Arquitetura para publicação de dados sobre biodiversidade em instituições de pesquisa

Architecture for Publication of Data on Biodiversity in Research Institutions

Eduardo Dalcin*

João Lanna**

Natália Queiroz***

Rafaela Campostrini Forzza****

RESUMO

Desde a Declaração de Berlin sobre o Acesso Aberto ao Conhecimento em Ciências e Humanidades, publicada em 2003, a demanda por uma "ciência aberta" cuja preocupação primordial é tornar a atividade de pesquisa mais transparente, mais cola-borativa e mais eficiente, tem crescido na comunidade acadêmica. Aliado a isso, vem se consolidando a percepção de que o acesso e compartilhamento de dados de pesquisa contribui de forma significativa para que a ciência avance e maximize os investimentos aplicados em programas de pesquisa. Neste sentido este estudo apresenta uma proposta composta de repositórios digitais e ferramentas computacionais voltadas para publicação e compartilhamento de recursos de informação em institutos de pesquisa. A arquitetura proposta, baseada em ferramentas livres e de código aberto mostrou-se adequada à gestão e

ABSTRACT

Since the Berlin Declaration on Open Access to Knowledge in Science and Humanities published in 2003, the demand for an "open science" whose primary concern is to make research activity more transparent, more collaborative and more efficient, has grown at the academy. Added to this, the perception that the access and sharing of research data contribute significantly to science advance and maximize the investments applied in research programs has been consolidated. In this sense, the present work presents a proposal composed of digital repositories and computational tools aimed at publishing and sharing of information resources in research institutes. The proposed architecture, based on free and open-source tools, proved adequate for the management and publication of information resources in research institutions. However, this approach

* Doutor em Informática aplicada à Biodiversidade pela Universidade de Southampton, no Reino Unido. Instituto de Pesquisas Jardim Botânico do Rio de Janeiro. Endereço: Rua Pacheco Leão 915 sl 203, Rio de Janeiro - RJ, 22460-030. E-mail: edalcin@jbrj.gov.br

** Mestre em Ecologia pela Universidade Federal de Ouro Preto-MG. Instituto de Pesquisas Jardim Botânico do Rio de Janeiro. Endereço: Rua Pacheco Leão 915 sl 203, Rio de Janeiro - RJ, 22460-030. E-mail: joaomlanna@gmail.com

*** Mestre em Sistemas e Computação pelo Instituto Militar de Engenharia. Instituição: Instituto de Pesquisas Jardim Botânico do Rio de Janeiro. Endereço: Rua Pacheco Leão 915 sl 203, Rio de Janeiro - RJ, 22460-030. E-mail: queiroz.nati@gmail.com

**** Doutora em Ciências Biológicas (Botânica) pela Universidade de São Paulo. Instituição: Instituto de Pesquisas Jardim Botânico do Rio de Janeiro. Endereço: Rua Pacheco Leão 915, Rio de Janeiro - RJ, 22460-030. E-mail: rafaela@jbrj.gov.br

publicação de recursos de informação em instituições de pesquisa. Porém, esta abordagem apontou a necessidade de uma ferramenta de busca que integre as diferentes ferramentas, assim como da existência de um vocabulário controlado, capaz de indexar os recursos em seus diferentes contextos.

Palavras-chave: Dados Abertos; Ciência Aberta; Publicação de Dados Científicos.

pointed to the need for a search tool that integrates the different tools, as well as the existence of a controlled vocabulary, capable of indexing resources in their different contexts.

Keywords: Open Data; Open Science; Scientific Data Publishing.

INTRODUÇÃO

O Instituto de Pesquisas Jardim Botânico do Rio de Janeiro (JBRJ) tem como missão "*Promover, realizar e difundir pesquisas científicas, com ênfase na flora, visando à conservação e à valoração da biodiversidade, bem como realizar atividades que promovam a integração da ciência, educação, cultura e natureza*" (JBRJ, 2019). Neste sentido, em 2014 o JBRJ assinou uma Carta de Acordo com o Ministério da Ciência, Tecnologia e Inovação (MCTIC) para o desenvolvimento do Projeto "*Contribuições do Jardim Botânico do Rio de Janeiro à implementação do SiBBr – Sistema de Informação sobre a Biodiversidade Brasileira*". Este projeto, possuía quatro componentes, a saber:

- Repatriamento
- Flora do Brasil
- Herbário Virtual
- Integração de Dados

O componente de "Integração de Dados" tinha como objetivo "*Tornar o JBRJ uma instituição-modelo no tratamento, qualificação e compartilhamento de dados sobre biodiversidade, provendo dados qualificados sobre espécies, ocorrências, observações de campo e produção intelectual sobre a biodiversidade de plantas brasileira à sistemas voltados para integração, tratamento e análise de dados sobre biodiversidade, como o SiBBr, GBIF, Flora do Mundo Online, IUCN,etc.*" (DALCIN et al., 2017)

Desta forma, através dos recursos aportados para o desenvolvimento deste projeto, foi formada uma equipe de um coordenador e três bolsistas para o desenvolvimento e implementação desta arquitetura aqui apresentada.

