PUBLICADORES DE DADOS
DA GESTÃO ESTRATÉGICA À ABERTURA

OPEN KNOWLEDGE BRASIL

ESCOLA DE DADOS

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
DE ACORDO COM ISBD
Elaborado por Odilio Moreira Junior – CRB 8/9949

P976

Publicadores de dados [recurso eletrônico]: da gestão estratégica à abertura / Bernardo Pacheco Loureiro... [et al.] ; organizado por Open Knowledge Brasil. – São Paulo : Open Knowledge Brasil, 2021.

100 p. ; PDF ; 3,4 MB.

Inclui índice e bibliografia.
ISBN 978-65-993954-1-3 (Ebook)

1. Dados. 2. Dados abertos. 3. Publicação de dados. 4. Administração pública. 5. Transparência governamental. I. Loureiro, Bernardo Pacheco. II. Campagnucci, Fernanda. III. Svab, Haydée. IV. Oliveira, Leandro Bispo. V. Baptista, Vitor. VI. Open Knowledge Brasil. VII. Título.

2021-703

CDD: 005.13
CDU: 004.62

OPEN KNOWLEDGE BRASIL

Open Knowledge Brasil (OKBR) é uma Organização da Sociedade Civil (OSC) sem fins lucrativos e apartidária legalmente constituída no país desde 2013. Durante a última década, vem desempenhando papel-chave na promoção dos dados governamentais abertos, por meio de uma combinação de mobilização de pessoas para fortalecer o controle social, desenvolvimento de software e materiais de capacitação, e engajamento da comunidade de software livre em projetos de tecnologia cívica.

E-MAIL: **contato@ok.org.br**
SITE: **https://ok.org.br**

ESCOLA DE **DADOS**

Escola de Dados é um programa da OKBR que apoia comunicadores, organizações da sociedade civil e instituições a extraírem o máximo potencial dos dados abertos. É responsável pela organização da maior conferência de jornalismo de dados e métodos digitais da América Latina (o Coda.Br) e já formou colaboradores(as) de grandes empresas e servidores(as) públicos de todos os estados brasileiros.

E-MAIL: **escoladedados@ok.org.br**
SITE: **https://escoladedados.org**

Abril/2021

AUTORIA DOS CAPÍTULOS
Bernardo Loureiro
Fernanda Campagnucci
Haydée Svab
Leandro Oliveira
Natalia Langenegger
Vitor Baptista

EDIÇÃO E REVISÃO
Murilo Bansi Machado

EQUIPE PEDAGÓGICA
Adriano Belisário (coordenação)
Anicely Santos (assessoria)
Edilaine Santos (estágio)
Isis Reis (comunicação e identidade visual)

PROJETO GRÁFICO E DIAGRAMAÇÃO
Mórula_Oficina de Ideias

Este ebook conta com o apoio do "Fundo de Apoio a Instituições que Fortalecem a Gestão Pública no Enfrentamento dos Impactos da Covid-19", iniciativa da "Aliança para Lideranças de Impacto no Setor Público e no Terceiro Setor", formada pela Fundação Brava, Fundação Lemann, Instituto Humanize e República.org.

SUMÁRIO

APRESENTAÇÃO

Bem-vindas, publicadoras e publicadores!

Quem já precisou trabalhar com dados em uma organização sabe que são muitas as barreiras: departamentos e públicos que não se conversam, qualidade sofrível da informação, ferramentas inadequadas e normas que não ajudam a avançar.

Parte do problema é que, em geral, tomadores de decisão e formuladores de política não têm a visão de todo o processo de gestão dos dados. Podem até ter uma preocupação estratégica, mas não estão familiarizados com aspectos técnicos que seriam importantes para promover o uso e a publicação de dados na organização — e acabam tomando decisões ruins. O contrário também pode acontecer: pessoas técnicas que não falam a língua da gestão e não compreendem por que suas boas ideias não saem do lugar.

O curso "Publicadores de Dados — da gestão estratégica à abertura" surgiu para conectar esses fios soltos. Notamos que não havia, especialmente em língua portuguesa, materiais que abordassem toda a trilha da abertura de dados, unindo aspectos de governança e planejamento a noções técnicas de desenvolvimento e publicação.

Depois de promover duas edições do curso, do qual participaram mais de 700 gestores de todo o Brasil, dos mais variados tipos de organizações públicas, decidimos compilar os principais aspectos da formação neste e-book. Esperamos, assim, que esse conhecimento possa chegar a mais pessoas e inspirar processos de abertura de dados país afora.

Estes são os perfis de pessoas que participam do processo de abertura de dados em uma organização e que acreditamos que se beneficiarão desta leitura:

- **especialista, técnica**: tanto na área de TI quanto em políticas públicas, são aquelas que vão formular e executar as políticas e desenvolver sistemas;
- **(intra)empreendedora**: promovem a inovação dentro da organização e superam obstáculos para que as ideias cheguem às tomadoras de decisão;
- **tomadora de decisão**: ocupam posições-chave de liderança, seja em departamentos ou na alta administração; são as pessoas que precisam ser convencidas ou que, se já sensibilizadas, vão patrocinar projetos de abertura;
- **assessora de gabinete**: têm acesso direto às tomadoras de decisão e às especialistas, promovendo pontes dentro da organização;
- **assessora jurídica e controle interno:** normalmente vistos como barreiras, esses perfis podem se envolver desde o início na concepção de projetos, ajudando a dar mais segurança e fortalecendo a gestão do órgão.

A trilha de publicação que concebemos está dividida em cinco etapas. Sabemos que não se trata de um processo linear — muitos desses passos se retroalimentam, funcionando como um ciclo a ser aperfeiçoado:

Ao longo deste guia, você verá as etapas desta trilha na margem direita das páginas, para que possa se situar durante o processo.

No primeiro módulo, “Compreender”, abordaremos os motivos para abrir dados e apresentaremos os tipos, formatos e padrões internacionais. É aqui, também, que buscaremos alinhar uma visão estratégica sobre a governança de dados em qualquer nível da organização.

No segundo passo, “Planejar”, trataremos da elaboração de catálogos e planos de dados abertos colaborativos, além da adequação dos processos de abertura de dados à Lei Geral de Proteção de Dados (LGPD).

A terceira etapa, “Desenvolver”, introduz noções sobre armazenamento de dados e automatização dos fluxos de dados — extração, tratamento e carga, que formam a famosa sigla ETL, em inglês. Também é aqui que apresentaremos fundamentos para boas visualizações de dados em painéis.

Avançando para o quarto passo da trilha, “Abrir”, apresentaremos boas práticas para o desenvolvimento de repositórios e APIs — um mecanismo que permite a interoperabilidade de sistemas e automação do consumo de dados, dentro e fora da organização.

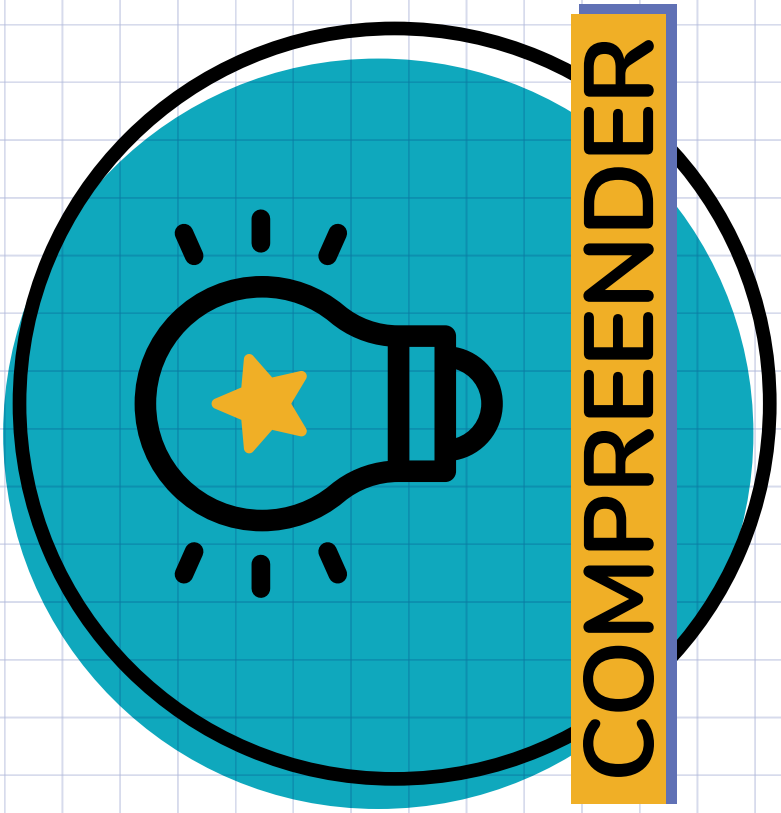
Finalmente, a trilha se encerra com “Conectar”. É aqui que falamos de ideias para ativar e impulsionar o ecossistema de inovação ao redor dos dados de sua organização. Essa etapa é o coração do processo de publicação de dados: é quando eles ganham sentido e se transformam em novos produtos e valores para a sociedade e para outros usuários do próprio setor público.

