OPEN KNOWLEDGE BRASIL

Querido Diário

1º Relatório Técnico de Atividades

RESUMO

Este relatório apresenta um detalhamento dos aspectos técnicos desenvolvidos e implementados para o projeto Querido Diário entre setembro de 2020 e maio de 2021 pelo Programa Ciência de Dados para Inovação Cívica da Open Knowledge Brasil.

INTRODUÇÃO

Neste relatório, a OKBR apresenta todos os avanços obtidos no desenvolvimento do projeto Querido Diário (QD). O objetivo geral da ferramenta é libertar os dados contidos nos diários oficiais municipais no Brasil, que registram a implementação de políticas públicas na esfera local, a publicação de editais de compras públicas, as obrigações municipais contidas na Lei de Responsabilidade Fiscal (LRF), entre outras informações relevantes para a sociedade civil sobre a atuação dos governos locais. A inserção da iniciativa no contexto da chamada Empatía foi fundamental para garantir os mecanismos básicos e os algoritmos de inteligência artificial que sustentarão a plataforma.

Com dimensões continentais, o Brasil possui cerca de 5.570 municípios, e cada governo municipal é responsável por publicar seu diário oficial em meio eletrônico; na prática, portanto, existem muitos sistemas e formatos de divulgação de dados, o que torna muito difícil a extração, compilação e utilização destas informações. O Querido Diário aborda justamente esse problema e oferece uma fonte integrada de acesso e tratamento dos dados, com plataforma de visualização, interface para acesso programático (API) e sistema de busca aberto ao público. A visão do QD é facilitar o acesso de qualquer tipo de pessoa usuária a qualquer dado municipal.

Para sustentar sua missão, o Querido Diário conta com uma série de ferramentas tecnológicas e colaboradores que permitem com que seus mecanismos estejam disponíveis de forma aberta e gratuita. Temos, por exemplo, uma equipe de voluntários responsável por desenvolver os algoritmos a fim de coletar os diários oficiais (cada publicação costuma ter uma estrutura distinta e necessita de códigos personalizados), que posteriormente são adicionados a uma base pública e analisados utilizando técnicas

de processamento de linguagem natural. Neste relatório, descrevemos nosso progresso para implementar esse ciclo de libertação dos dados.

MATURIDADE DO PROJETO

Existem quatro fases para a implementação da infraestrutura do projeto Querido Diário. Cada etapa requer um grupo de algoritmos executando funções específicas e, portanto, seu desenvolvimento é heterogêneo.

O primeiro grupo é composto por raspadores de arquivos de diários oficiais, que são responsáveis pela coleta dos dados que farão parte do nosso banco de dados. As pessoas voluntárias que nos auxiliam nesta tarefa escrevem o código dos raspadores e validam os dados obtidos.

> *Atualmente, nossos programas são capazes de coletar diários de **2.229 municípios**, cerca de **40% do total** de municípios brasileiros. Avançamos, portanto, mais de **7 vezes** o número de cidades referentes à cobertura do nosso alvo de 300 municípios.*

O **segundo grupo** compõe a API do Querido Diário, que já está disponível e permite consultar atos oficiais de 12 capitais brasileiras (João Pessoa, Teresina, Boa Vista, Maceió, Salvador, Goiânia, Rio de Janeiro, Natal, Campo Grande, Florianópolis, Palmas e Manaus). A API pode ser acessada pelo público em geral, mas também serve para alimentar a plataforma online e o sistema de busca de conteúdo de diários, que compõem a fase quatro. Apesar da capacidade de raspar os dados de 2.229 municípios, decidimos partir de um conjunto significativo de 12 cidades para a primeira fase de testes da plataforma, sobretudo para garantir o funcionamento da infraestrutura de processamento e armazenamento. Nas próximas fases da iniciativa, dependendo do financiamento que obtivermos, vamos seguir aumentando a quantidade de municípios para consulta na plataforma.

O **terceiro grupo** de algoritmos sendo desenvolvido para o projeto é composto pelos programas que fazem a extração do conteúdo dos diários, processam este conteúdo e identificam informações relevantes para composição do banco de dados.