REPOSITÓRIOS DIGITAIS E DADOS INSTITUCIONAIS

O JBRJ, assim como várias outras instituições, "produz e possui sob sua guarda um conjunto de recursos de informação variado sobre biodiversidade. Vamos considerar neste texto "***recursos de informação***" como sendo documentos (relatórios, dissertações, teses, publicações, etc), planilhas (dados estruturados, tabulados sob a forma de planilhas, arquivos de texto delimitados ou bancos de dados), apresentações (séries de "slides"), imagens (estáticas ou em movimento – vídeo), livros e publicações, mapas (em formato vetorial ou "raster") e objetos em coleções científicas" (DALCIN, 2016) (Figura 1).

Figura 1. Recursos de informação da DIPEQ-JBRJ (DALCIN, 2016)

Dentro dos diferentes projetos e atividades de pesquisa desenvolvidas pela Diretoria de Pesquisas da Instituição (DIPEQ-JBRJ), o processo de avaliação de risco de extinção, levado à cabo pelo Centro Nacional de Conservação da Flora (CNCFlora), foi selecionado como estudo de caso para este trabalho, tendo em vista a variedade de recursos de informações utilizados ou gerados por esta iniciativa (e.g. status de conservação das espécies da flora brasileira e desenvolvimento de planos estratégicos para conservação de espécies CNCFLORA 2019).

Repositórios digitais são coleções de dados e informações em formato digital, que podem ser construídas de diferentes formas e com diferentes propósitos ou, segundo a definição do Digital Repositories JISC Briefing Paper (2005), "um repositório digital é aquele onde conteúdos, recursos, estão armazenados e podem ser pesquisados e recuperados para uso posterior. Um repositório suporta mecanismos de importação, exportação, identificação, armazenamento e recuperação de recursos digitais" (MARTINS, 2008).

Muitos dos repositórios digitais institucionais são voltados para publicação *on-line* da produção científica. No caso do CNCFlora, além desta demanda, os analistas responsáveis por elaborar a avaliação de risco de extinção das espécies, com base na metodologia da IUCN, necessitam organizar e publicar um conjunto diverso de recursos de informação para cada espécie avaliada. Dentre estes recursos de informação, podemos destacar, além dos documentos não estruturados (relatórios, artigos científicos, etc.), conjuntos de imagens (de espécies ou do ambiente que ocorrem e as ameaças a sua conservação), mapas contendo a distribuição da ocorrência e dados estruturados, como planilhas contendo informações sobre cada uma das espécies avaliadas.

ARQUITETURA DE REPOSITÓRIOS DIGITAIS

O objetivo da arquitetura proposta é viabilizar o compartilhamento de diferentes tipos de recursos de informação sobre biodiversidade relacionados com os projetos,

coleções e laboratórios da instituição, oferecendo estes dados em formatos padronizados para *download* e integração com outros sistemas, como o Sistema de Informação sobre a Biodiversidade Brasileira - SiBBr (GADELHA *et al.* 2014). Esta arquitetura possibilita uma busca integrada pelas diferentes ferramentas e tipos de dados (Figura 2) .

A arquitetura proposta segue os seguintes princípios:

- As ferramentas utilizadas na arquitetura devem ser de código livre e aberto;
- Todo o dado disponível deve ter um conjunto de metadados associado, em formato aberto, padronizado e também digitalmente acessível;
- Todo o dado e metadado disponível deve ser associado a um responsável pela sua qualidade, consistência e integridade;
- Todo o dado disponível é considerado público, e seu uso é definido pela licença citada nos metadados correspondentes.

Com base nestes princípios, procedeu-se uma prospecção de ferramentas disponíveis no mercado, onde foram selecionadas as seguintes ferramentas:

- **Ckan**
  - O Ckan (WINN, 2013) acrônimo de Comprehensive Knowledge Archive Network é uma ferramenta de código livre e aberto baseada na web, criada para a o armazenamento, gestão e publicação de dados abertos. É utilizado em vários sites do governo federal, inclusive no Portal Brasileiro de Dados Abertos.
- **DSpace**
  - O DSpace (SMITH *et al.*, 2003) foi desenvolvido para possibilitar a criação de repositórios digitais com funções de armazenamento, gerenciamento, preservação e visibilidade da produção intelectual. Os repositórios DSpace permitem o gerenciamento da produção científica em qualquer tipo de material digital, dando-lhe maior visibilidade e garantindo a sua acessibilidade ao longo do tempo. O DSpace é largamente utilizado em todo mundo. No Brasil, vem recebendo apoio e divulgação do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) desde 2004, sendo utilizado como repositório digital por diversas instituições públicas, possibilitando sua adoção em forma consorciada federada (IBICT, 2019).
- **ResourceSpace**
  - O ResourceSpace (JIGNESH & VIRAL, 2012) é uma ferramenta também de código livre e aberto para gerenciamento e publicação de acervos digitais, especialmente voltado para imagens, áudio e vídeo.
- **GeoNode**
  - O GeoNode (CORTI *et al.* 2019) é uma plataforma livre para a catalogação e publicação de dados espaciais, em formato vetorial ou matricial. Os dados e metadados catalogados são oferecidos também como *web services* para acesso em outras aplicações.