Estamos muito felizes em ter você trilhando conosco esse caminho da abertura de dados e da gestão baseada em evidências. Desejamos uma excelente jornada!

Fernanda Campagnucci
DIRETORA-EXECUTIVA DA OPEN KNOWLEDGE BRASIL

Adriano Belisário
COORDENADOR DA ESCOLA DE DADOS

Todos os links apresentados ao longo deste guia estão ativos na data de seu lançamento, em abril de 2021. Encontrou algum link quebrado? Avise-nos em **escoladedados@ok.org.br**


COMPREENDER

CAPÍTULO I

FORMULANDO POLÍTICAS PÚBLICAS BASEADAS EM DADOS

Fernanda Campagnucci

A jornada da publicação de dados em uma organização começa com o alinhamento de uma *visão estratégica* entre as pessoas envolvidas: das áreas técnicas, de gestão de políticas públicas, as que tomam decisões, usuárias. Você pode já estar *convencido* de que manter os dados organizados, seguros e disponíveis é um trabalho necessário e imprescindível. **Mas dificilmente sairá do lugar se cada um desses atores decidir caminhar em uma direção distinta**. Para isso, uma boa governança será fundamental — vamos falar mais sobre o tema neste capítulo.

Seja você uma pessoa tomadora de decisão ou operadora técnica, os benefícios de publicar dados — dentro e fora de sua organização! — precisam estar na ponta da língua. Lembre-se: o importante é começar. Quando os primeiros resultados aparecem, fica muito mais fácil fazer as pessoas enxergarem o valor que você já enxerga.

Por que abrir dados? Aqui vão cinco razões

1. **Você pode fazer mais com... *mais*.** Ganho de eficiência é um argumento imbatível. Principalmente se envolve redução de tempo — e, quase sempre, de custos — para fazer uma mesma rotina. Mas, quando você disponibiliza dados de qualidade, também ganha *mais recursos*: no mínimo, mais olhos, braços, cabeças. Em outras palavras: aproveita a inteligência coletiva, e às vezes chega a resultados mais vantajosos, que sequer planejava.
2. **Mais confiança das partes interessadas.** É verdade que a abertura de dados pode chamar atenção para problemas — é a primeira ressalva que você vai ouvir de quem pode tomar a decisão de abri-los. Mas é possível aproveitar o gesto da transparência para gerar mais confiança: "estamos abrindo porque queremos encontrar os problemas". Sem dados abertos, também é muito mais difícil se defender das *fake news* ou da desinformação.
3. **A confiabilidade *dos dados* melhora**. As pessoas que formulam políticas também podem errar menos. Se você está esperando ter dados melhores para poder abri-los, esqueça: *não vai acontecer*. Até porque os erros podem estar na origem, na produção dos dados, e só serão descobertos quando a base estiver sob a "luz do sol". De erros de digitação a fraudes, muito pode ser rapidamente corrigido quando os dados forem conferidos e usados por mais pessoas.

4. **A lei precisa ser cumprida**. A legislação determina a disponibilização de dados em *formato aberto*. É o que diz a Lei de Acesso à Informação (Lei Federal nº 12.527/2011), à qual as entidades e órgãos públicos de todos os níveis e poderes estão sujeitos. Seu artigo 8º, que trata de transparência ativa, determina que devem ser abertos "dados gerais para o acompanhamento de programas, ações, projetos e obras". Estabelece, ainda, que os sites públicos devem permitir baixar relatórios em formato aberto e não proprietário, além de permitir o acesso automatizado (§ 3º, incisos II e III).

5. **Portas abertas para a inovação**. Dados abertos são o insumo para a construção de ferramentas úteis para acessar serviços públicos e para desenvolver a ciência. Não acredite em processos de "transformação digital" que não trazem de forma bem articulada uma política de abertura de dados, de códigos e de colaboração — eles não conseguem transformar, de fato, a organização. No melhor dos casos, têm fôlego curto. No pior, geram dependência de padrões tecnológicos fechados e sua organização fica refém de um ou poucos fornecedores que, com frequência, são caros e ineficientes.

AINDA NÃO ESTÁ CONVENCIDO? BINGO!

Você não está sozinho na jornada de convencer colegas e lideranças a abrir dados. No mundo inteiro, pessoas ouvem argumentos semelhantes para mantê-los fechados. Daria para preencher uma cartela de bingo! Alguns são infundados, outros, preocupações legítimas — mas para tudo há uma resposta. Navegue pelo **Bingo dos Dados Abertos**[1] quando precisar de inspiração.

[1] Acesse: http://bit.ly/bingo-dados

Como formular e avaliar políticas com dados?

Construir *políticas públicas baseadas em evidências* é, na prática, elaborar e conduzir projetos e ações usando dados para entender *o que* funciona, *para que* funciona, e *por que* funciona. O uso de dados é uma ferramenta essencial na gestão pública, tendo em mente que há limites — um contexto e atores interessados que influenciam nas formulações.

Os dados podem estar presentes em todo o ciclo das políticas públicas. Na imagem abaixo, estão destacados os principais momentos e formas em que os dados são úteis.

FONTE: Adaptado do livro "Políticas Públicas", de Marta M. Assumpção Rodrigues (2010).

Construindo indicadores

Uma das principais razões para usar dados é compreender e comunicar o impacto de suas ações. Os indicadores são ferramentas úteis para identificar e medir esses esforços. Tanto do processo, para verificar se está num bom caminho; quanto em termos de resultados. Em suma, o ideal é construir um *sistema de indicadores* que deem conta dessas diferentes dimensões. Isolados, eles não têm muita serventia. Veja exemplos de indicadores para monitorar a eficiência, a eficácia e a efetividade de projetos e ações.

O QUE VOCÊ QUER MEDIR?

Como nos exemplos que aparecem na imagem, indicadores podem ter vários formatos: calculados, como taxas e médias; ou mesmo o número absoluto. Também é possível criar índices, que agregam diferentes aspectos a serem medidos (veja, por exemplo, o Índice de Transparência da Covid-19[2], desenvolvido pela Open Knowledge Brasil para monitorar o grau de abertura de dados na pandemia).

Governança de dados: não saia de casa sem ela

A governança de dados é o *conjunto de práticas e regras para organizar o uso e o controle dos dados* em uma organização. Quando se trata do processo de publicação de dados, uma boa governança ajuda a definir quais dados abrir, como e por quem devem ser abertos. Existem várias "camadas":

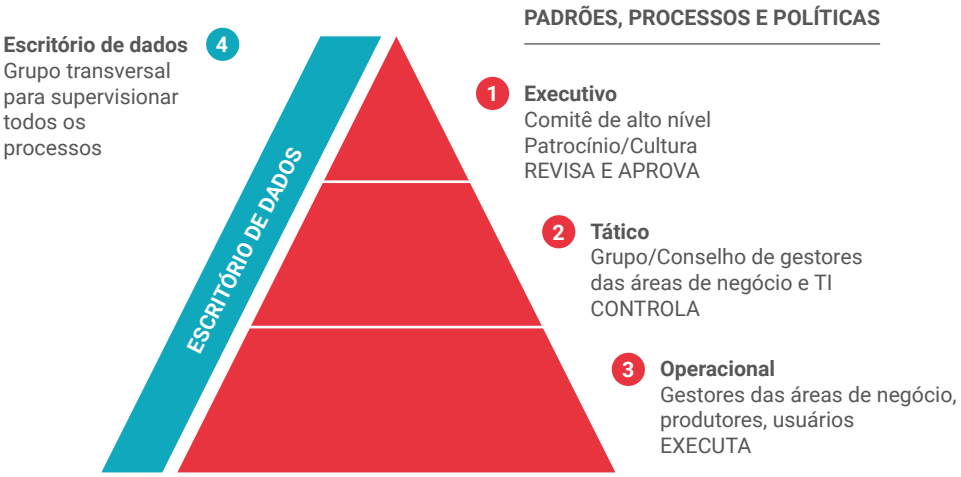
- gestão e liderança;
- aspectos jurídicos, com normas e políticas;
- aspectos técnicos, com a definição de padrões; e
- desenvolvimento de capacidades.

Trata-se de um processo de constante evolução: sua organização pode atingir diferentes graus de maturidade da governança de dados. Um bom jeito de começar ou aprimorar processos existentes é **mapear os principais problemas de dados**. Faça entrevistas com os usuários internos, desenhe jornadas que vão da coleta à análise de determinado conjunto de dados. Dessa forma, será possível priorizar os desafios mais comuns e elaborar um plano de ação para lidar com eles (com normas, padrões, novos procedimentos ou capacitação, por exemplo).

[2] Acesse: http://bit.ly/itc-19

É imprescindível haver uma boa articulação entre as diversas áreas de um órgão. Uma armadilha comum é deixar que o "pessoal da TI" seja o único responsável por tomar decisões. Há vários modelos possíveis para organizar essa articulação (comitês, grupos, escritórios de dados), mas é fundamental que haja esse olhar "matricial", como na imagem abaixo.

PESSOAS E PAPEIS



FONTE: Adaptado do livro "Governança de dados: práticas, conceitos e novos caminhos", de Carlos Barbieri (2019).