O **quarto e último grupo** são os algoritmos que compõem o mecanismo de busca, que foram prototipados e agora se encontram na fase de teste, e a plataforma de visualização disponível à pessoa usuária. Com estes algoritmos, permitimos que pessoas que não tenham nenhuma familiaridade com linguagem de programação ou processamento de dados sejam capazes de consumir o conteúdo do QD a partir de filtros simples como palavras-chave e localização. A plataforma conterá informações do Censo, que mapeia a cobertura dos diários oficiais pelo QD, os resultados de busca na API, um glossário de termos, uma seção de análises feitas pela equipe da OKBR e muitos outros elementos que maximizem o uso e disseminação do projeto.

Neste relatório, apresentamos também o **primeiro caso de uso da API** do Querido Diário. A jornalista Beatriz Farrugia analisou, através dos atos oficiais do município de Manaus, disponibilizados pela API, o papel da prefeitura no colapso do sistema de saúde do município devido à pandemia de COVID-19.

IMPLEMENTAÇÃO DOS ALGORITMOS E INFRAESTRUTURA

Coleta Automatizada dos Dados

A partir da implementação dos raspadores de arquivos feitos pela nossa comunidade, nós estruturamos uma **rotina diária de coleta de dados** que funciona de forma **automática e agendada**. Contamos com 12 capitais brasileiras sendo atualizadas diariamente até o momento. Dentro desta estrutura, nós temos controle de eficácia e recebemos alertas de sucesso ou falha na coleta de dados para cada município. Este programa está hospedado com o apoio da empresa *Zyte*, em sua infraestrutura otimizada para a raspagem de dados em massa e automática.

API Pública

Com o resultado da coleta automatizada, conseguimos enviar, via um pipeline, todos os arquivos para o serviço Digital Ocean Spaces. A API usa uma estrutura REST e é construída com recursos da FastAPI e Swagger, que permitem fácil acesso, rapidez e estabilidade nas consultas aos dados.

Através da API, é possível buscar documentos disponíveis utilizando combinações de filtros por município e por data, assim como também é possível utilizar o mecanismo de busca lexical no conteúdo dos diários. O sistema open source que permite buscar termos, indexar documentos e ranqueá-los foi desenvolvido pela empresa Elastic e se chama elasticsearch (ES). O ES constrói e mantém o índice de documentos, procura o termo no índice e retorna os documentos mais relevantes.

A API já está em funcionamento, ainda que com uma cobertura limitada de 12 capitais brasileiras que foram incluídas na coleta automatizada mencionada anteriormente. Como explicamos na seção sobre a maturidade do projeto, esta quantidade de cidades se manterá durante o período de testes de infraestrutura e até que tenhamos financiamento suficiente para fornecer mais municípios de maneira contínua. Atualmente, o acesso é gratuito para os membros de nossa comunidade, e esta mesma API servirá a plataforma de visualização e consulta.

O esforço de adicionar municípios à coleta automatizada e disponibilizá-los através da API é contínuo e constante, e mantemos o plano mencionado no relatório anterior de incluir cerca de 15 municípios até o lançamento da plataforma de visualização para o público geral. Com o lançamento da plataforma, esperamos despertar o interesse de um volume maior de colaboradores e financiadores, e com isso, acelerar o ritmo de disponibilização de municípios.

Extração e Processamento de Conteúdo

Para que os documentos encontrados pelos raspadores tenham seu conteúdo disponibilizado, nossos programas extraem e processam o texto e identificam informações relevantes para composição da base de dados.

O conteúdo dos diários é extraído através da ferramenta open source Apache Tika, desenvolvida pela Apache. Com o Tika, é possível extrair conteúdo de documentos de diversos formatos e convertê-los para TXT. Essa conversão é imprescindível para que seja possível utilizar o conteúdo dos diários em análises empíricas.

Para melhorar os resultados disponibilizados para as pessoas usuárias, há o trabalho de "fatiar" o conteúdo dos diários oficiais para melhor retratar as informações contidas nestes documentos. No momento, somos capazes de extrair conteúdo textual e

fazer buscas de informações usando regras lógicas simples (identificadores numéricos, cidades, períodos do ano).