Dentro da arquitetura, houve a necessidade de integrar um sistema "proprietário", o JABOT2. O JABOT2 (DA SILVA *et al.* 2017) é o sistema desenvolvido para suprir

demandas de digitalização das coleções biológicas do JBRJ. Hoje abriga 100% do acervo do herbário, xiloteca, carpoteca, *spirit* (coleção armazenada em álcool 70%), banco de sementes e de DNA, além de grande parte da coleção viva, disponíveis na web e integra ferramentas de busca espacial.

Com essas ferramentas foi possível, não só, cobrir todos os tipos de recursos de informação utilizados pelo CNCFlora em suas atividades, mas também atender eventuais demandas institucionais e de pesquisadores, projetos de pesquisa e laboratórios que produzem ou tem sob sua guarda recursos de informação sobre biodiversidade.

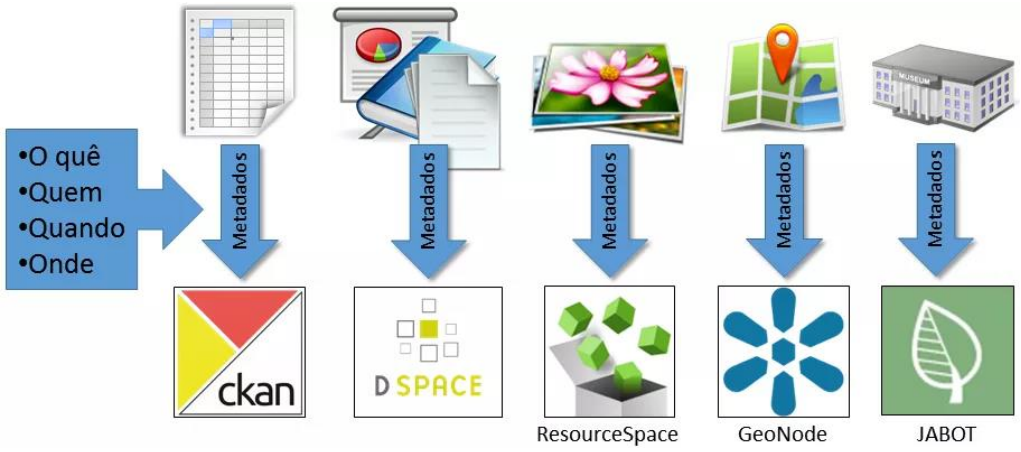

Figura 2. Arquitetura de ferramentas para publicação de dados da DIPEQ-JBRJ (DALCIN, 2016)

A estas ferramentas, foram acrescentadas um conjunto de páginas, desenvolvidas em HTML; e uma "Wiki", com base na ferramenta DokuWiki, para compor o Portal de Dados institucional (http://dados.jbrj.gov.br), conforme a Figura 3.