UM CASO INSPIRADOR: SECRETARIA MUNICIPAL DE EDUCAÇÃO DE SÃO PAULO

Para extrair informações estratégicas dos sistemas, como a demanda por creche em cada uma das diretorias regionais, os técnicos da Secretaria precisavam puxar dados de um sistema antigo realizando mais de 60 tarefas manuais. Esse tipo de relatório era feito trimestralmente.

Por meio de uma cooperação técnica sem transferência de recursos, em um processo liderado por técnicos da Pasta, foram implementadas ferramentas livres como o *Metabase* e o *Airflow* (leia mais sobre elas no Capítulo 6). Automatizada, a rotina passou a ser realizada diariamente.

Os diretores regionais puderam, com isso, ter números confiáveis para planejamento e abertura de vagas nos locais de maior demanda. Isso permitiu priorizar as áreas mais vulneráveis e reduzir o déficit histórico de vagas. A organização dos dados também permitiu dar mais transparência ao processo para as famílias, com uma ferramenta de consulta chamada **"Vaga na creche"**[3].

LEIA MAIS SOBRE ESSA INICIATIVA: **http://bit.ly/patio-digital** (pág. 85)

[3] Acesse: http://bit.ly/vaga-creche-sp

PARA SABER MAIS

LIVRO “Indicadores sociais no Brasil”, de Paulo Januzzi (2012).

LIVRO “Governança de dados: práticas, conceitos e novos caminhos”, de Carlos Barbieri (2019).

GUIA “Modelo de maturidade dos dados abertos”, de Leigh Dodds e Andrew Newman (2015). Acesse: http://bit.ly/guia-maturidade

GUIA "The Third Wave of Open Data", do The GovLab (2021). Acesse: http://bit.ly/3rdwave-od

CAPÍTULO II

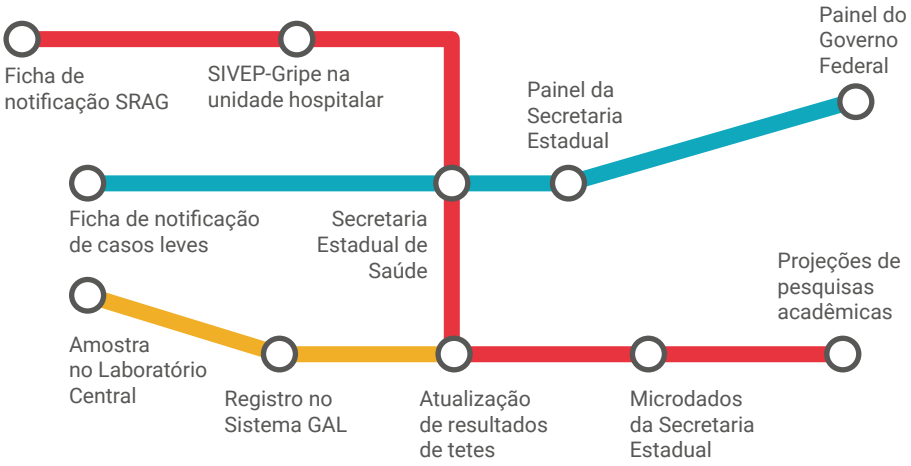
CONHECENDO FORMATOS, PADRÕES E TECNOLOGIAS PARA ABRIR DADOS

Fernanda Campagnucci

Você já se perguntou qual é a definição de *dados*? De forma simples, podemos dizer que são **valores** atribuídos a **aspectos** de objetos. Parece uma sutileza, mas é algo fundamental para se lembrar em todo trabalho com dados: você não está captando a realidade em si, mas aspectos que escolhe captar, quantificar. Tudo pode virar dados, mas, no final das contas, temos que ter em mente que são **construções humanas**, suscetíveis a erros, incertezas e diferentes interpretações. Inclusive causados pelo seu instrumento de medição — um termômetro mal calibrado, um sensor mal configurado.

Pensar que dados são construções humanas é muito útil para quem precisa lidar com eles em organizações. Imagine um trilho de metrô, em que cada estação é um momento no qual aquele dado passa por transformações de departamentos distintos. E as linhas, assim como as bases de dados, podem se cruzar, se fundir. Veja a imagem abaixo que, de forma simplificada, mostra a "jornada" dos dados da Covid-19 por diferentes sistemas e finalidades.

OS DADOS TÊM UMA JORNADA



DICA!

Se você quer compreender melhor os problemas de dados (como sugerimos no capítulo anterior, ao falar de governança), desenhar a jornada de uma base de dados pode ser um processo muito revelador. Junte uma equipe, faça um painel na parede, em que as pessoas podem colar post-its e registrar observações mais detalhadas sobre cada uma das etapas de coleta, transformação e uso dos dados. Você certamente levantará pontos a melhorar, ou descobrirá caminhos que os dados percorrem sem que você tenha ciência deles.

Tipos e formatos de dados

Os dados podem ter tipos distintos — é preciso fazer operações adequadas a cada um deles:

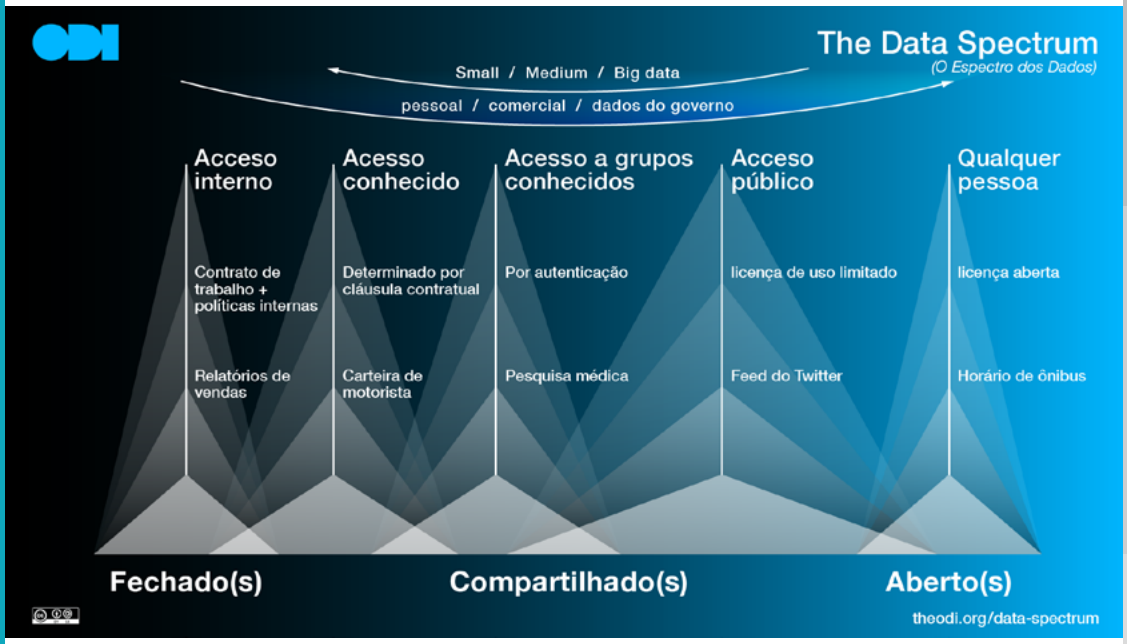
- **Categóricos.** Dividir os dados em "fatias" nos ajuda a estudar um fenômeno ou conjunto de objetos. Por exemplo, categorias de raça/cor/etnia: preta, parda, amarela, branca, indígena.
- **Discretos**. Dados numéricos, inteiros, que usamos para contar, basicamente. A idade das pessoas e os tamanhos de calçados são exemplos.
- **Contínuos**. São resultados de medição. Neste caso, todos os valores, inteiros e frações, são possíveis. O tamanho do pé é um exemplo.

Também existe uma variedade de formatos de dados. Aqui não se trata de formatos de arquivo, mas de "formas" em que você vai encontrá-los.

- **Sequências ordenadas**. Uma sequência de valores, ou itens, com uma única dimensão. Em algumas linguagens de programação, ela vai se chamar "array", como no Python, por exemplo, enquanto em outras será concebida como "vetor", como no R. O conceito fundamental permanece: o que temos é uma lista. Em Python, o seguinte comando cria uma *array* chamada frutas: *frutas = ["Banana", "Maçã", "Uva"]*.
- **Tabela**. Esta talvez seja a forma mais comum de pensarmos os dados. Nas tabelas, é como se tivéssemos sequências ordenadas empilhadas, ou seja, a nossa sequência de dados agora tem duas dimensões (colunas e linhas). Por isso, às vezes, podemos chamar esta estrutura de "arrays bidimensionais" ou matrizes.
- **Microdados**. É um tipo específico de tabela, em que cada "caso" é uma linha. Os dados sobre casos de coronavírus, por exemplo: uma tabela que registre a situação de cada paciente, suspeito ou confirmado, além de sua idade e sexo.
- **Redes e grafos**. A estrutura básica de uma rede é composta por nós (em geral, representados por pontos) e arestas (representadas por linhas que conectam os pontos). Esta forma de modelar os dados é especialmente útil quando precisamos analisar relações entre entidades.
- **Dados geográficos**. Ainda que possam ser representados por tabelas, eles demandam componentes relacionados ao espaço físico — no mínimo, um par de coordenadas geográficas (latitude, longitude). Para produzir mapas com os dados, é preciso ter uma projeção cartográfica, que permite transpor os dados para esse tipo de representação.
- **Texto**. Sequências de palavras também podem ser processadas como dados, a partir de técnicas conhecidas como Processamento de Linguagem Natural.
- **Imagens**. Com recursos da chamada computação visual, imagens em fotos ou vídeos também podem ser interpretadas como dados estruturados. Utilizando softwares como o *OpenCV*, por exemplo, é possível identificar e contar elementos, como objetos, pessoas ou seus atributos, em uma série de imagens ou vídeos.