Esse trabalho de fatiamento dos diários municipais requer, também, a identificação das seções que compõem um diário e os termos que identificam cada seção. Embora ainda não tenhamos implementado estas ferramentas no âmbito do primeiro *Minimum Viable Product* - o MVP #1, já concluímos o trabalho necessário para reconhecimento de seções relevantes, a saber:

Seção	Relevância
Atos normativos	Geram obrigações ou direitos aos cidadãos fruto de atividade do Poder Executivo (ou Legislativo) municipal.
Atos de pessoal	Dispõem sobre mudanças nas atividades dos funcionários municipais, contemplando ingresso em carreira pública, exoneração, concessão de afastamento temporário etc.
Contas públicas	Reportam as obrigações de dívida dos municípios e peça orçamentária em respeito à Lei de Responsabilidade Fiscal
Audiências públicas	Informam a população sobre audiências públicas referentes à atividade do governo municipal e implementação de políticas públicas locais.
Licitações	Funcionam como o principal instrumento de contratação de bens e serviços privados no Brasil pelo setor público.
Procedimentos disciplinares	Relatam os processos investigativos e disciplinares sobre conduta de servidores públicos ou contratantes do poder executivo municipal.
Procedimentos tributários	Informam os acórdãos entre o fisco municipal e os contribuintes em processos administrativos tributários e afetam a arrecadação do município.
Conselhos Municipais	Realizam a fiscalização de políticas públicas por parte da sociedade civil.

A divisão dos diários nestas seções pertence à segunda etapa do QD, a ser iniciada na segunda metade de 2021.

Mecanismo de Busca

Nas entregas passadas, estruturamos e testamos o mecanismo de busca do Querido Diário, que está descrito a seguir.

Um mecanismo de busca textual pode assumir diversos formatos. O mais comum deles é a correspondência integral entre termos de busca e termos presentes nos documentos alvo. Esses são os algoritmos de busca lexical, em que a presença do termo de busca no documento alvo determina o documento como resultado positivo da busca. Por exemplo, uma busca pela palavra "casa" retornará todos os documentos em que esta palavra estiver presente.

A grande vantagem destes instrumentos é a baixa complexidade e facilidade de implementação: nós conseguimos definir o que é a correspondência de termos (existência da mesma sequência de caracteres na consulta e no alvo) e a implementação consiste na varredura do documento, caracter por caracter, e marcação do documento quando a sequência de caracteres da busca for encontrada. A desvantagem destes algoritmos é que a busca é rígida e não permite a correspondência entre termos declinados em gênero, número ou grau.

> *Por exemplo, uma busca pela palavra "casarão" não retorna documentos que contenham somente a palavra "casa". Por esse motivo,* ***optamos por um sistema de busca duplo, com funcionalidade lexical e semântica****.*

A busca semântica consiste na correspondência semântica entre os termos de busca e o documento alvo. Essa busca é mais complexa que a versão lexical pois exige a avaliação da relação entre palavras (casa e casarão, por exemplo) e a conversão das palavras em vetores semânticos para execução da query de consulta. A desvantagem da busca semântica é o custo computacional, pois se exige uma etapa adicional de conversão de texto em vetor semântico (tanto dos termos de busca quanto do

documento alvo), e o custo temporal, com base na construção lenta dos vetores do corpo de documentos nos quais serão realizadas as buscas.

Para implementação deste sistema, nós precisamos de (i) uma plataforma de busca, (ii) um conversor de palavras em vetores, (iii) um índice de documentos (o alvo) e (iv) um mecanismo de ranqueamento dos documentos por relevância. Nossa tarefa, ao longo dos últimos meses, foi formular essa estrutura e testá-la.

Como explicado anteriormente, o sistema que escolhemos para buscar termos, indexar documentos e ranqueá-los se chama elasticsearch (ES). O ES constrói e mantém o índice de documentos, procura o termo no índice e retorna os documentos mais relevantes. Para a conversão dos termos em vetores semânticos, nós adotaremos o modelo de linguagem do Google, BERT, em sua versão em português, BERTimbau. Este algoritmo de conversão também está disponível publicamente e será introduzido como intermediário entre a busca e o retorno dos resultados.

A equipe do Querido Diário já testou este sistema em dois bancos de dados. O primeiro foi um banco de manchetes curtas de jornais brasileiros disponível no site de competições computacionais Kaggle. O segundo foi a amostra própria dos diários oficiais de 15 municípios e 2.300 páginas. O desempenho nos dados do Kaggle foi excelente, mas nos diários foi médio. Esse resultado, contudo, ainda foi positivo, pois a complexidade da estrutura dos diários oficiais indicava desempenho inicial pior do que o obtido.

Em especial, o resultado é encorajador, pois ainda temos um menu de melhorias diagnosticadas, entre elas: a quebra do conteúdo do diário em partes homogêneas (fatiamento descrito acima), o uso de versão mais completa do BERTimbau (apenas utilizamos a versão simples, mais leve, para teste), o treino do modelo de linguagem e a sumarização do conteúdo de modo a diminuir a complexidade dos vetores semânticos.

