# LAGUNA: INFRAESTRUTURA INFORMACIONAL ABERTA E LAGO DE DADOS

Patricia da Silva Neubert
Janinne Barcelos
Fabio Lorensi do Canto
Priscila Machado Borges Sena

## INTRODUÇÃO

O projeto Laguna é uma iniciativa coordenada pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) em colaboração com diversas instituições federais brasileiras. Seu objetivo principal é desenvolver uma Infraestrutura Informacional Aberta (IIA) para consolidar e disponibilizar dados do ecossistema brasileiro de Ciência, Tecnologia e Inovação (CT&I), baseada na arquitetura de um lago de dados (*data lake*) (Carvalho Segundo *et al*., 2023a e 2023b).

Um *data lake* é um repositório centralizado que armazena dados em seu formato bruto, sem restrições de estrutura ou tipo. Essa abordagem permite a integração de dados de diversas fontes, sejam elas estruturadas (como tabelas de bancos de dados), semiestruturadas (como arquivos JSON ou XML), ou não estruturadas (como e-mails, documentos, imagens, áudio e vídeo) (Alserafi *et al*, 2016; Khine; Wang, 2018; O´Leary, 2014). Diferente dos sistemas de armazenamento tradicionais que requerem uma estrutura rígida de dados, um lago de dados mantém os dados em seu estado original até que sejam necessários para análises ou processamento, permitindo maior flexibilidade para o uso de tecnologias avançadas, como aprendizado de máquina e inteligência artificial (Derakhshannia *et al.*, 2020).

Entre as vantagens do lago de dados está a capacidade de coletar e atualizar dados diretamente em seu formato nativo nas fontes selecionadas, eliminando etapas preliminares de padronização, o que reduz custos e agiliza o processamento (Fang, 2015). Além disso, o lago de

dados permite que grandes volumes de informações sejam armazenados de maneira econômica, utilizando armazenamento em nuvem escalável, como o *Amazon Web Services* (AWS), *Google Cloud* ou *Microsoft Azure*, que oferecem suporte a petabytes de dados com baixo custo. Isso facilita a realização de análises extensivas e o desenvolvimento de modelos de aprendizado de máquina, proporcionando insights valiosos para a tomada de decisões estratégicas (Fadnis, 2024).

O Laguna foi inicialmente projetado para atender a duas frentes de serviços estratégicos do Ibict: o BrCris (*Current Research Information System do Brasil*) e a construção de repositórios de dados de pesquisa, uma demanda crescente de instituições de ensino e pesquisa no país. O BrCris (Ibict, 2023) integra informações de repositórios e bases de dados nacionais e internacionais, como a Plataforma Lattes, a Plataforma Sucupira, a Wikidata e a CrossRef, entre outras, criando uma infraestrutura unificada que facilita a gestão e a análise da produção científica brasileira. Paralelamente, a construção de repositórios de dados de pesquisa visa oferecer suporte às instituições que necessitam armazenar, organizar e disseminar dados científicos de maneira eficiente, promovendo a transparência e o acesso aberto conforme os princípios FAIR (*Findable, Accessible, Interoperable, Reusable*).

A fragmentação dos registros de atividades científicas em diferentes sistemas de informação representa um desafio significativo para a obtenção de um panorama claro e abrangente do cenário de Ciência, Tecnologia e Inovação (CT&I). Muitos desses sistemas

operam de forma isolada, com baixa interoperabilidade, o que dificulta a integração e o uso eficiente dos dados para análises robustas e para a formulação de políticas públicas baseadas em evidências. A falta de integração e a existência de barreiras à interoperabilidade tornam o uso desses dados limitado, prejudicando sua utilidade como fontes confiáveis de informação sobre o desenvolvimento científico e tecnológico de um país.