Princípios e padrões internacionais

Outra distinção comum dos dados é sobre seu tipo de acesso: **fechados, compartilhados, abertos**. Pense nos dados ocupando um *espectro* de acordo com seu grau de disponibilidade pública, como nesta imagem do Open Data Institute:

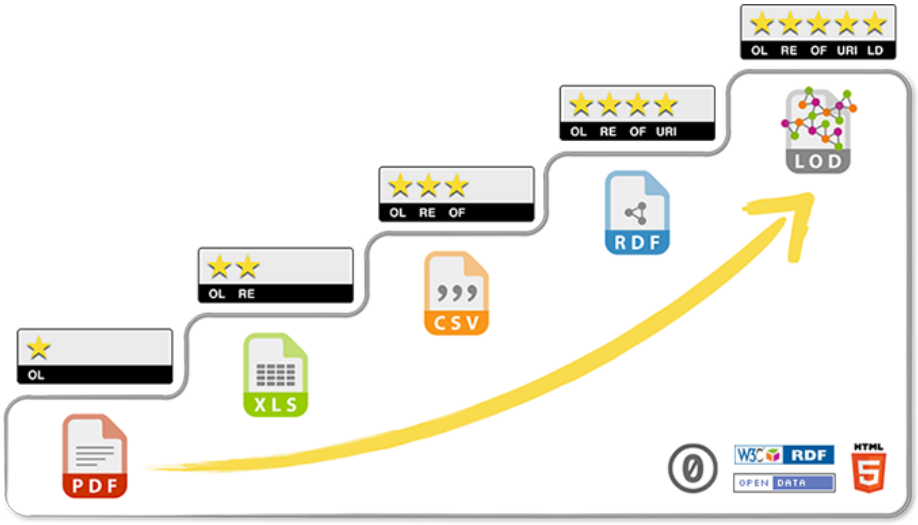


Nosso principal enfoque, neste guia sobre publicação de dados, são os **dados abertos**. Mas, neste caso, estamos tratando de **padrões** e **princípios** — o termo não significa, apenas, que os dados estão disponíveis abertamente.

Mesmo para dados que são restritos a pessoas autorizadas na organização, é muito útil manter padrões de dados abertos, pois eles são mais *interoperáveis*. Isso significa que permitem que os sistemas conversem melhor entre si.

De forma bem simplificada, uma definição de dados abertos é de que qualquer pessoa possa acessar, usar e compartilhar.

Tim Berners Lee, o inventor da Web, criou um modelo chamado “Cinco Estrelas dos Dados Abertos”, que você vê na imagem abaixo, em forma de uma escada. A cada degrau, fica mais fácil conectá-los com outras bases.



- ★ Acessível na web, em qualquer formato, com licença aberta;
- ★★ Estruturados e legíveis por máquinas (ex. uma planilha Excel em vez de um PDF);
- ★★★ Como o anterior, mas em formato não proprietário (ex.: CSV em vez de Excel);
- ★★★★ Todas as anteriores, e com uso de padrões da W3C (RDF e SPARQL) para identificação dos dados;
- ★★★★★ Todas as anteriores, mas com conexão com outros dados que proporcionem contexto.

Um arquivo com apenas uma estrela, como uma tabela escaneada e disponibilizada na internet sob uma licença livre, com muita dificuldade poderá ser usado para gerar valor à sociedade. Já aqueles com duas ou três estrelas se tornam mais acessíveis e proporcionam mais possibilidades de reutilização.

As bases de dados classificadas com quatro ou cinco estrelas impõem certa dificuldade técnica aos publicadores – conhecimento a respeito dos padrões RDF, por exemplo –, mas representam as melhores práticas para estimular o reúso de dados abertos pelo ecossistema de inovação. Na seção "Para saber mais", apresentamos alguns guias para aplicar esses conceitos.

Com as definições e princípios apresentados, podemos afirmar que **somente a partir das três estrelas** uma base de dados está em conformidade com os padrões de **dados abertos da legislação nacional** (como a Lei de Acesso à Informação).

Tecnologias livres e abertas

Não são só os dados que podem ser abertos: os *softwares* também. Os sistemas de código aberto (*open source*) têm seu código-fonte acessível, além de liberdades para uso, cópia e distribuição (que podem variar de acordo com a licença), e a possibilidade de trabalhos derivados. Existe uma discussão sobre a diferença entre software livre e *open source*, mas, para simplificar, estamos usando o termo "código aberto" como um sinônimo que abarca essas características gerais.

Você pode usar softwares de código aberto em sua organização ou desenvolver os seus próprios sistemas nessa modalidade. Alguns exemplos de diferentes aplicações (alguns deles abordaremos ao longo deste guia): Python, R, Apache Airflow, Metabase, CKAN, Wordpress.

Os benefícios são semelhantes aos que vimos no caso dos dados abertos:

- **colaboração**: outras pessoas podem estudar e propor melhorias;
- **transparência**: para governança compartilhada e também dos investimentos realizados;
- **inovação**: mais funcionalidades podem ser criadas, e experimentações podem ser feitas; e
- **interoperabilidade**: sua capacidade de comunicação com outros sistemas é maior.

Há também desafios, como criar capacidades para sustentação e para manter comunidades atuantes em torno do software. Mas, mesmo na ausência de pessoal qualificado, você pode estabelecer contratos que prevejam a abertura do código-fonte e sua manutenção.

TRANSPARÊNCIA ALGORÍTMICA

Algoritmos são as regras, ou o "passo a passo", com as quais os sistemas são instruídos para a tomada de decisões. No setor público, são aplicados para definir serviços, direitos, cálculos de impostos devidos etc.

Dependendo da maneira como são implementados, e do tipo de dado que alimenta esses sistemas, há riscos de gerar distorções e desigualdades, impactando grupos vulneráveis ou minoritários. Por isso, a transparência torna-se essencial. Não só dos códigos (que muitas vezes não são compreensíveis), mas dos dados utilizados, dos objetivos do sistema e dos resultados obtidos.

PARA SABER MAIS

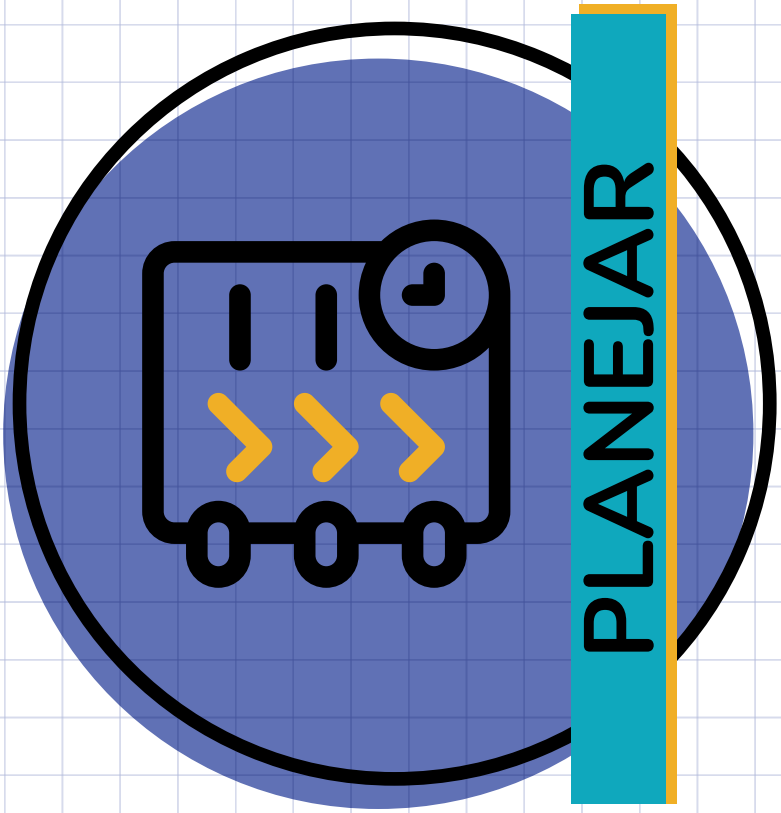
ESTUDO (em inglês) "A Governance Framework For Algorithmic Accountability And Transparency", do Parlamento Europeu (2019). Acesse: http://bit.ly/algor-account

DOCUMENTO "Boas práticas de dados na web", do W3C (2017). Acesse: http://bit.ly/praticas-dados

LIVRO "Dados abertos conectados", de Seiji Isotani e Ig Ibert Bittencourt (2015). Acesse: http://bit.ly/dadosabertoscon

SITE "Open Data Handbook". Acesse: http://bit.ly/opendatah

SITE "Open Source Definition". Acesse: http://bit.ly/osdefinition


PLANEJAR

CAPÍTULO III

ORGANIZE-SE PARA ABRIR!