PLATAFORMA DE VISUALIZAÇÃO

Estágio de desenvolvimento

Planejada sob arquitetura *mobile first*, a aplicação é totalmente responsiva, possibilitando uma boa experiência de navegação tanto em computadores como em dispositivos móveis.

Figura 1 - Telas iniciais para computadores (esquerda) e celulares (direita).

Exibição dos graus de maturidade dos municípios e da plataforma

Gostaríamos de informar ao visitante do site o estágio de maturidade do projeto. Essa é uma forma de gerenciar as expectativas das pessoas usuárias e mantê-las engajadas na utilização da plataforma.

Atualmente, existem três graus de maturidade de um município na plataforma Querido Diário, a saber:

- Nível 0 - não possuímos acesso à fonte de publicação do diário oficial deste município (não sabemos onde são publicados ou não há versão digital disponível);
- Nível 1 - temos acesso à fonte de publicação do diário oficial deste município;
- Nível 2 - temos o script para coletar os arquivos e armazená-los em nossa base de dados;
- Nível 3 - o conteúdo dos diários oficiais do município está disponível na plataforma Querido Diário.

A concepção dos graus de maturidade se deve ao paralelismo das diferentes frentes do projeto, uma vez que o processamento dos dados ocorre sobre municípios que já estão mapeados e que contam com raspadores previamente desenvolvidos. É necessário explicar ao usuário que, dos 5.570 municípios, apenas uma parte está mapeada, outra parte derivada desta conta com raspadores, e uma parte ainda menor está integrada à plataforma de visualização.

Ao utilizar a plataforma, é possível selecionar o município em que é desejada a realização da busca e imediatamente visualizar o grau de maturidade do município, como apresentado na Figura 2:

Figura 2 - Seleção de município e apresentação do nível de abertura na versão para computadores.

Estrutura do mecanismo de busca

Aqui, nós disponibilizamos às pessoas usuárias do site a possibilidade de encontrar conteúdo dos diários com base em termos de busca livres, como apresentado na Figura 3. Além disso, a usuária poderá utilizar filtros para restringir o espaço da busca, com a escolha de datas e municípios. Este recurso utiliza as funcionalidades de busca e destaque dos termos, ambas já contempladas pela API.

Figura 3: Exemplo de termo de busca e filtro por data na versão para celulares.

Exibição dos resultados

O retorno da busca consiste principalmente na apresentação do nível de acesso do município e da lista de arquivos em que os termos tenham sido encontrados. Estes arquivos estarão ranqueados por relevância ou ordenados por data, e as usuárias terão a opção de acesso ao arquivo no próprio resultado da busca.

Os resultados da busca serão exibidos seguindo a divisão dos graus de maturidade descritos anteriormente. Se o conteúdo do diário oficial de um município está integrado à plataforma, são exibidos os trechos dos documentos referente ao termo e às datas informados nos filtros de busca. Caso o conteúdo não esteja integrado, serão listados todos os arquivos que atendem os filtros, mas não será possível realizar a busca textual. Todos os arquivos são disponibilizados em seu formato original e também em TXT.

Na Figura 4 são apresentados exemplos de retorno da busca caso não haja arquivos disponíveis para o município. As informações como instruções de como contribuir com o projeto, o significado do nível de acesso e endereços de publicação mapeados (apenas no nível 1) do município buscado, são apresentadas.