O Laguna visa superar essas limitações adotando uma abordagem de lago de dados, que centraliza dados provenientes de diversas fontes, permitindo seu tratamento, análise e reutilização para múltiplos propósitos (Ibict, 2023). Ao consolidar esses dados em um único repositório, o Laguna facilita a criação de indicadores e relatórios detalhados sobre a produção científica brasileira, essencial para o desenvolvimento e a avaliação de políticas públicas eficazes em CT&I. Além disso, ao promover a convergência de dados e informações, o Laguna se alinha com o movimento de Ciência Aberta, ampliando o acesso e a transparência dos dados científicos. Para tanto, o Laguna baseia-se na criação de uma Infraestrutura Informacional Aberta (IIA).

## INFRAESTRUTURA INFORMACIONAL ABERTA (IIA) E DO LAGO DE DADOS

O desenvolvimento da Infraestrutura Informacional Aberta (IIA) e do lago de dados no âmbito do Projeto Laguna conta com uma equipe multidisciplinar responsável por diversos procedimentos técnicos e operacionais. Esses procedimentos envolvem a aplicação de métodos

computacionais avançados para a coleta, tratamento, organização, análise e disponibilização de dados sobre as atividades científicas, tecnológicas e de inovação no Brasil.

Dado que os registros das atividades científicas brasileiras estão fragmentados em vários sistemas públicos e institucionais, como as plataformas Lattes e Sucupira, é necessário um esforço considerável para coletar esses dados. Além desses sistemas, são amplamente utilizados mecanismos de busca, bases de dados e diretórios que agregam registros de produções científicas brasileiras em diversos formatos e tipos. Entre essas fontes, destacam-se a Biblioteca Digital de Teses e Dissertações (BDTD), CrossRef, *Directory of Open Access Journals* (DOAJ), Sistema Regional de Información en Línea para Revistas de América Latina, el Caribe, España y Portugal (Latindex), *Google Scholar Metrics*, *OpenAlex*, *OpenCitations*, OpenAIRE, Portal ISSN e Wikidata.

Para alimentar o lago de dados, é fundamental que essas fontes de informação ofereçam os dados em formatos que permitam exportação, seja por meio de acesso aberto ou por extração de dados a partir de acesso restrito, utilizando VPN institucional ou mediante acordos específicos. Em todos os casos, é essencial agregar valor aos dados integrados, garantir a continuidade na atualização da IIA e promover seu uso nos serviços desenvolvidos a partir dessa integração. Ademais, as diferentes fontes e conjuntos de dados requerem da equipe estratégias variadas para estudo e tratamento, incluindo o desenvolvimento de soluções próprias. Nesse contexto, o projeto tem dado ênfase ao uso de fontes abertas e acessíveis, como o OpenAlex (Neubert *et al*., 2024).

Uma vez coletados, os dados passam por diversos métodos de tratamento e processamento para maximizar seu valor (Fang, 2015; Giebler *et al.*, 2019). Esse processamento é dividido em várias etapas: (a) seleção e separação dos dados relevantes; (b) transformação e integração para assegurar compatibilidade e conectividade entre diferentes fontes; (c) organização, classificação e indexação para facilitar a recuperação eficiente; e (d) recuperação e visualização, visando disponibilizar os dados de maneira clara e acessível para os usuários finais (Carvalho Segundo *et al.*, 2023b). Tais métodos são descritos no Quadro 1, que destaca as práticas e as técnicas adotadas para garantir a qualidade, a integridade e a interoperabilidade dos dados integrados no sistema.

Quadro 1 - Descrição dos procedimentos para criação da IAA e lago de dados

| **Procedimento** | **Descrição** |
|---|---|
| Seleção das fontes dos dados | Identificação, estudo e seleção das fontes sobre atividade em CT&I brasileira, nas quais são priorizados os repositórios e bases de dados de CT&I que atendam total ou parcialmente os princípios FAIR. |
| Coleta dos dados | Os dados serão coletados por meio de API's públicas ou ferramentas de busca e extração disponíveis em cada uma das fontes selecionadas; se necessário, são utilizadas ferramentas de extração de dados web (web scraping tools) existentes, ou desenvolvidas especificamente para o projeto. |