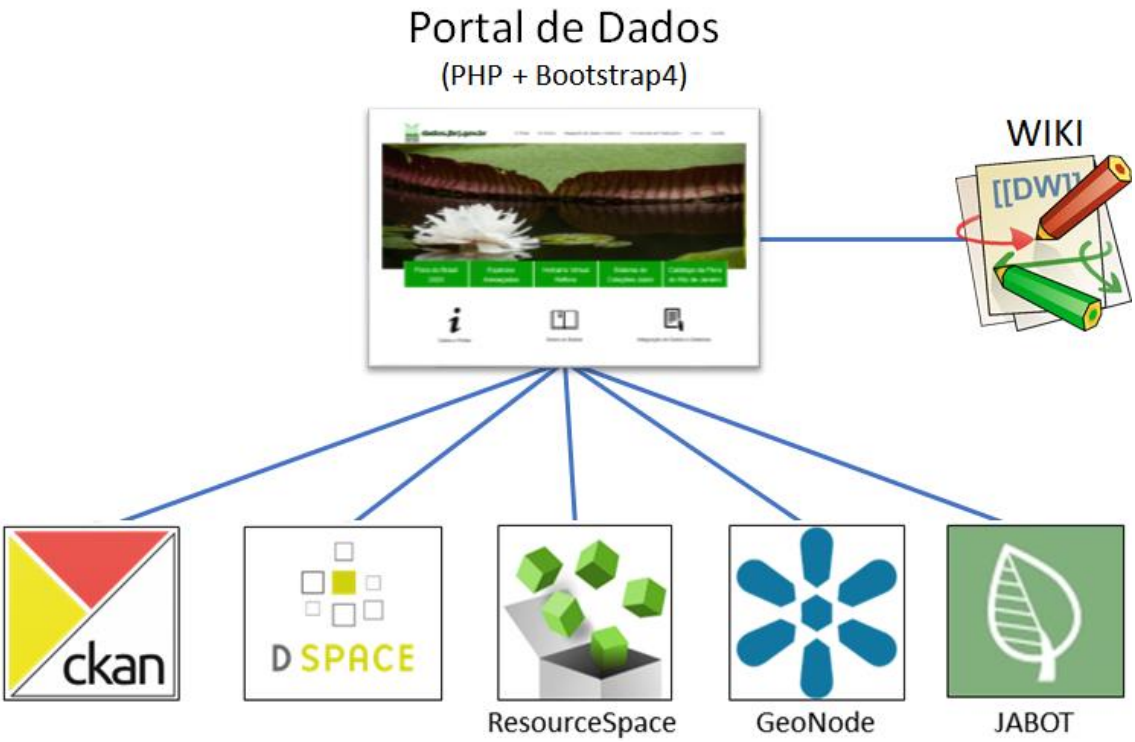

Figura 3. Arquitetura do Portal de Dados com ferramentas associadas

PADRÕES DE METADADOS ADOTADOS

Para catalogação e descrição dos recursos de informação em cada uma das ferramentas, foram utilizados padrões de metadados específicos (Tabela 1).

Ferramentas	Padrão de Metadados Adotado
Ckan	Ecological Metadata Language (EML) (FEGRAUS *et al.*, 2005) / Dublin Core (DC) (TANSLEY *et al.*, 2003)
DSpace	Dublin Core (DC) (TANSLEY *et al.*, 2003)
Geonode	Perfil de Metadados Geoespaciais do Brasil (Perfil MGB) (INDE, 2015)
JABOT2	Darwin Core Archive (DwC-A) (WIECZOREK *et al.*, 2012)
ResourceSpace	XMP, IPTC e EXIF

Tabela 1. Ferramentas de metadados e os padrões adotados

O código-fonte do Portal de Dados e a customização feita nas ferramentas associadas, assim como a especificação do hardware necessário para sua implementação, está disponível em https://github.com/DIPEQ-JBRJ/portal_de_dados_jbrj.

RESULTADOS E DISCUSSÃO

Com o Portal de Dados e as ferramentas associadas em "produção", ou seja, recebendo conteúdo através da catalogação dos recursos de informação correspondentes e sendo acessadas e consultadas pelos usuários, ficou clara a necessidade de alguns ajustes na arquitetura proposta. Trazemos à discussão dois dos aspectos mais importantes desta necessidade de ajuste.

A busca por recursos de informação em repositórios heterogêneos

No estabelecimento de um conjunto de ferramentas heterogêneas para gestão (catalogação e publicação on-line) de recursos de informação também heterogêneos, nos deparamos com o problema ilustrado pelo seguinte caso: um usuário deseja saber quais os recursos de informação que a instituição tem em seus repositórios sobre, por exemplo, o "pau-brasil".

Na forma em que a arquitetura está proposta, o usuário teria que acessar cada ferramenta, uma a uma, fazendo a pesquisa "pau-brasil" (Figura 4).

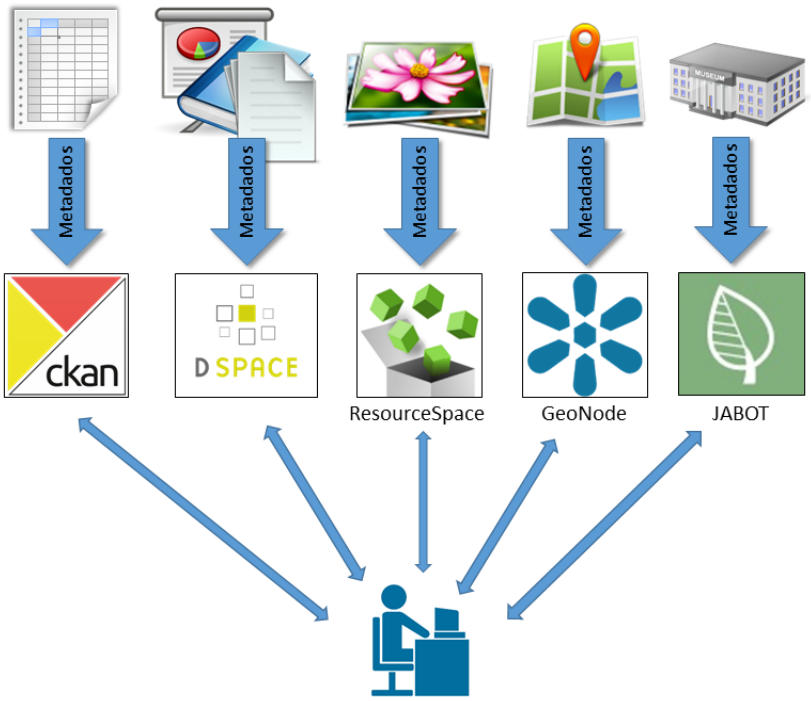

Figura 4. Busca por recursos de informação em repositórios heterogêneos e distribuídos

Entretanto, a situação ideal seria oferecer ao usuário uma *interface* "*Google like*" onde sua consulta seria feita de forma automática em todos os repositórios.