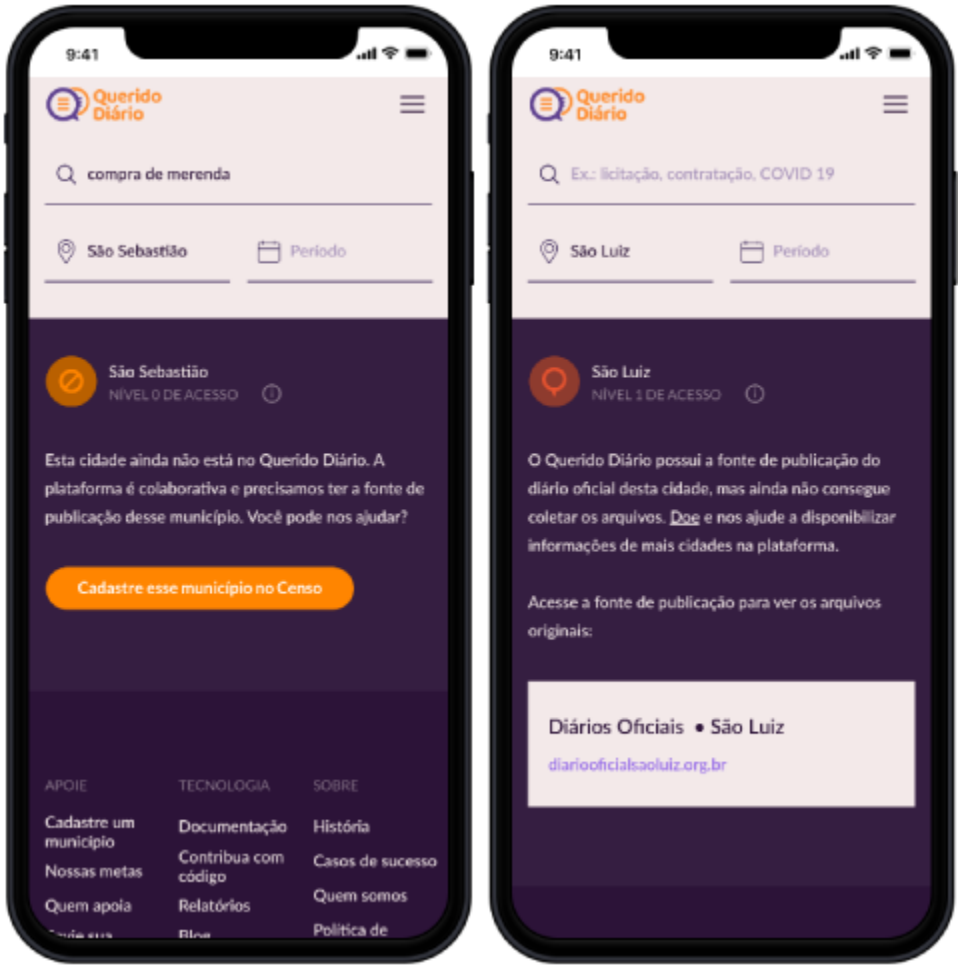

Figura 4 - Busca em municípios sem arquivos na base de dados, níveis 0 (esquerda) e 1 (direita), na versão para celulares.

Na Figura 5, são apresentados exemplos onde a busca retorna arquivos em municípios de níveis 2 e 3. No nível 2, como o conteúdo dos arquivos não está disponível para o mecanismo de busca, todos os arquivos do município buscado são apresentados. No nível 3, são apresentados os trechos nos arquivos onde o termo

buscado é encontrado de forma aproximada e apenas esses arquivos são disponibilizados, no formato original ou em TXT.