Fernanda Campagnucci

Agora que sua organização já começou a alinhar uma **visão estratégica sobre os dados** e as principais partes interessadas já compreenderam a importância de uma **boa governança**, é hora de avançar em dois instrumentos fundamentais para a gestão e a abertura de dados: **os catálogos (ou inventários) de dados** e os **planos de dados abertos**.

Nunca é demais lembrar: todas as etapas do processo da abertura são incrementais e possuem graus de maturidade que podem ser permanentemente aprofundados. O importante é começar — a própria jornada de construção desses dois instrumentos contribuirá para uma governança de dados mais sólida.

Catálogos de dados

Você sabe quais são e onde estão todos os dados de sua organização? A verdade é que a maioria das pessoas deve responder que não: no máximo, conhece bem as bases de dados de seu setor. Não só em governos, mas também empresas, é frequente que a lógica de "silos" impere. Cada departamento, por mais que haja uma missão comum, tende a se fechar em suas rotinas e a guardar informações para si. Mesmo que isso implique retrabalho, ineficiência e falhas de comunicação.

Os catálogos são pressupostos para a governança e excelentes pontos de partida para superar a cultura de silos. Também são o primeiro passo para a construção de planos de dados e de repositórios de dados abertos.

Um catálogo de dados é uma **lista que registra todas as bases de dados** de uma organização. Deve ser feito de forma estruturada: é como se fosse uma base de dados sobre as bases de dados. Aqui usamos o termo "catálogo" no sentido de **inventário**, para diferenciar dos **repositórios de dados**, que armazenam ou permitem acessar as bases de dados em si (e que veremos mais adiante neste guia).

Com a entrada em vigor da legislação de proteção de dados pessoais (de que trataremos no próximo capítulo) e com a necessidade crescente de garantir a segurança cibernética, o catálogo de dados torna-se uma ferramenta mais importante do que nunca.

POR QUE SUA ORGANIZAÇÃO PRECISA DE UM CATÁLOGO DE DADOS?

- Proporciona uma boa **governança** de dados, inclusive para a comunicação com pessoas externas à organização;
- Permite a **cooperação** entre setores, entregando melhores serviços e políticas;
- Aumenta a **consistência** dos dados, evitando sobreposições de bases;
- É essencial para o **planejamento** das políticas de abertura de dados, pois dá uma visão geral sobre os ativos da organização;
- Possibilita o controle interno, tais como *checklists* para garantir a segurança e a privacidade, e para medir o grau de transparência da organização.

O que há em um catálogo?

Os catálogos podem ser adaptados de acordo com a realidade de cada organização mas, em geral, eles contêm os seguintes **metadados** de cada uma das bases:

- Título;
- Descrição de conteúdo;
- Periodicidade;
- Fonte dos dados;
- Mantenedor; nos termos da LGPD:
  - Controlador
  - Operador
  - Encarregado;
- Dados pessoais / Nível de proteção

Você voltará a ver esse assunto no capítulo 8 deste guia, quando apresentamos as questões essenciais sobre documentação.

E o que deve ser objeto do catálogo? Idealmente, **todas** as bases de dados existentes na organização. Estas podem estar em sistemas ou planilhas armazenadas em departamentos específicos. Mas, se precisar começar com um escopo menor, mire nos dados que sejam efetivamente consultados, atualizados ou usados por diferentes pessoas: técnicos, gestores, população.

E isso vale para qualquer nível da organização em que você exerça alguma influência: pode ser a Administração como um todo (Prefeitura, Estado, Governo Federal) ou um órgão específico (uma secretaria, um ministério) ou um departamento (um setor, coordenadoria, diretoria). Qualquer que seja o nível, um catálogo compartilhado fará toda a diferença.

Um catálogo será útil em qualquer nível. Quanto maior a abrangência, melhor o resultado para a organização.

CASO PARA INSPIRAÇÃO: CATÁLOGO MUNICIPAL DE BASES DE DADOS (CMBD) DE SÃO PAULO

A cidade determinou a criação do CMBD em 2014, com o Decreto Municipal nº 54.779/2014. Os responsáveis eram o DEINFO (departamento da então Secretaria Municipal de Desenvolvimento Urbano, que detinha grande quantidade de dados de toda a prefeitura); e a Controladoria Geral do Município, responsável pela implementação da Lei de Acesso à Informação.

Uma primeira versão do inventário, com 591 bases, foi publicada na forma de planilha no Portal da Transparência da cidade. O CMBD tem sido atualizado periodicamente, e em 2020 contava com 938 bases. Essa "base sobre as bases" se encontra no Portal de Dados Abertos[1] da cidade. Um Guia[2] foi produzido para orientar os setores.

[1] Acesse: http://bit.ly/cmbd-sp

[2] Acesse: http://bit.ly/guia-cmbdsp

As principais etapas para produção de um catálogo são:

1. **Identificar os responsáveis**. Quem lidera, quem é o gestor de dados em cada departamento e os pontos focais.
2. **Criar o instrumento de coleta**. Definir os campos do formulário ou planilha; testar e aprimorar (ou não será útil, nem cumprido!). Comece simples.
3. **Formar os responsáveis**. Com oficinas, videotutoriais, manuais ou guias. Cuidado com a abordagem: os setores devem ser envolvidos e engajados.
4. **Preenchimento**. Estabeleça fases, metas, prazos. Coteje o resultado com relatórios de sistemas e bancos de dados.
5. **Manutenção e atualização**. Crie processos e rotinas para revisão periódica.

Uma base normativa (orientação, portaria ou decreto) é importante para consolidar o processo e os papéis.

Planos colaborativos de dados abertos

Um Plano de Dados Abertos (PDA) é o instrumento que efetivamente implementa a Política de Transparência Ativa. Este deve ser feito de forma **colaborativa** porque, dessa maneira, garante que as prioridades de abertura de dados se **alinhem à demanda** dos diferentes tipos de usuários. Também minimiza a descontinuidade entre diferentes gestões. Finalmente, com os PDAs as pessoas podem participar de todo o ciclo de vida de dados.

Vamos às principais etapas para a construção do plano colaborativo de dados abertos:

1. **Defina os responsáveis** e alinhe a visão internamente.
2. **Use seu inventário e outros *inputs*.** Dependendo da área de política pública, os planos decenais e outros instrumentos podem ser uma fonte importante (por exemplo, para definir indicadores prioritários para abertura). Pedidos de e-SIC mais frequentes também são estratégicos para compreender a demanda.

3. **Engaje o público**. Promova consultas públicas, oficinas, bate-papos.
4. **Priorize!** Estabeleça os critérios relevantes para a organização. A Controladoria Geral da União, por exemplo, sugere uma matriz em seu Manual de Elaboração de PDAs (veja referência ao final deste capítulo).
5. **Defina um cronograma** de abertura, estratégias e responsáveis.
6. **Publique e monitore!**

CASOS INTERNACIONAIS: NOVA IORQUE (EUA) E TORONTO (CANADÁ)

Em Nova Iorque, a Open Data Law[3] de 2012 estabeleceu a abertura até 2018 de todas as bases da cidade com informação pública. Para monitorar, institui planos de *compliance* anuais, além de apontar coordenadores de dados abertos em cada órgão. Um inventário das bases[4] foi publicado em 2017. Dois anos depois, construiu uma visão de longo prazo[5] para a década seguinte. Entre as principais diretrizes, está o codesenho da política com usuários, além da noção de dados abertos como uma plataforma cívica para resolução de problemas da cidade.

Toronto segue uma linha semelhante. Seu "plano diretor" de dados abertos de 2018 traz uma visão de cinco anos, também prezando pelo codesenvolvimento com cidadãos e a resolução de problemas. Tem um enfoque importante na melhoria da qualidade dos dados. O plano propõe uma "trilha" de etapas parecida com a deste guia: compreender, planejar, desenvolver, abrir e conectar.

Abrir os dados e torná-los mais úteis e melhores: este é o plano!