Uma vez que todas as ferramentas possuem "APIs" - Application Programming Interface - um conjunto de rotinas e padrões estabelecidos por um software para a utilização das suas funcionalidades por aplicativos que não pretendem envolver-se em detalhes da implementação do software, mas apenas usar seus serviços (INTERFACE, 2019), é possível adicionar essa funcionalidade, fazendo com que um "Motor de Busca" (*search engine*) seja responsável por receber a consulta do usuário e devolver os resultados, em uma *interface* unificada (Figura 5).

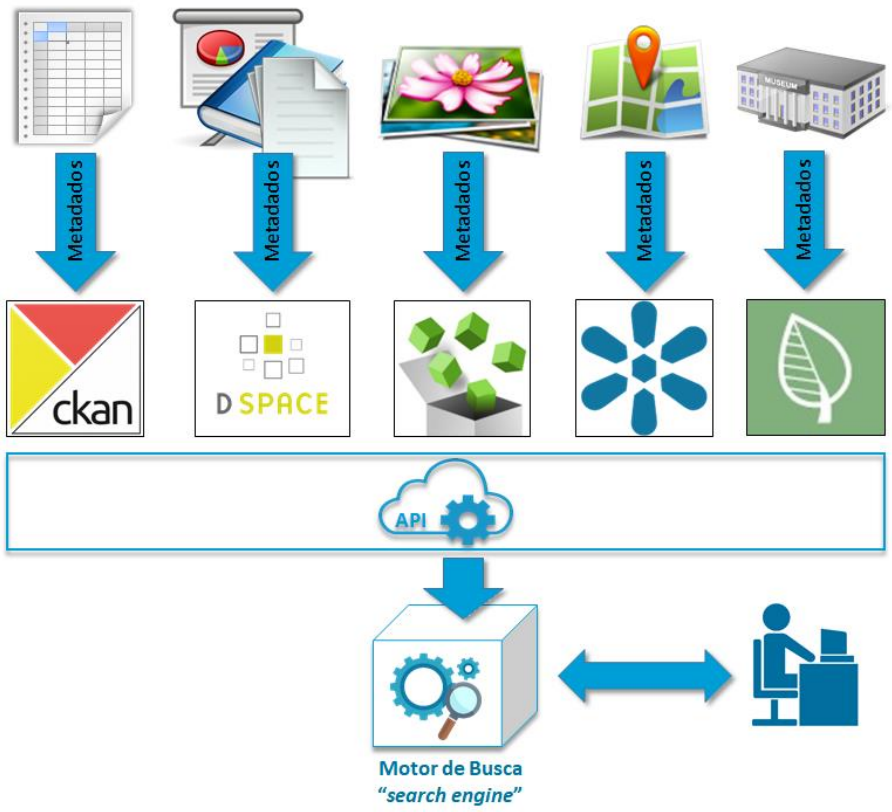

Figura 5. Implementação de uma Ferramenta de Busca Integrada na arquitetura proposta (DALCIN, 2016)

Uma vez instalado e configurado o Motor de Busca, baseado na ferramenta "Elastic Search" (DIVYA & GOYAL, 2013), foi desenvolvida uma "*interface*" para ofertar ao usuário a consulta integrada em todos os repositórios (Figura 6).

