. **Open science definition**. 2019. Disponível em: <https://www.fosteropenscience.eu/foster-taxonomy/open-science-definition>. Acesso em: 21 mai. 2019. 5

Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP. **Policy for open access to publications resulting from aid and FAPESP fellowships**. 2019. Disponível em: www.fapesp.br/12632. Acesso em: 26 mar. 2019. 2

Harvard University. **Harvard dataverse**. 2019. Disponível em: <https://dataverse.harvard.edu/>. Acesso em: 20 mai. 2019. 12

HERTERICH, P.; DALLMEIER-TIESSEN, S. Data citation services in the high-energy physics community. **D-Lib Magazine**, v. 22, n. 1/2, 2016. 10

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS. **Programação de monitoramento da Amazônia e demais biomas**. 2019. Disponível em: <http://terrabrasilis.dpi.inpe.br/downloads/>. Acesso em: 18 jun. 2019. 17

Invenio. **Invenio powering open science**. 2019. Disponível em: <https://invenio-software.org/>. Acesso em: 10 mar. 2019. 10, 11

KING, G. An introduction to the dataverse network as an infrastructure for data sharing. **Sociological Methods & Research**, v. 36, n. 2, p. 173–199, 2007. 12

LYNCH, C. A. Institutional repositories: essential infrastructure for scholarship in the digital age. **Portal Libraries and the Academy**, v. 3, p. 327–336, 04 2003. 1, 10

MACIEL, D.; NOVO, E.; CARVALHO, L. S. de; BARBOSA, C.; FLORES JUNIOR, R.; LOBO, F. Retrieving total and inorganic suspended sediments in amazon floodplain lakes: a multisensor approach. **Remote Sensing**, v. 11, p. 1744, 07 2019. 55

MARTONE, M. E. Data citation synthesis group: Joint declaration of data citation principles. In: . San Diego CA: FORCE11, 2014. 13

OGC. **OGC standards and supporting documents**. 2017. Disponível em: <http://www.opengeospatial.org/standards>. Acesso em: 16 mai. 2019. 7

OPEN KNOWLEDGE FOUNDATION. **The open definition**. 2019. Disponível em: <https://opendefinition.org/>. Acesso em: 22 mai. 2019. 5

OWNCLOUD. **OwnCloud overview**. 2019. Disponível em: <https://owncloud.org/>. Acesso em: 07 jun. 2019. 29

PIERRO, B. de. **Scientific communication without barriers**. 2019. Disponível em: <https://revistapesquisa.fapesp.br/2019/02/08/comunicacao-cientifica-sem-barreiras/>. Acesso em: 26 mar. 2019. 2

PONTIKA, N.; PEARCE, S.; KNOTH, P.; CANCELLIERI, M. Fostering open science to research using a taxonomy and an elearning portal. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE TECHNOLOGIES AND DATA DRIVEN BUSINESS. **Proceedings...** Graz, Austria, 2015. 5

Portal Brasileiro de Dados Abertos. **dados.gov.br**. 2019. Disponível em: <http://dados.gov.br/>. Acesso em: 20 mar. 2019. 14

RAJABIFARD, A.; MANSOURIAN, A.; ZOEJ, M. J. V.; WILLIAMSON, I. Developing spatial data infrastructure to facilitate disaster management. In: . [S.l.: s.n.], 2004. 7

RESEARCH DATA ALLIANCE. **About RDA**. 2020. Disponível em: <https://www.rd-alliance.org/about-rda>. Acesso em: 20 mar. 2020. 6

SAEZ, R. V.; FUENTES, C. M. Open science now: a systematic literature review for an integrated definition. **Journal of Business Research**, v. 88, p. 428 – 436, 2018. ISSN 0148-2963. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0148296317305441>. 1, 5

Scimago Institutions Rankings. **Scimago journal & country rank**. 2019. Disponível em: <https://www.scimagojr.com/countrysearch.php?country=br>. Acesso em: 26 mar. 2019. 1

SICILIA, M.-A.; GARCíA-BARRIOCANAL, E.; SáNCHEZ-ALONSO, S. Community curation in open dataset repositories: insights from zenodo. In:

Procedia Computer Science. 2017. v. 106, p. 54–60. Disponível em: <www.sciencedirect.com/science/article/pii/S1877050917302776>. 10

WAINWRIGHT, M. **Using CKAN: storing data for re-use**. 2012. Disponível em: <https://ckan.org/files/2012/08/OKF-OR12-poster.pdf>. Acesso em: 21 mar. 2019. 14, 15

WEIGEL, T. **Persistent identifiers for Earth science data management**. Tese (Doutorado) — Universität Hamburg, Von-Melle-Park 3, 20146 Hamburg, 2015. Disponível em: <http://ediss.sub.uni-hamburg.de/volltexte/2016/7844>. 3, 6, 9, 65

WILKINSON, M.; DUMONTIER, M.; AALBERSBERG, I. J.; APPLETON, G.; AXTON, M.; BAAK, A.; BLOMBERG, N.; BOITEN, J.-W.; SANTOS, L. O. Bonino da S.; BOURNE, P.; BOUWMAN, J.; BROOKES, A. J.; CLARK, T.; CROSAS, M.; DILLO, I.; DUMON, O.; EDMUNDS, S.; EVELO, C.; FINKERS, R.; MONS, B. The fair guiding principles for scientific data management and stewardship. **Scientific Data**, v. 3, 03 2016. 1, 6

WOELFLE, M.; OLLIARO, P.; TODD, M. H. Open science is a research accelerator. **Nature Chemistry**, v. 3, p. 745 EP –, Sep 2011. Disponível em: <https://doi.org/10.1038/nchem.1149>. 1, 5