| Procedimento | Descrição |
|---|---|
| Tratamento dos dados | Após a coleta, é realizada a seleção e a separação dos dados, por meio de filtragens e categorizações. É por meio desta etapa que as informações auxiliares (a chamada informação de overhead) são eliminadas, e os arquivos coletados são desmembrados entre os diversos tipos de entidades descritas em seu conteúdo (também chamado de payload). |
| Integração | Os dados já desmembrados, classificados e categorizados, são adaptados e validados, formando relações com registros de outras fontes. Um registro coletado de uma fonte A tem um atributo comum com o registro coletado da fonte B, podendo ser estabelecida uma vinculação entre ambas, com um determinado grau de confiabilidade. Os demais atributos dos registros podem ser mesclados de forma a resultar em um só registro enriquecido, eliminando-se réplicas. Um esquema de validação pode ser criado de modo a se descartar registros malformados, redundantes, inconsistentes ou ambíguos. |
| Estruturação | Os registros são organizados, classificados e indexados, de acordo com os seus atributos. As classificações servem de base para construção de interfaces de busca, webservices e dashboards de visualização. |
| Disponibilização | Com base nos dados tratados, podem ser construídas novas métricas e indicadores, |

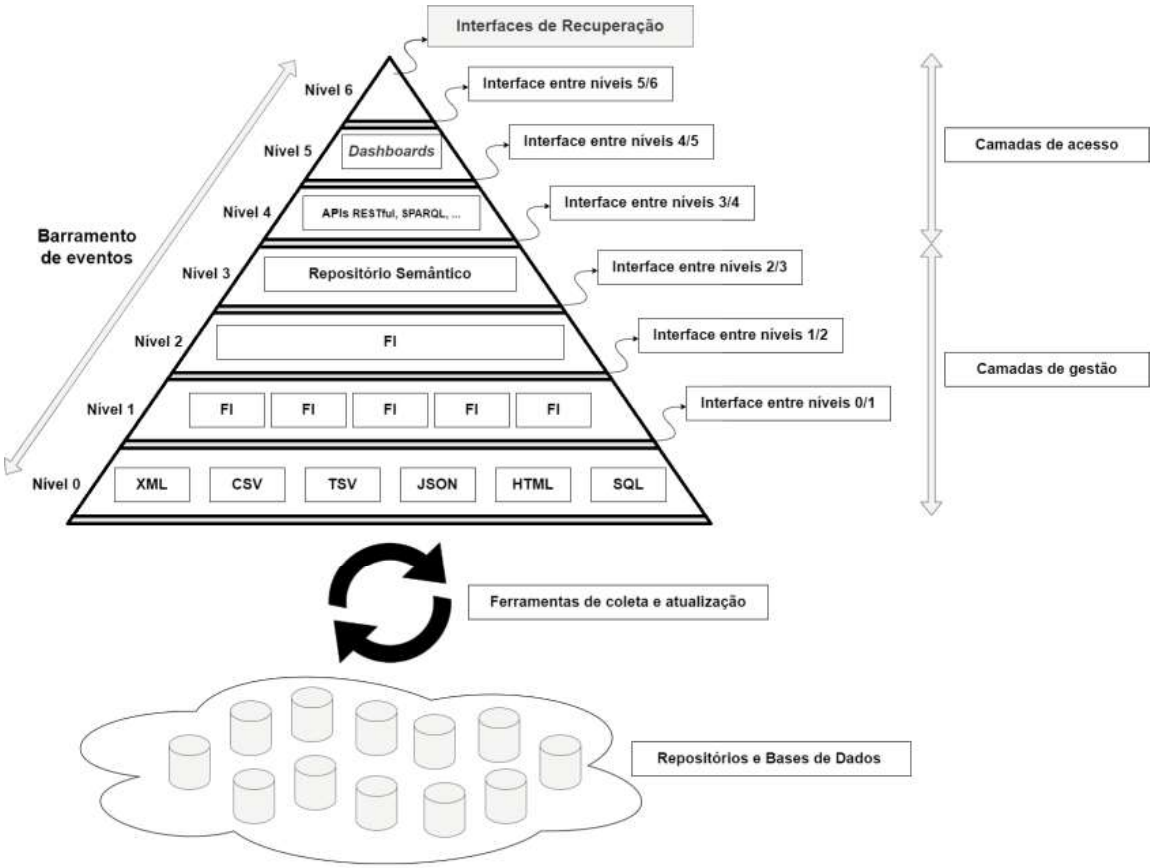
| Procedimento | Descrição |
|---|---|
| | adaptados à produção nacional. Para visualização, serão utilizadas ferramentas de exibição de redes de colaboração, de dados geoespaciais, de séries temporais, de esquemas dinâmicos de tabulação, entre outros. |

