OPEN KNOWLEDGE BRASIL

Querido Diário

Proposta de cooperação acadêmica

Apresentação

O Querido Diário[1] (QD) é um projeto de código aberto[2] mantido pela Open Knowledge Brasil[3] (OKBR) cujo objetivo é acessar, libertar e integrar informações contidas nos diários oficiais dos 5.570 municípios brasileiros, para aprimorar o controle social da administração pública e aproximar os cidadãos das ações de seus governantes.

Este documento tem por objetivo **abrir um diálogo entre o Querido Diário e Faculdades, Universidades, Centros Técnicos e Centros de Pesquisa** buscando facilitar cooperações que levem estas instituições a utilizar, de maneira cada vez mais presente em seu cotidiano, as informações disponíveis dentro do conteúdo dos Diários Oficiais do Brasil, através do ecossistema de ferramentas do QD.

Assim, este documento apresenta a fase atual de desenvolvimento do projeto, possíveis usos dos dados disponibilizados pelo QD, alguns desafios a serem superados no campo de Ciência de Dados e a descrição dos modelos de cooperação propostos.

Há muito espaço para aperfeiçoamento do QD e, por ser um projeto de código aberto, acreditamos que o contato com a academia, além de compor a tão rica cultura de colaboração com a sociedade civil, também fortalece as partes mutuamente, oferecendo desafios a serem vencidos por estudantes ou pesquisadores que buscam **aplicar em projetos reais e interessantes os conhecimentos adquiridos durante suas formações** ao mesmo tempo que aprimora o potencial de consolidar o projeto como uma importante ferramenta sócio-técnica brasileira.

[1] Querido Diário
[2] Repositórios de código do Querido Diário
[3] Open Knowledge Brasil

Sumário

Como está o projeto atualmente?

A ferramenta de buscas do Querido Diário (QD), lançada em junho de 2021, disponibiliza a consulta à base de conteúdo textual dos diários oficiais, trazendo como resposta um trecho do texto do diário onde a consulta foi encontrada, alguns metadados sobre a publicação em que o texto foi encontrado, a opção de baixar a versão original (PDF) e a versão em formato aberto (TXT), que não possui formato estruturado. Os dados são obtidos via técnicas de raspagem (*web scraping*) diariamente e disponibilizados para consultas via interface web[4] (para uso pontual, amigável) ou API pública[5] (para uso programático, robusto), exemplificadas nas Figuras 1 e 2.