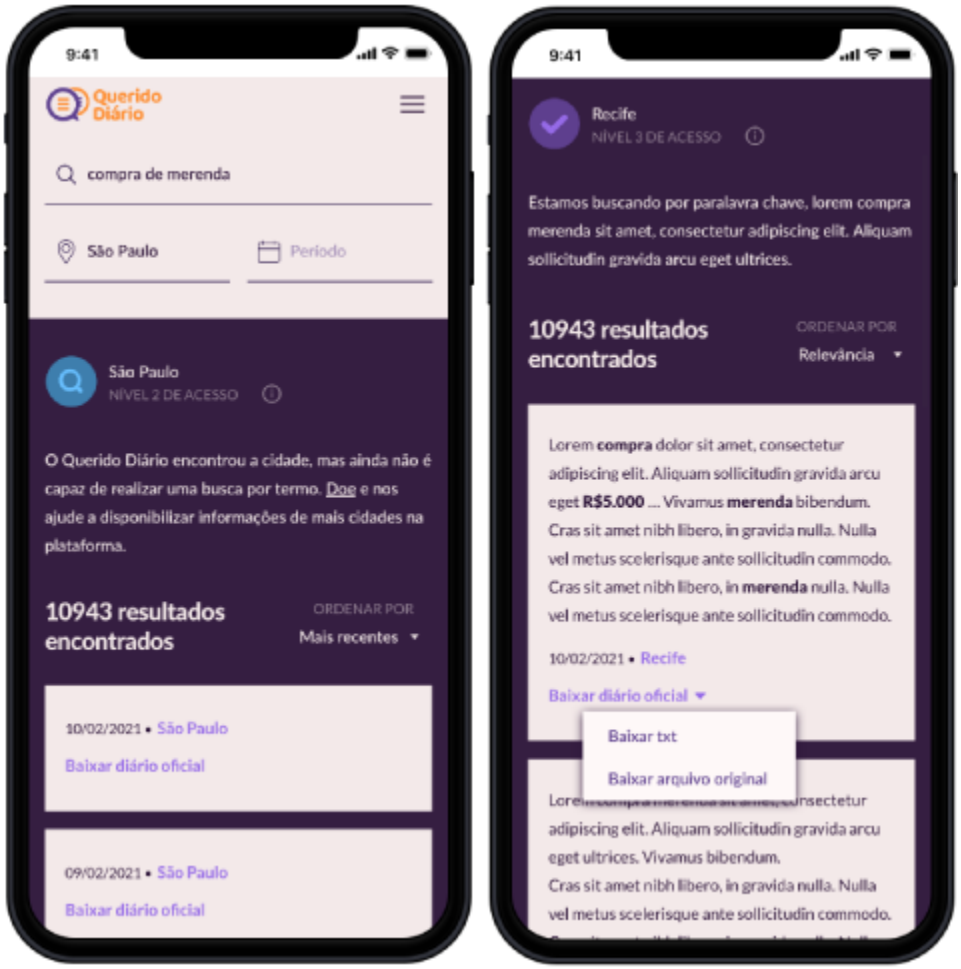

Figura 5 - Busca dos municípios nos níveis 2 e 3.

CASO DE USO: COLAPSO DO SISTEMA DE SAÚDE EM MANAUS

Análise dos atos oficiais e linha do tempo do colapso

Através de um parceria entre a OKBR e a jornalista Beatriz Farrugia, mestranda da Universidade de Birmingham, em regime de estágio voluntário, foi realizado o primeiro uso concreto da API.

A jornalista conseguiu identificar os atos oficiais da prefeitura de Manaus entre 17 de novembro de 2020 e 26 de janeiro de 2021 que teriam impacto no isolamento social e na resposta à COVID-19. Desta análise é possível acompanhar o desenvolvimento de eventos como o seguinte:

> *"Em 26 de novembro de 2020, foi publicado no Diário Oficial o edital de contratação, por seis meses, de 20 profissionais técnicos de patologia clínica para atuar nos Estabelecimentos Assistenciais de Saúde (EAS) no combate à COVID-19. No dia seguinte, no entanto, o Diário Oficial estabeleceu os procedimentos para a reabertura de restaurantes, padarias e supermercados, medidas que afrouxam o isolamento social".*

A reportagem aponta o atraso da resposta da prefeitura e traçou uma linha do tempo desses momentos-chave em paralelo ao número de casos e óbitos por COVID-19. O texto ainda não foi publicado, mas com ele já pudemos validar o tipo de impacto que este projeto pode ter.

CONCLUSÃO

Conforme indicado neste relatório, os algoritmos de inteligência artificial e outros mecanismos necessários ao funcionamento da plataforma de pesquisa Querido Diário, com toda a estrutura de coleta e processamento de texto, foram concluídos com sucesso no período desta primeira fase do projeto. Todos os links dos repositórios, bem como a documentação da API desenvolvida, foram indicados ao longo desta publicação para fins de verificação.

O desenvolvimento dos algoritmos utilizados é contínuo e cíclico: nossa equipe constantemente revisa, testa e melhora os programas existentes. Porém, este projeto proporcionou a criação de um MVP robusto e funcional que certamente servirá como referência para experimentos de inovação cívica e governo digital no cenário brasileiro.

SOBRE A OKBR

A Open Knowledge Brasil (OKBR), também conhecida como Rede pelo Conhecimento Livre, é uma organização da sociedade civil sem fins lucrativos e apartidária que atua no país desde 2013. Desenvolvemos e incentivamos o uso de tecnologias cívicas e de dados abertos, realizamos análises de políticas públicas e promovemos o conhecimento livre para tornar a relação entre governo e sociedade mais transparente e participativa.

Saiba mais no site: http://br.okfn.org

EQUIPE RESPONSÁVEL

COORDENAÇÃO GERAL
Fernanda Campagnucci

COORDENAÇÃO DE INOVAÇÃO CÍVICA
Giulio Carvalho

REVISÃO
Ariane Alves e Murilo Machado

CONTATO PARA IMPRENSA
imprensa@ok.org.br