Fonte: adaptado de Carvalho Segundo (2023).

Após a obtenção dos dados a partir das fontes de informação externas, todos os procedimentos são realizados por meio de interfaces de processamento organizadas em seis níveis, conforme ilustrado na estrutura de processamento do lago de dados na Figura 1.

No Nível 0 da pirâmide, encontram-se os conjuntos de dados na forma em que foram coletados, nos mais variados formatos (como XML, CSV, entre outros). A interface entre os Níveis 0 e 1 transforma esses dados em Formato Intermediário (FI). Desta forma, no Nível 0 há uma normalização estrutural de representação de dados. No entanto, ainda podem existir redundâncias e falta de normalização ou validação quanto ao preenchimento dos campos mapeados. A interface entre os Níveis 1 e 2 realiza a remoção das redundâncias por meio de ações de deduplicação, mantendo a representação via FI. Nesta etapa, também são realizadas ações de normalização e validação dos campos, garantindo a consistência e precisão dos dados à medida que avançam para o próximo nível.

Figura 1 - Estrutura de processamento de dados em níveis do lago de dados




Fonte: Carvalho Segundo (2023).

Já a interface entre os Níveis 2 e 3 exerce a transformação do FI para um padrão de *dados ligados*, onde todos os dados são armazenados no formato de Triplas, na configuração [*sujeito*, *predicado*, *objeto*]. Nesse formato, o *sujeito* é sempre um identificador, o *predicado* é uma propriedade e o *objeto* pode ser um valor literal ou um novo identificador. No nível 3, a representação por meio de Triplas recebe um tratamento fino de representação das propriedades e preenchimento dos valores dos campos,

que são balizados via vocabulários semânticos (Hatch; Brown, 1995).

Por fim, nos Níveis 4, 5 e 6, são desenvolvidas, respectivamente: Interfaces de Programação de Aplicações (APIs) para o consumo dos dados tratados, que permitem o acesso estruturado e seguro aos dados por diferentes aplicações e sistemas; painéis de indicadores (*dashboards*), que fornecem uma interface visual para a visualização e monitoramento dos dados processados, facilitando a análise e a tomada de decisões com base em *insights* e métricas derivadas dos dados; e interfaces de recuperação da informação agregada, projetadas para permitir buscas complexas, consultas avançadas e o cruzamento eficiente de grandes volumes de dados de diferentes fontes.

## ACESSO AO LAGO DE DADOS

Na primeira fase do projeto, o acesso ao lago de dados está limitado aos pesquisadores do Laguna e de outros projetos mantidos pelo Ibict, tais como o BrCris, o OasisBR, a Biblioteca Digital Brasileira de Teses e Dissertações (BDTD), o Diretório das revistas científicas eletrônicas brasileiras (Miguilim) e o indicador persistente *Decentralized Archival Resource Key* (dARK). Estão sendo realizadas diversas atividades de integração entre esses projetos, visando a testar metodologias de processamento e de análise a partir de conjuntos de dados relevantes em uma perspectiva nacional.

Nesse contexto, foi desenvolvido em conjunto com o projeto BrCris um modelo de integração de dados das

Plataformas Lattes e OpenAlex por meio de identificadores persistentes de publicações (DOI) e de autores (ORCID), bem como do registro ISSN de periódicos. Esse modelo possibilitou o enriquecimento dos registros disponibilizados no BrCris, mantendo a origem dos dados do Lattes (fonte principal), mas agregando dados do modelo do OpenAlex (coautoria, fontes das publicações, vinculação institucional) enquanto fonte secundária.