[3] Acesse: http://bit.ly/open-data-law

[4] Acesse: http://bit.ly/da-inventory

[5] Acesse: http://bit.ly/next-dec-data

PARA SABER MAIS

MANUAL "Elaboração de planos de dados abertos", da Controladoria Geral da União. Acesse: http://bit.ly/manual-plano

GUIA "Fundamentos para publicação de dados na web", do Ceweb.br. Acesse: http://bit.ly/fundamentos-dados

GUIA (em inglês): "New York City Open Data Playbook", da cidade de Nova Iorque (EUA), para orientar os planos de *compliance* anuais. Acesse: http://bit.ly/nyc-playbook

DOCUMENTO (em inglês): "Open Data Master Plan (2018-2022)", da cidade de Toronto (Canadá). Acesse: http://bit.ly/plano-toronto

CAPÍTULO IV

ADEQUANDO O PROCESSO DE ABERTURA DE DADOS À LEI GERAL DE PROTEÇÃO DE DADOS (LGPD)

Natalia Langenegger

Como vimos no primeiro capítulo, a abertura de dados apresenta incontáveis benefícios. Mas, entre os dados implicados nas iniciativas de abertura, pode haver dados pessoais (que vinculam ou permitem a vinculação a uma pessoa física). Ainda que, por diversas vezes, eles sejam imprecisamente referidos como dados públicos por serem divulgados em portais de transparência, muitos são dados pessoais e devem ser utilizados conforme legislação específica.

Importante destacar, no entanto, que a **possibilidade de se extraírem dados pessoais de bases de dados governamentais, ainda que estes exijam cuidados adicionais, não deve ser utilizada como obstáculo para a transparência governamental**. Como veremos, não somente a legislação permite a divulgação de dados pessoais quando houver interesse público, mas também apresenta balizas para guiar essa divulgação.

Regulação aplicável à divulgação de dados pessoais

A abertura de dados possui respaldo nos princípios constitucionais de acesso à informação e da publicidade dos atos administrativos (artigos 5º, 33 e 37 da Constituição Federal), que foram regulados pela Lei de Acesso à Informação (LAI).

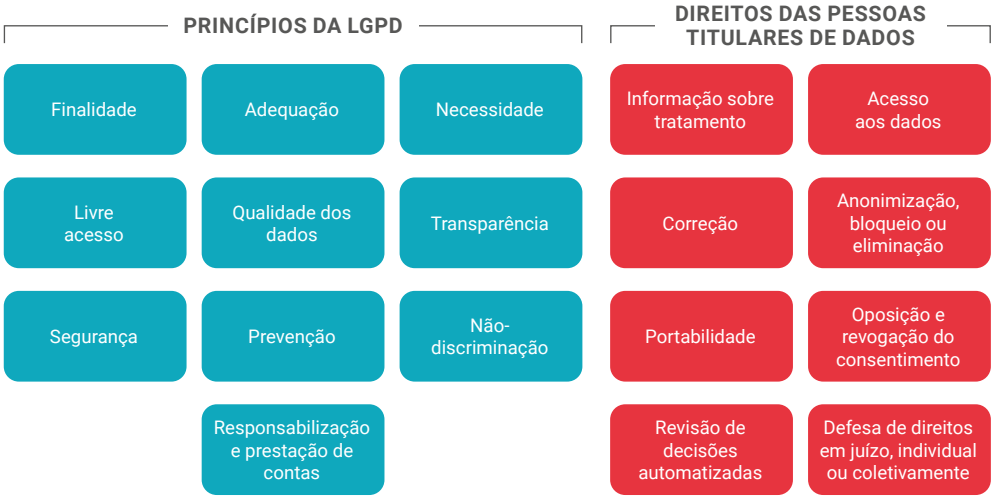
Essa obrigação de transparência governamental tem exceções, como a divulgação de informações pessoais. O objetivo é proteger as pessoas contra usos de seus dados de formas que possam lhes prejudicar. No entanto, essa restrição também é limitada pelos casos em que a divulgação de dados pessoais é de interesse público (artigo 31 da LAI).

Salvo quando a publicação de dados pessoais está prevista em lei, a avaliação deverá ser feita pelo próprio gestor público e poderá ser referendada pelo Poder Judiciário. Por exemplo, a publicação de salários de

servidores públicos foi objeto de ação judicial, oportunidade na qual os tribunais decidiram pelo interesse público da divulgação e argumentaram haver medidas técnicas capazes de mitigar os riscos à privacidade.[1]

À época desses julgamentos, ainda não havia lei específica para tanto, cenário que foi alterado pela Lei nº 13.709/2018, a Lei Geral de Proteção de Dados (LGPD). Sancionada em 2018, a LGPD fornece medidas jurídicas para o devido tratamento de dados pessoais. Ela estabelece princípios (ex: finalidade, necessidade e transparência) e regras (ex: observância de uma base legal e adoção de medidas de segurança da informação) que deverão ser observados no tratamento de dados pessoais, bem como prevê direitos aos titulares dos dados (ex: acesso, correção e portabilidade). A imagem abaixo resume essas previsões.

PRINCÍPIOS E DIREITOS DO TITULAR



[1] Agravo Regimental na Suspensão de Segurança 3.902, do Supremo Tribunal Federal. Acesse: http://bit.ly/agravo-stf

Também recomenda àqueles que usam dados pessoais em suas operações a adoção de uma estrutura de governança e de procedimentos para assegurar aderência de suas práticas à LGPD. Isso significa, por exemplo:

- contar com a nomeação de uma **pessoa encarregada** para atuar como canal de comunicação com indivíduos e autoridades;
- **mapear** atividades de tratamento de dados realizadas; e
- elaborar **relatório de impacto** para atividades que possam prejudicar direitos dos titulares dos dados.

Para o poder público, a lei exige que qualquer atividade com dados pessoais observe o **interesse público** e as **competências legais** do órgão. Essa exigência deverá ser compreendida em conjunto com os princípios constitucionais da legalidade e da impessoalidade, que exigem aos agentes públicos se guiarem pela Constituição Federal e pela e legislação vigente.

A LGPD também prevê a **relevância** na divulgação dos dados em posse do poder público, ao **(i)** estabelecer que entes públicos deverão manter dados em formato interoperável e estruturado para viabilizar a descentralização da atividade pública e a disseminação das informações; e **(ii)** estabelecer que dados cujo acesso é público devem ser usados em observância à finalidade[2], à boa-fé[3] e ao interesse público, que justificaram sua disponibilidade.

[2] A finalidade exige a utilização de dados pessoais para propósitos legítimos, específicos e informados ao titular, não sendo possível tratar os dados posteriormente de forma incompatível com as finalidades que justificaram sua coleta.

[3] Isso significa o compromisso de não tratar dados de maneira ilegal ou que viole direitos do titular, além de agir dentro das expectativas legítimas do titular em relação àqueles dados.

Como divulgar dados em observância à LGPD?

Assim, a presença ou a possibilidade de extração (em virtude de cruzamento de dados ou da fragilidade de medidas de anonimização) de dados pessoais em determinada base de dados não significa que ela não deva ser divulgada.

Sempre que o gestor público se confrontar com situações nas quais informações de indivíduos poderão ser expostas, deverá avaliar o interesse público na divulgação e observar o disposto na LGPD. Isso significa:

- **Avaliar se a base de dados possui dados pessoais, identificados ou identificáveis**. Em caso afirmativo, será necessário avaliar se: **(i)** há interesse público na divulgação; **(ii)** as técnicas de anonimização prejudicariam o interesse público – se não, elas devem ser utilizadas.
- Havendo interesse público na divulgação, **elaborar relatório de impacto à proteção de dados pessoais**. Esse documento deverá identificar, entre outros:
  - Os dados pessoais contidos na base de dados (ou que poderão ser dela extraídos), a natureza dos dados (ex: dados sensíveis, como raça ou orientação sexual) e quem são titulares dos dados (ex: beneficiários de programas sociais);
  - A justificativa do interesse público e da necessidade de divulgação de cada um dos dados. Os não necessários ao interesse público devem ser removidos (ex: a publicação do salário de servidores não exige a divulgação de seus endereços);
  - A ponderação dessas informações a fim de identificar se os riscos às pessoas titulares de dados são superiores ao interesse público.
- **Informar titulares de dados** sobre a possibilidade **(i)** de seus dados serem publicados em portais de dados abertos, e **(ii)** de exercerem os direitos previstos na LGPD, como correção de dados imprecisos.

A avaliação entre interesse público e riscos a direitos de titulares de dados pessoais poderá variar conforme o caso concreto. Em algumas situações, o interesse público deverá prevalecer em relação à privacidade e, em outras, os prejuízos demasiados a direitos de titulares poderão falar mais alto.

EXEMPLO 1 A divulgação de dados de pessoas beneficiárias de programas sociais (como NIS, nome e cidade de residência) pode ser fundamentada pela possibilidade de a sociedade contribuir com a fiscalização sobre a devida alocação de verbas públicas. Por outro lado, titulares de dados terão o seu direito à privacidade afrontado. Devemos considerar que esses indivíduos são geralmente mulheres que enfrentam dificuldades no exercício de direitos básicos, como acesso ao mercado formal de trabalho, além de esses dados serem recorrentemente utilizados de forma discriminatória ou para a prática de fraudes. Nesse caso, portanto, os riscos gerados à privacidade tendem a se sobrepor ao interesse público subjacente, e a quantidade de informações sobre cada indivíduo poderia ser revista[4].

EXEMPLO 2 Do mesmo modo, a divulgação de nome e salário de servidores públicos pode ser fundamentada pela possibilidade de a sociedade contribuir com a fiscalização sobre a devida alocação de verbas públicas. Por outro lado, titulares de dados terão afrontado seu direito à privacidade. Entretanto, considerado o contexto, o interesse público tende a se sobrepor à privacidade.