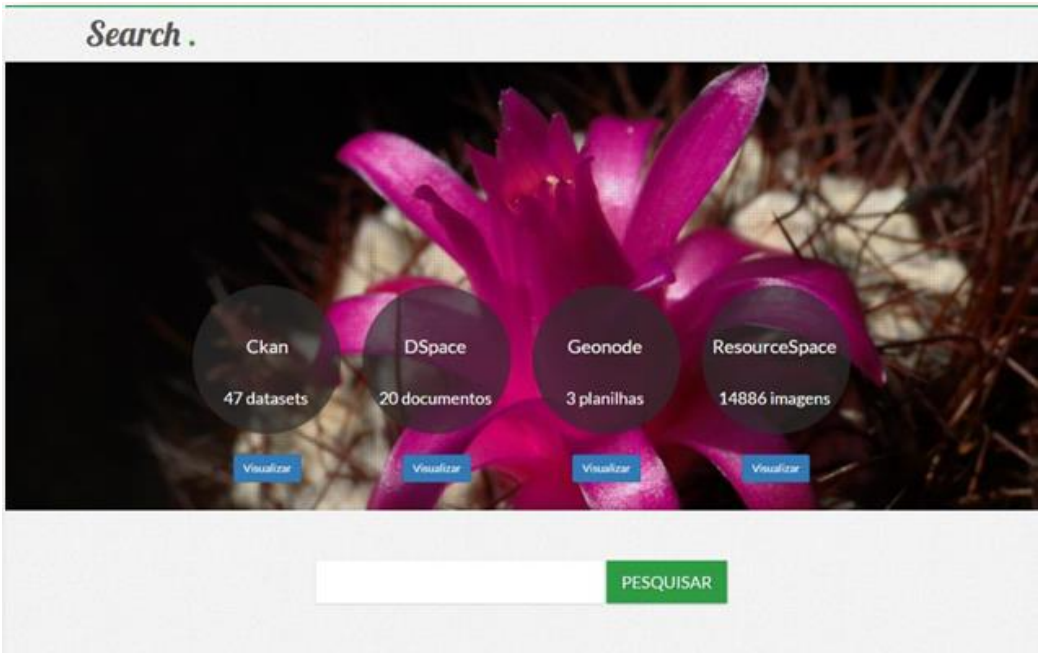

Figura 6. Interface da Ferramenta de Busca Integrada

O desafio da indexação de recursos de informação sobre biodiversidade

A indexação é um método de busca para combinar o tratamento do conteúdo dos documentos e a recuperação pelo usuário. Ele descreve e identifica o documento por meio de conceitos que o representam, assim como auxilia na busca e no acesso à informação armazenada. Ainda, ao se tratar dos sistemas de informação, a indexação é a responsável por condicionar os resultados das estratégias de busca, ou seja, o bom desempenho da indexação refletirá sobre a recuperação da informação presente nas bases de dados (PEREIRA *et al.* 2015).

As informações sobre biodiversidade têm quatro contextos fundamentais: o taxonômico, o espacial, o temporal e o temático (DALCIN, 2016).

Figura 7 - Contextos da informação sobre biodiversidade (DALCIN, 2016)

O contexto taxonômico associa o recurso de informação a um grupo taxonômico específico, como uma espécie de animal ou vegetal, ou um outro grupo taxonômico supra-específico, como uma família de plantas, por exemplo.

O contexto espacial associa o recurso a um ponto, geralmente uma latitude e longitude, ou mesmo uma região ou divisão geopolítica, como país, estado e município. Da mesma forma, o contexto temporal associa o recurso a uma data ou período de tempo.

O contexto temático associa o recurso a uma característica ou tema específico, como por exemplo "espécies de plantas comestíveis" ou "animais ameaçados de extinção".

Um exemplo prático poderia ser um documento que trata das espécies de BROMELIACEAE ameaçadas de extinção que ocorrem no Parque Nacional de Itatiaia. Este recurso tem os seguintes contextos para indexação:

Contexto	Valor
Taxonômico	BROMELIACEAE
Espacial	Parque Nacional de Itatiaia
Temático	Espécies Ameaçadas de Extinção

Tabela 2. Exemplos de contexto em biodiversidade

Entretanto, em todos os contextos encontramos termos diferentes que se referem à mesma entidade do mundo real (Figura 8).

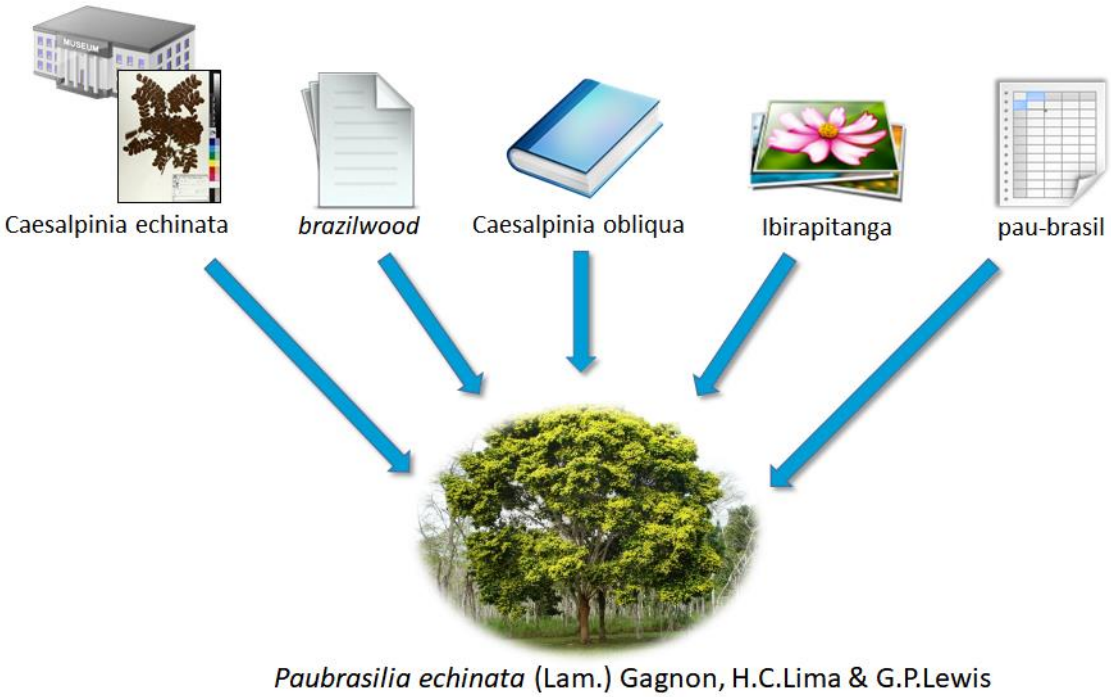

Figura 8 - Diferentes termos usados para descrever a mesma entidade (DALCIN, 2016)

Na figura 8, temos diferentes termos do contexto taxonômico que referenciam uma mesma e única espécie (taxon). Desta forma, caso um usuário deseje saber quais os recursos de informação que o Jardim Botânico do Rio de Janeiro possui sobre o "pau-brasil", os repositórios devem retornar recursos indexados por todos os termos associados. Fica claro então a necessidade de integrar à arquitetura inicialmente proposta uma ferramenta adicional para armazenar os termos relacionados.