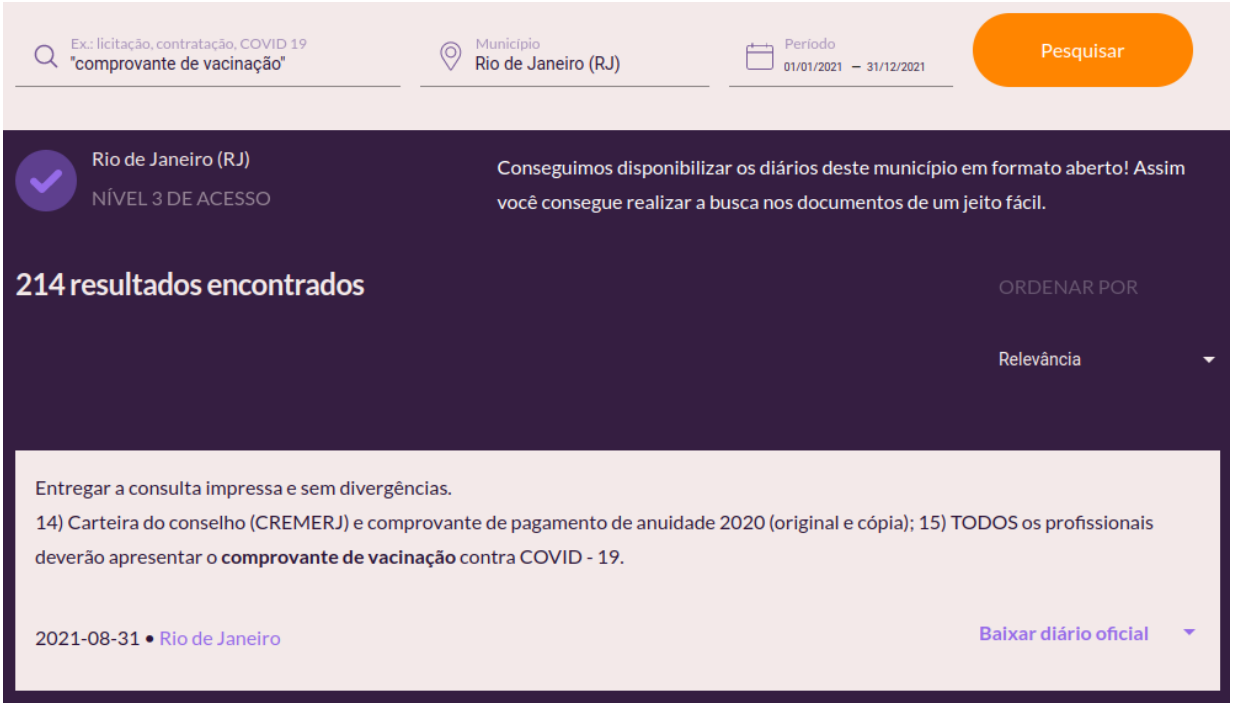

Figura 1: Exemplo de busca na interface web do Querido Diário

[4] Querido Diário
[5] Documentação da API do Querido Diário

Curl

```
curl -X GET "https://queridodiario.ok.org.br/api/gazettes/3304557?since=2021-01-01&until=2021-12-31&querystring=%22comprovante%20de%20vacina%C3%A7%C3%A3o%22&size=10&fragment_size=500&number_of_fragments=1&pre_tags=%3Cb%3E&post_tags=%3C%2Fb%3E&sort_by=relevance" -H "accept: application/json"
```

Request URL

```
https://queridodiario.ok.org.br/api/gazettes/3304557?since=2021-01-01&until=2021-12-31&querystring=%22comprovante%20de%20vacina%C3%A7%C3%A3o%22&size=10&fragment_size=500&number_of_fragments=1&pre_tags=%3Cb%3E&post_tags=%3C%2Fb%3E&sort_by=relevance
```

Server response

Code	Details
200	**Response body**


```
{
  "total_gazettes": 214,
  "gazettes": [
    {
      "territory_id": "3304557",
      "date": "2021-08-31",
      "url": "https://querido-diario.nyc3.cdn.digitaloceanspaces.com/3304557/2021-08-31/40bd3c4bcb246c2b71b96e27507711acef32db94",
      "territory_name": "Rio de Janeiro",
      "state_code": "RJ",
      "highlight_texts": [
        "Entregar a consulta impressa e sem divergências.\n14) Carteira do conselho (CREMERJ) e comprovante de pagamento de anuidade 2020 (original e cópia);\n15) TODOS os profissionais deverão apresentar o <b>comprovante</b> <b>de</b> <b>vacinação</b> contra COVID - 19."
      ],
      "is_extra_edition": false,
      "file_raw_txt": "https://querido-diario.nyc3.cdn.digitaloceanspaces.com/3304557/2021-08-31/40bd3c4bcb246c2b71b96e27507711acef32db94.txt"
    },
    {
      "territory_id": "3304557",
      "date": "2021-06-14",
      "url": "https://querido-diario.nyc3.cdn.digitaloceanspaces.com/3304557/2021-06-14/0f426f9a6f2286f05ffccde9cf87ef06b6d7b8b6"
```


Download

Figura 2: Exemplo de busca na API do Querido Diário

Atualmente, o projeto está em sua segunda fase de desenvolvimento (junho/2021 a maio/2022). Nesta fase, conseguimos libertar cerca de 104 mil diários oficiais, integrando 27 cidades à plataforma, totalizando um volume de dados de aproximadamente 370 GB.

Também disponibilizamos, por meio da API, a possibilidade de filtro de resultados da busca que sejam relacionados à temática ambiental e o reconhecimento de termos

relevantes para a temática no texto. Estas funcionalidades são os primeiros frutos de uso de Ciência de Dados para enriquecimento da base de dados do projeto.

Além disso, também via API, é possível realizar uma consulta à base de dados da Receita Federal, facilitando o cruzamento de dados do QD com o de empresas cadastradas. Este é o primeiro passo para facilitar a integração do projeto com outras bases de dados públicas.

Outras informações sobre o funcionamento do QD podem ser conferidas em: https://queridodiario.ok.org.br/tecnologia.

Como os dados fornecidos pelo Querido Diário podem ser utilizados como insumos para análises?

Em seus primeiros passos, o projeto já proporcionou interessantes resultados registrados em matérias jornalísticas[6]. E no âmbito acadêmico, os dados do QD já foram utilizados por alunos da UFRN[7] e do IME-USP[8] para análise de publicações relacionadas à pandemia de COVID-19.

Estas análises demonstram algumas possibilidades de uso dos dados. Dentre as áreas do conhecimento que podem ser utilizadas para aproveitar os dados, destacamos:

- ➔ Análise de Dados (estruturados e não-estruturados);
- ➔ Visualização de Dados;
- ➔ Processamento de Linguagem Natural;
- ➔ Aprendizagem de Máquina;
- ➔ Gestão Pública.