[4] Confira também o estudo "Bolsa família: pensando a privacidade das titulares". Acesse: http://bit.ly/bolsafam-dados

Como demonstramos, a existência de dados pessoais em bases de dados públicas não deve ser utilizada como justificativa para a opacidade governamental. A legislação não somente permite a divulgação de dados pessoais quando houver interesse público, como também apresenta balizas para guiar a divulgação dos dados de maneira adequada. Em síntese, ao publicar uma base de dados pública, deve ser avaliada a possibilidade de serem utilizadas técnicas de anonimização, bem como a preponderância do interesse público em relação aos direitos de titulares de dados.

PARA SABER MAIS

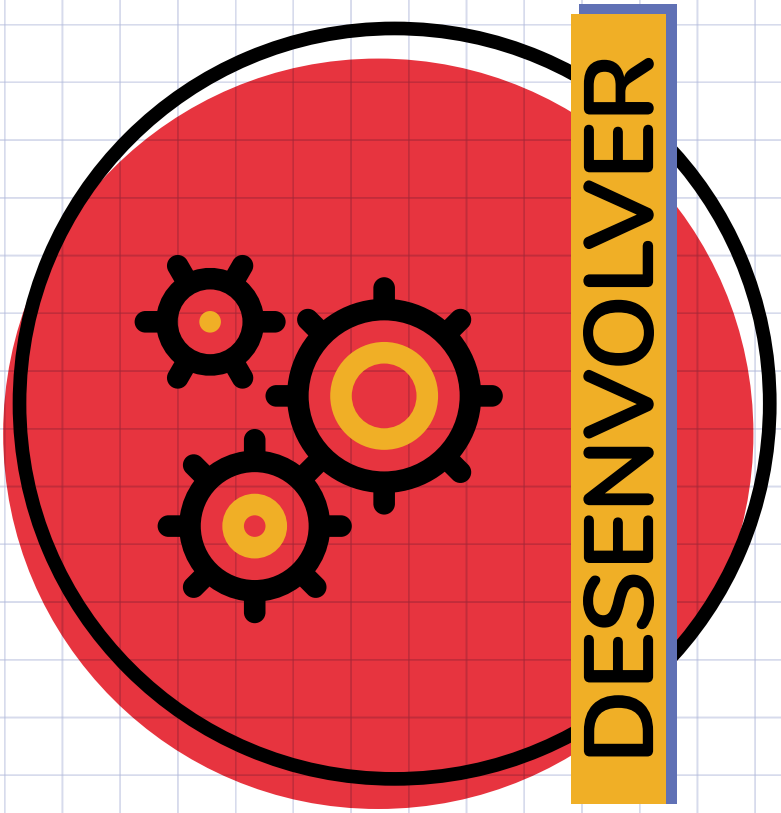
ARTIGO "Cidadania, tecnologia e governo digital: proteção de dados pessoais no Estado movido a dados", de Miriam Wimmer (2020). Acesse: http://bit.ly/mwimmer (pág. 27)

MANUAL "Manual de proteção de dados pessoais para gestores e gestoras públicas educacionais", do CIEB (2020). Acesse: http://bit.ly/manual-cieb

ARTIGO (em inglês): "Open Data, Privacy, And Fair Information Principles: Towards a Balancing Framework", de Frederik Borgesius e outros (2015). Acesse: http://bit.ly/fborgesius

ARTIGO (em inglês): "The Relationship Between Open Data Initiatives, Privacy, And Government Transparency: a Love Triangle?", de Tanja Jaatinen (2016).

ARTIGO (em inglês): "Towards a Modern Approach to Privacy-Aware Government Data Releases", de Micah Altman e outros (2015). Acesse: http://bit.ly/priv-gov-data


DESENVOLVER

CAPÍTULO V

ARMAZENANDO OS DADOS EXISTENTES

Leandro Oliveira

Depois de compreender a necessidade e a importância de abrir dados, e de planejar como sua organização pode fazer isso, é hora de iniciar a execução. Com isso, a discussão sobre como armazenar informações se torna crucial — e o desafio para armazenar e manter dados seguros é cada dia mais complexo.

Armazenamento de dados se refere ao uso de mídia de gravação para guardar informações usando dispositivos. As soluções existentes são baseadas em serviços físicos que podem utilizar um servidor, unidades de mídia, ou serviços em computadores de terceiros, as chamadas soluções "na nuvem".

Lembre-se: problemas relacionados à imprecisão ou à perda de dados podem acarretar prejuízos diversos, como indisponibilidade de serviços, impacto na credibilidade da organização, problemas jurídicos, entre outros.

Neste capítulo, vamos apresentar opções de armazenamento compatíveis com diferentes tipos de projetos e dar dicas para que sua organização tire o melhor proveito dos dados.

Por que nos preocupar em armazenar e manter dados?

A cada dia, passamos a lidar com um volume maior de dados. Para dar conta disso, governos e grandes organizações costumam dispor de provedores de nuvem com servidores escaláveis e armazenamento "infinito", pagando pela utilização.

É de suma importância estar sempre atualizando os servidores de armazenamento para acompanhar o crescimento do volume de dados, além de garantir o acesso às informações de forma satisfatória.

Aqui, a complexidade envolve não apenas as tecnologias empregadas para o armazenamento, mas também sua capacidade. Este infográfico[1] mostra a evolução do tamanho dos arquivos digitais desde os disquetes até o cenário para o qual as organizações já devem se preparar: petabytes e exabytes de informações!

Como armazenar dados com segurança?

Podemos separar as diversas soluções de armazenamento da seguinte forma:

- **Storage**
  - Armazenamento físico — local
  - Armazenamento físico — em rede
  - Armazenamento em nuvem
- **Banco de dados**
  - Armazém de dados (do inglês, *data warehouse*)

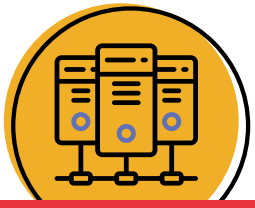
Armazenamento físico — local se refere a soluções de uso individual, disponíveis em computador, tablet, celular etc. As mais comuns atualmente são:

- **Meio óptico:** CDs, DVDs, Blu-Rays;
- **Meio magnético:** HDs;
- **Meio eletrônico:** Pendrives, SSDs, cartões de memória.

Cada um desses meios oferece espaço e velocidade de acesso/gravação predeterminados. Embora ainda estejam presentes em algumas organizações, os de meio óptico vêm perdendo espaço, já que são considerados defasados, por conta de sua capacidade de armazenamento limitada e da baixa durabilidade da mídia.

[1] Acesse: https://bit.ly/armaz-info

Com o **Armazenamento físico — em rede**, também conhecido como *Network Attached Storage* ou NAS, os dados ficam acessíveis para redes internas, em que um ou mais servidores com discos dedicados armazenam dados e os compartilham com vários "clientes" (na linguagem de computação, são computadores ou outros dispositivos que consomem recursos de um servidor) conectados a uma mesma rede.



SERVIDOR é um computador com alta capacidade de armazenamento e processamento, capaz de fornecer serviços de forma centralizada para uma rede de outros computadores ou dispositivos conectados. Existem vários tipos de servidores e funções que podem desempenhar. Por exemplo, podem armazenar e "servir" (compartilhar) arquivos e dados; hospedar e executar as aplicações necessárias para a publicação de um site; gerenciar e enviar e-mails, fila de impressão; ou mais de uma tarefa simultaneamente.

Diferente das soluções locais, como HDs ou pendrives, que podem ser conectadas a apenas um dispositivo por vez, a solução de armazenamento em rede oferece suporte a vários dispositivos simultaneamente.

Já no **armazenamento em nuvem**, os dados são armazenados e distribuídos em uma estrutura de servidores de grande escala, com alto poder de processamento e armazenamento. Em geral, são providos por grandes empresas, como Amazon, Google e Microsoft, que garantem recursos operacionais sob demanda, disponibilizando painéis em que é possível gerenciar esses recursos.

FERRAMENTA	CÓDIGO ABERTO	CUSTO	INFRAESTRUTURA
Owncloud	Sim	Gratuito (exceto custo de infra)	Da organização
Nextcloud	Sim	Gratuito (na versão auto-hospedada)	Da organização ou parte do serviço (Nextcloud Enterprise)
Google Cloud Platform (GCS)	Não	Pago	Parte do serviço
Amazon Web Services	Não	Pago	Parte do serviço

De maneira geral, os melhores serviços de armazenamento em nuvem combinam:

- Simplicidade e baixa necessidade de configuração;
- Interface intuitiva;
- Custos que variam de acordo com a necessidade.

O importante é proteger os arquivos, realizar a gestão de backups e proporcionar tranquilidade à organização, sabendo que os dados estão seguros e protegidos.

E qual a melhor opção de armazenamento?