Na prospecção das ferramentas disponíveis encontramos o TemaTres (https://www.vocabularyserver.com) - um servidor de vocabulário controlado de código aberto e gratuito que, através de sua API, pode ser integrado à arquitetura proposta possibilitando que o termo de pesquisa definido pelo usuário na interface

de busca seja associado aos termos correspondentes e, através do motor de busca (Figura 9).

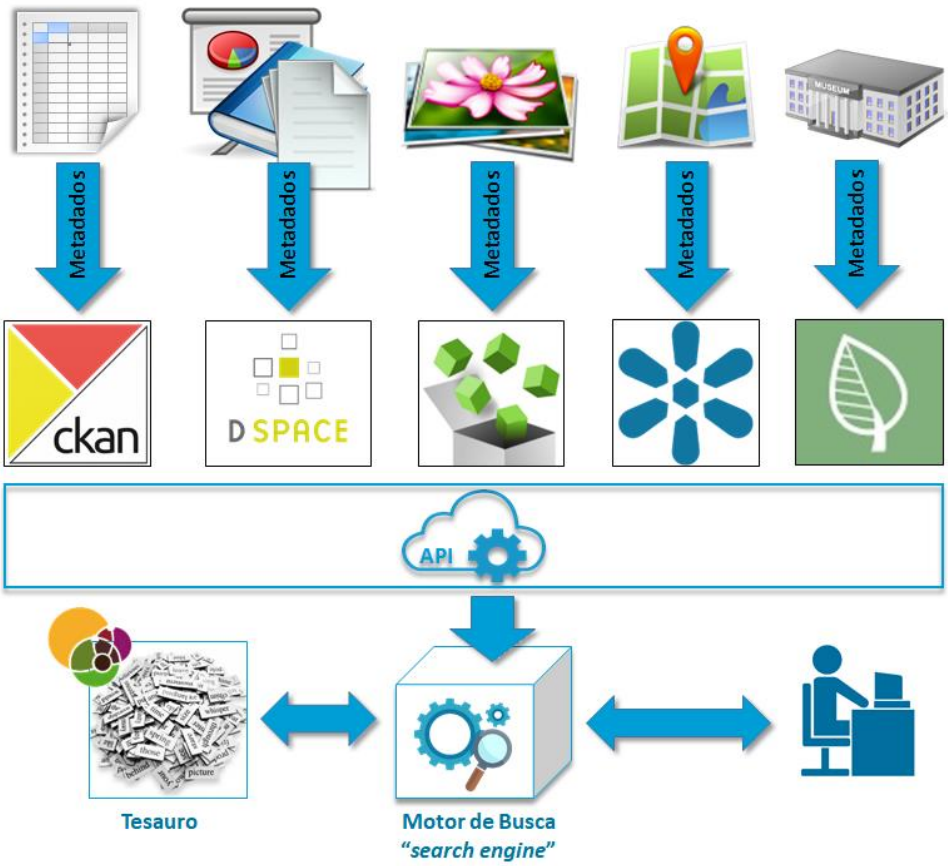

Figura 9 - Esquema completo da arquitetura proposta (DALCIN, 2016)

Acreditamos que o modelo descrito na figura 9 pode ser aplicado em diferentes institutos de pesquisa que lidam com recursos de informação heterogêneos e que desejam publicar estes recursos de informação *on-line* de forma aberta. A arquitetura proposta adota ferramentas gratuitas, de código aberto, e de uso largamente difundido por demais instituições de pesquisa ao redor do mundo formando uma comunidade robusta de suporte, facilitando sua adoção e manutenção.

CONCLUSÕES

A adoção de um conjunto heterogêneo de ferramentas livres, gratuitas e de código aberto para criação de repositórios de recursos de informação deu celeridade com baixo custo para a implementação de um portal de dados institucional. Entretanto, a busca por recursos dispersos nos diferentes repositórios demandou o desenvolvimento de ferramentas específicas para este fim, assim como a definição de um conjunto de vocabulários comuns às diferentes ferramentas visando a classificação dos mesmos nos diferentes contextos. Contudo, apesar desta limitação, o portal de dados causou um impacto positivo na instituição, que pôde avançar em Políticas do Governo Federal relacionadas com Dados Abertos e, mais recentemente, Ciência Aberta.

AGRADECIMENTOS

A Coordenação de Tecnologia da Informação e Comunicação (CTIC) do JBRJ, ao *SiBBr – Sistema de Informação sobre a Biodiversidade Brasileira* e ao MCTIC pelo financiamento concedido. Rafaela Campostrini Forzza é Bolsista de Produtividade do CNPQ e Cientista do Nosso estado da FAPERJ.