Outra atividade de integração foi realizada com o projeto Miguilim, diretório de revistas científicas brasileiras. Os dados das revistas foram cruzados com diversas bases de dados de periódicos, visando a extração de dados relacionados às publicações nacionais. Foram identificados aspectos como indexação, identificadores de impacto, de produtividade e de citação, classificação em áreas do conhecimento, entre outros.

Além disso, a equipe do Laguna tem participado de reuniões técnicas do Ibict com outras instituições de ensino e pesquisa do país, com a finalidade de estabelecer colaborações nas áreas de pesquisa e avaliação científica, indicadores, repositórios e Ciência Aberta. Essas colaborações podem ser informais, visando a extração de conjuntos de dados do Laguna para utilização em pesquisas. Podem ainda ser formalizadas por meio de protocolos de intenção e acordos de cooperação técnica.

Está prevista também no projeto a disponibilização de conjuntos de dados tratados e enriquecidos no repositório de dados de pesquisa do Ibict, o Alea. Serão selecionados conjuntos de dados relativos à produção brasileira em CT&I, especialmente os conjuntos extraídos de fontes abrangentes tais como OpenAlex, Lattes,

Plataforma Sucupira, OpenAIRE, DOAJ, entre outras. A disponibilização desses conjuntos terá a finalidade de incentivar a utilização dos dados por pesquisadores brasileiros.

## RESULTADOS ESPERADOS

O principal resultado esperado com a execução desse projeto é o desenvolvimento de um amplo repositório de dados abertos relativos à CT&I brasileira, com infraestrutura de lago de dados, armazenamento e processamento distribuídos. Essa infraestrutura permitirá o tratamento e a análise de conjuntos de dados, gerando novos conhecimentos e aplicações estratégicas. O repositório será composto por diversas fontes integradas e padronizadas, possibilitando consultas e análises com ferramentas de visualização, além de exportações por meio de APIs em formatos padronizados e certificados.

As informações serão classificadas e padronizadas, permitindo sua importação por outras ferramentas de análise ou repositórios locais. Dessa forma, o projeto busca consolidar uma base de dados de alta qualidade, integrada e acessível, promovendo o uso eficiente dos dados para o avanço da CT&I no Brasil.

Sob o aspecto tecnológico, espera-se que a equipe adquira *know-how* em tecnologias computacionais de alto desempenho, como inteligência artificial, big data e computação em nuvem. Esse conhecimento será estratégico para aplicação em outros contextos nacionais, ampliando a capacidade técnica e operacional em projetos futuros.

O projeto também visa gerar impactos sociais em médio e longo prazo, principalmente por meio do desenvolvimento de uma infraestrutura informacional robusta para CT&I no Brasil. Essa infraestrutura otimizará o aproveitamento dos resultados de pesquisa e subsidiará a formulação de políticas públicas em áreas estratégicas para o desenvolvimento nacional.

Os impactos econômicos esperados incluem a otimização da distribuição de recursos públicos — humanos e materiais — em projetos de CT&I, o aumento da visibilidade e do aproveitamento de resultados de pesquisa, bem como a redução de custos com assinaturas de serviços informacionais oferecidos por empresas comerciais.

Por fim, a partir dos indicadores gerados pelo lago de dados e do tratamento e agregação dos dados, prevê-se o desenvolvimento de produtos e serviços de apoio à pesquisa em CT&I brasileira, como sistemas de recomendação e identificação. Esses produtos e serviços deverão potencializar a inovação e o acesso a dados estratégicos no país.

## REFERÊNCIAS

ALSERAFI, Ayman *et al*. Towards Information Profiling: Data Lake Content Metadata Management. *In:* INTERNATIONAL CONFERENCE ON DATA MINING WORKSHOPS (ICDMW), 16., 2016. **Anais** [...]. Barcelona: IEEE, 2016. p. 178-185. DOI: 10.1109/ICDMW.2016.0033.