Por pertencerem a um projeto de código aberto, todos os dados do QD podem ser acessados e consumidos para a utilização em análises e experimentações livremente.

[6] "Retrospectiva 2021: relembre as reportagens e outros casos de uso do Querido Diário"
[7] "Querido Diário é utilizado em estudo da UFRN sobre gastos públicos com a Covid-19"
[8] "OKBR e IME-USP encerram semestre de parceria com muitos avanços para o Querido Diário"

Que tipos de estudos podem ser realizados para melhorar o projeto?

O Querido Diário é fundamentado pela ideia de suprir falhas de abertura de dados na esfera municipal do poder executivo, onde não há padrão para publicação de informações e a grande quantidade de municípios torna a escala desta ideia ainda maior. Assim, o projeto atua obtendo as publicações em formato mais comum de publicação (formato para impressão, normalmente PDF), extraindo o conteúdo textual dos documentos e então extraindo informações a partir de um formato aberto (TXT). Essa abordagem faz com que o projeto já possa ser utilizado para realizar buscas e análises, como apresentado anteriormente.

Porém, para que o projeto consiga lidar com desafios mais complexos e forneça soluções cada vez mais úteis, avanços técnicos são necessários. O QD é terreno fértil para estudos que buscam solucionar problemas reais e atuais, principalmente (mas não exclusivamente) ao abordar tarefas de Processamento de Linguagem Natural em Português. Assim, elencamos algumas ideias:

- **Segmentação em seções**, como no estudo feito em disciplina de pós-graduação na UNICAMP[9] que buscou segmentar os Diários Oficiais com auxílio do modelo de linguagem BERTimbau[10]. Importante tarefa que tornaria possível encontrar conteúdo específico em seções como “Licitações Públicas”;
- **Sumarização de conteúdo**. Tornaria tarefas de leitura mais objetivas e facilita a indexação e categorização de diários;
- **Busca semântica**. Permitiria que buscas realizadas dependam menos do conhecimento técnico para escolha de ótimas palavras-chave;
- **Recomendação de conteúdo**. Para além da “busca semântica”, tornaria possível identificar documentos semanticamente próximos;
- **Reconhecimento de entidades nomeadas**. Enriqueceria os dados disponibilizados de forma raramente encontrada em publicações de diários.

[9] “Segmentation of Legal Documents: unsupervised approach with BERT embeddings”
[10] BERTimbau - BERT para Português

Qual o formato da cooperação?

A cooperação pode ser construída de múltiplas formas. Seja por meio de projetos de disciplinas, pesquisas de graduação, pós-graduação, mestrado ou doutorado, projetos de extensão ou grupos de estudo que visam a produção de conteúdos que possam reafirmar a importância da libertação dos dados presentes nos Diários Oficiais, sejam artigos e publicações científicas ou projetos de fim de curso. Caso seja do interesse das responsáveis, a OKBR buscará auxiliar a atividade por alguns meios:

- ➔ Enumerar um ou mais desafios específicos para direcionamento, tendo em vista o contexto no qual será(ão) inserido(os);
- ➔ Acolhimento de sugestões e dúvidas que sejam identificadas durante o desenvolvimento da atividade através do servidor no Discord[11] e dos repositórios no GitHub mantidos pela OKBR para interação com a comunidade de código aberto;
- ➔ Incentivo e apoio na divulgação dos resultados obtidos.

Para a OKBR e o projeto Querido Diário, qualquer resultado – mesmo os não atingidos por identificação de novas barreiras não previstas, que podem, inclusive, advir de limitações da ferramenta – tem uma enorme validade para a evolução da iniciativa.

Contato

Para dar seguimento à cooperação, preencha este formulário com informações básicas sobre o projeto que utilizará o Querido Diário.

O uso do Querido Diário é livre, e contribuições espontâneas são bem-vindas, não estando condicionadas à aprovação por meio de contato. Esta chamada tem como propósito tornar mais próxima a relação do Querido Diário com grupos interessados, permitindo-nos conhecer o trabalho que será desenvolvido, assim como as pessoas envolvidas, e dando-nos a oportunidade de auxiliar no processo de desenvolvimento.

A chamada não possui prazo para encerramento. Ao receber sua resposta, retornaremos em cerca de 7 dias úteis propondo uma conversa para aprofundamento, caso necessário, ou indicando contrapartidas que possamos oferecer.

Aguardamos seu contato!

[11] Discord da OKBR