Artigo recebido em 22/06/2019 e aprovado em 25/10/2019.

REFERÊNCIAS

CNCFLORA. *Home.* Disponível em: http://cncflora.jbrj.gov.br/portal. Acesso em: 18 jun. 2019.

CORTI, P.; BARTOLI, F.; FABIANI, A.; GIOVANDO, C.; KRALIDIS, A. T.; TZOTSOS, A. *GeoNode:* an open source framework to build spatial data infrastructures (No. e27534v1). [S.l.]: PeerJ Preprints, 2019.

DALCIN, E. "Uma arquitetura para publicação de informações sobre biodiversidade." *In:* BIODIVERSIDADE, dados e metadados. 17 set. 2016. Disponível em: http://eduardo.dalc.in/uma-arquitetura-para-publicao-de-informaes-sobre-biodiversidade/. Acesso em: 18 jun. 2019.

DALCIN, E.; LANCELLOTTI, L. L. O.; ROCHA, M. S.; RAMOS, D. R. M.; RABELO, C. L.; SILVA JUNIOR, C. M. da. *Plano de dados abertos 2017 - 2018.* 2017. Disponível em: http://dspace.jbrj.gov.br/jspui/handle/doc/73. Acesso em: 20 out. 2019.

DIVYA, M. S.; GOYAL, S. K. ElasticSearch: an advanced and quick search technique to handle voluminous data. *Compusoft*, v. 2, n. 6, 171, 2013.

FEGRAUS, E.; ANDELMAN, S.; JONES, M.; SCHILDHAUER, M. Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation. *Bulletin of the Ecological Society of America*, v. 86, n. 3, p. 158-168, 2005.

GADELHA, L.; GUIMARÃES, P.; MOURA, A.; DRUCKER, D.; DALCIN, E.; GALL, G.; LEO, W. SiBBr: uma infraestrutura para coleta, integração e análise de dados sobre a biodiversidade brasileira. *In:* BRESCI BRAZILIAN E-SCIENCE WORKSHOP, 8., 2014. *Proceedings* [..]. [*S.l.: s.n.*], 2014.

IBICT. *DSPACE.* Disponível em: http://www.ibict.br/tecnologias-para-informacao/dspace. Acesso em: 21 jun. 2019.

INDE. *Catálogo de metadados.* Disponível em: http://www.inde.gov.br/geo-servicos/catalogo-de-metadados. Acesso em: 18 jun. 2019.

INTERFACE. *In:* WIKIPÉDIA: a enciclopédia livre. Wikimedia, 2019. Disponível em: https://pt.wikipedia.org/wiki/Interface_de_programa%C3%A7%C3%A3o_de_aplica%C3%A7%C3%B5es. Acesso em: 18 out. 2019.

JABOT2. *Home.* Disponível em: http://aplicacoes.jbrj.gov.br/jabot/v2/consulta.ph. Acesso em: 18 jun 2019.

JBRJ. *Quem somos.* Disponível em: http://jbrj.gov.br/institucional/quem-somos. Acesso em 18 out 2019.

JIGNESH, A.; VIRAL, A. Developing digital archive of documents, images and videos using digital assets management system resource Space. *International Journal of*

Information Library and Society, n. 1, p. 44-53, 2012.

MARTINS, A.; NUNES, M. B.; RODRIGUES, E. *Repositórios de informação e ambientes de aprendizagem:* criação de espaços virtuais para a promoção da literacia e da responsabilidade social. 2008. (Newsletter da Rede de Bibliotecas Escolares, n. 3).

PEREIRA, F. A.; KRZYZANOWSKI, R. F; MORAIS, T. F.; CALHERANI, J. A importância da prática de indexação para a recuperação da informação: relato da BV-FAPESP. *Revista Brasileira de Biblioteconomia e Documentação,* São Paulo, v. 11, n. esp., p. 374-390, 2015.

SAYÃO, L. F.; SALES, L. F. *Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores.* Rio de Janeiro: Centro de Informações Nucleares (CIN); Instituto de Engenharia Nuclear (IEN), 2015.

SILVA, E. da; *et al.* Jabot-sistema de gerenciamento de coleções botânicas: a experiência de uma década de desenvolvimento e avanços. *Rodriguésia*, v. 68, n. 2, 2017.

SMITH, M. *et al. DSpace:* an open source dynamic digital repository. 2003.

TANSLEY, R. *et al.* The DSpace institutional digital repository system: current functionality. *In:* ACM/IEEE-CS JOINT CONFERENCE ON DIGITAL LIBRARIES, 3., 2003. *Proceedings* […]. [*S.l.*]: IEEE Computer Society, 2003. p. 87-97.

WIECZOREK, John *et al.* Darwin core: an evolving community-developed biodiversity data standard. *PLoS ONE* 7, n. 1, 2012.

WINN, J. *Open data and the academy:* an evaluation of CKAN for research data management. [*S.l.: s.n.*], 2013.