CARVALHO SEGUNDO, Washington Luís Ribeiro de *et al*. Fontes de dados sobre periódicos científicos: a proposta do projeto

laguna de dados. *In:* CONFERÊNCIA LUSÓFONA DE CIÊNCIA ABERTA, 14., 2023. **Anais** [...]. Natal, UFRN, 2023c.

CARVALHO SEGUNDO, Washington Luís Ribeiro de *et al*. Tratamento de dados FAIR no Projeto Laguna. *In:* CONFERÊNCIA LUSÓFONA DE CIÊNCIA ABERTA, 14., 2023. **Anais** [...]. Natal, UFRN, 2023a.

CARVALHO SEGUNDO, Washington Luís Ribeiro de. Construindo uma infraestrutura aberta de dados de pesquisa no Brasil. *In*: ENCONTRO DA REDE BRASILEIRA DE REPOSITÓRIOS DIGITAIS, 1., 2022. **Anais** [...]. Rio de Janeiro: Fiocruz/Icict; Ibict, 2022.

DERAKHSHANNIA, Marzieh *et al*. Data Lake Governance: towards a systemic and natural ecosystem analogy. **Future Internet**, Basel, v. 12, n. 8, p. 1-16, 27 jul. 2020. DOI: 10.3390/fi12080126.

FADNIS, Bhushan. The Data Lakes: a leap forward future of data warehousing. **International Journal of Innovative Science And Research Technology**, Jaipur, p. 3063-3067, 15 jun. 2024. DOI: 10.38124/ijisrt/IJISRT24MAY2158.

FANG, Huang. Managing data lakes in the big data era: What's a data lake and why has it became popular in data management ecosystem. *In*: INTERNATIONAL CONFERENCE ON DATA MINING WORKSHOPS (IEEE), 2015. **Anais** [...]. Shenyang, China: IEEE computer society, 2015.

GIEBLER, Corinna *et al*. Leveraging the Data Lake: current state and challenges. **Lecture Notes In Computer Science**, Cham, p. 179-188, 2019. DOI: 10.1007/978-3-030-27520-4_13.

HATCH, Evelyn; BROWN, Cheryl. **Vocabulary, semantics, and language education**. New York: Cambridge University Press,1995.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA (IBICT). **Ecossistema de informação da pesquisa científica brasileira, BrCris**. Brasília, Ibict, 2023. Disponível em: https://brcris.ibict.br. Acesso em: 09 ago. 2024.

KHINE, Pwint Phyu; WANG, Zhao Shun. Data lake: a new ideology in big data era. *In*: ANNUAL INTERNATIONAL CONFERENCE ON WIRELESS COMMUNICATION AND SENSOR NETWORK (WCSN), 4., 2017. **ITM Web Of Conferences**, Wuhan, v. 17, p. 1-10, 2018. DOI: 10.1051/itmconf/20181703025.

NEUBERT, Patricia *et al*. OpenAlex como fonte de dados para sistemas nacionais de informação científica: a experiência do projeto laguna. *In*: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA - WIDAT, 7., 2024. **Anais** [...]. Porto Velho: Ibict, 2024. DOI: 10.22477/vii.widat.184.

O'LEARY, Daniel E. Embedding AI and Crowdsourcing in the Big Data Lake. **IEEE Intelligent Systems**, [s. l.], v. 29, n. 5, p. 70-73, set. 2014. 2014.82. DOI: 10.1109/MIS.2014.82.

Como citar este capítulo

NEUBERT, Patricia da Silva; BARCELOS, Janinne; CANTO, Fabio Lorensi do; SENA, Priscila Machado Borges. Laguna: Infraestrutura Informacional Aberta e lago de dados. *In*: AMARO, Bianca; CAMPOS, Fhillipe de Freitas; BARCELOS, Janinne. (org.). **Infraestruturas de Ciência e de Acesso Aberto no Brasil**: iniciativas do Instituto Brasileiro de Informação em Ciência e Tecnologia. Brasília, DF: Editora Ibict, 2025. Cap. 16, p. 263-277. DOI: 10.22477/9788570132543.cap16