Cada uma das soluções apresentadas acima possui diversas vantagens e desvantagens. Vamos analisá-las:

SOLUÇÕES	VANTAGENS	DESVANTAGENS
Físico — Local 	• Uso pessoal	• Potenciais problemas de segurança e integridade das informações; • Dificuldade para compartilhamento; necessário outro dispositivo para backup
Físico — Rede 	• Controle de gestão de acessos, acesso remoto, sem necessidade de internet para funcionar	• Flexibilidade de investimento (custo alto para compra de equipamentos); • Segurança sob responsabilidade da própria organização
Nuvem 	• Acesso através de qualquer dispositivo, investimento e custos de utilização sob demanda, infraestrutura sob responsabilidade do provedor	• Possível indisponibilidade devido a problemas na região onde os servidores se encontram; • Performance pode ser reduzida

AINDA TEM DÚVIDAS? VAMOS FAZER UM BREVE EXERCÍCIO

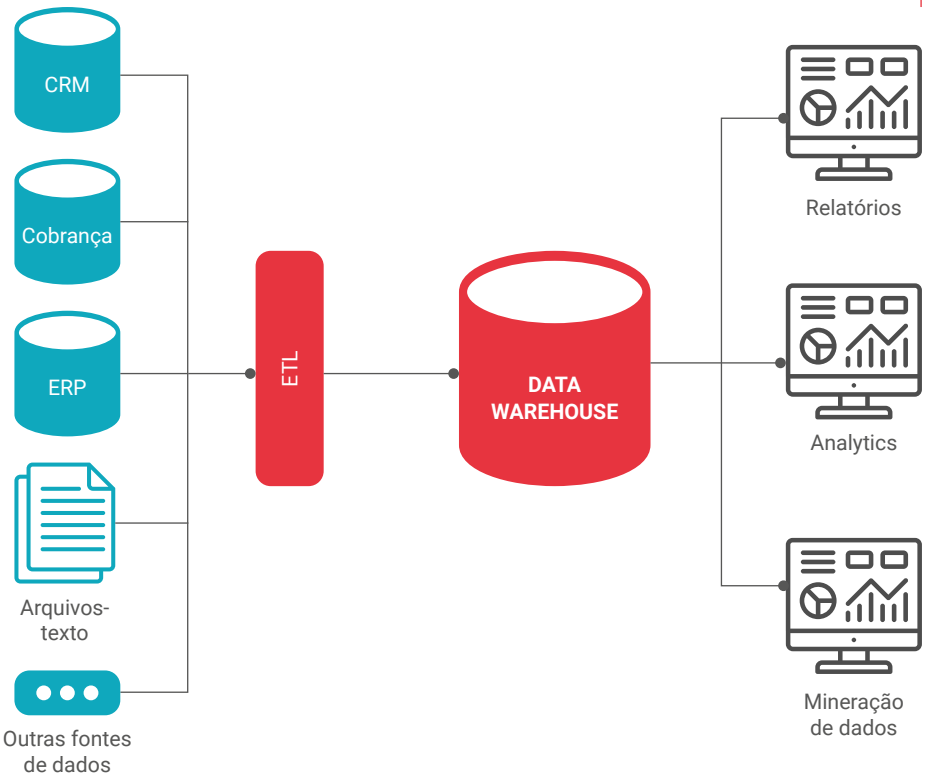
Imagine que sua organização conta com aproximadamente 500 pessoas trabalhando espalhadas por 2 bairros ou cidades diferentes, e que elas precisam trocar informações entre si. Além disso, a organização deve disponibilizar dados e relatórios para o público geral. Qual a melhor solução para esse caso?

Podemos fazer uma combinação. O armazenamento em rede pode ser eficiente para a troca de informações entre departamentos, sem expor esses dados na internet. Já quanto aos dados e relatórios a serem publicados, podemos utilizar uma solução em nuvem para armazenar as informações e disponibilizá-las por meio de um portal da transparência.

Armazém de dados: por que precisamos de um?

Um banco ou base de dados são conjuntos de arquivos relacionados entre si que podem ter registros sobre pessoas, lugares ou coisas. São coleções organizadas de dados que se relacionam de forma a criar algum sentido (informação) e proporcionar mais eficiência durante uma pesquisa.

Já um armazém de dados (*data warehouse*) é um banco de dados de alta capacidade que integra informações de diversos outros bancos menores, criando uma coleção de dados orientada por assuntos e variante no tempo, e com objetivo de dar suporte aos processos de tomada de decisão. Usando esse tipo de banco, é possível elaborar relatórios, criar investigações, montar dashboards, além de fazer análises preditivas.



E por que organizações precisam de um armazém de dados? Decisões baseadas em dados fragmentados, obtidos pelos sistemas tradicionais, não oferecem uma informação consistente, caso não exista uma forte integração entre eles. A imagem acima mostra como se posiciona um armazém de dados perante outros bancos de dados e storages, e como a informação chega para o usuário final.

Os dados, antes de serem armazenados, são transformados (filtrados, normalizados, reorganizados e sumarizados) para que constituam uma base confiável e íntegra.

Os benefícios de um *data warehouse* incluem:

- Manter o histórico de dados;
- Integrar os dados de vários sistemas, permitindo uma visão consolidada de toda a operação, principalmente quando a organização possui várias áreas com sistemas diferentes;
- Melhorar a qualidade dos dados, criando uma padronização de códigos e descrições;

- Apresentar informações de forma consistente;
- Fornecer um único modelo de dados para toda a organização, independente da fonte;
- Reestruturar os dados para melhorar o desempenho de consulta, sem afetar os sistemas em operação.

UM CASO PRÁTICO DE ARMAZÉM COM PUBLICAÇÃO DE DADOS: DATA.RIO

O Instituto Pereira Passos (IPP) já havia lançado, em 2001, a publicação de um "Armazém de Dados" para dar transparência a estatísticas, mapas, estudos e pesquisas com foco na cidade do Rio de Janeiro. Em 2017, o Armazém de Dados foi reformulado e passou a ser chamado Data.Rio. Há dados estruturados em diversas categorias temáticas e diferentes periodicidades de publicação, inclusive em tempo real (por exemplo, transporte). Conheça: **https://www.data.rio/**

Vimos, neste capítulo, que organizar os dados e mantê-los íntegros e disponíveis é uma etapa essencial para seguir na trilha de publicação. Seja por meio de um armazém de dados ou por outra solução, será possível utilizar as informações de maneira estratégica, distribuindo-as internamente em painéis para gestores de diferentes áreas ou dando a máxima transparência para dados públicos. E o mais importante: mantendo a consistência e integridade das informações, que devem ser as mesmas para todas as pessoas.

PARA SABER MAIS

INFOGRÁFICO "Do bit ao Yottabyte: conheça os tamanhos dos arquivos digitais". Acesse: http://bit.ly/bit-yotta

ARTIGO "O que é armazenamento de dados e qual sua importância?", de Denis Zeferino. Acesse: http://bit.ly/armaz-dados

CAPÍTULO VI

PROCESSANDO DADOS COM FERRAMENTAS DE TIPO ETL (EXTRACT, TRANSFORM, LOAD)

Bernardo Loureiro

Você já se viu tendo que baixar várias planilhas de um site, talvez todo dia ou toda semana? Depois, já teve que fazer processos repetitivos, como abrir as planilhas e juntá-las em uma só? Se sim, você já fez uma espécie de ETL, mesmo que de forma simples e manual.

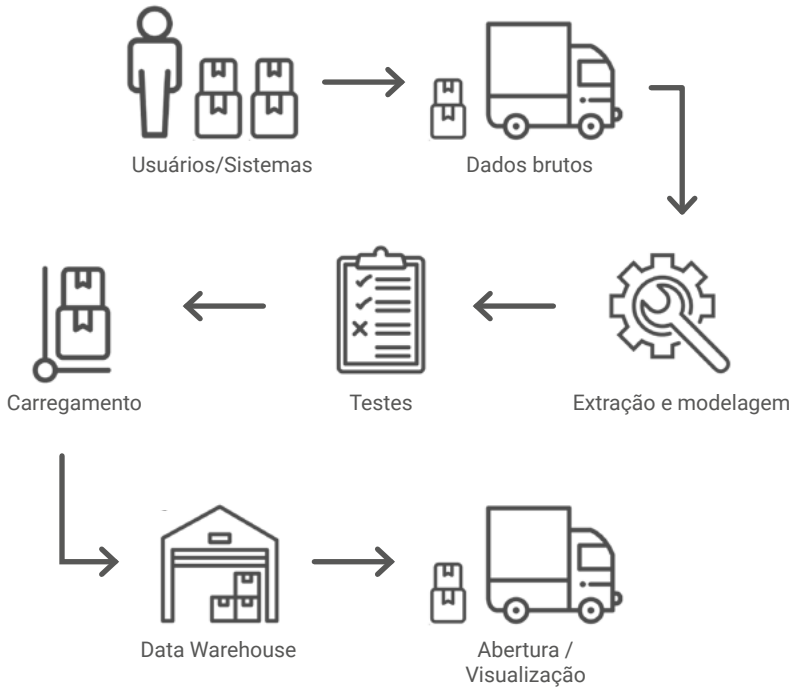
À primeira vista, processos de ETL (*extract, transform and load*, ou extrair, transformar e carregar) podem parecer complexos. No entanto, mostraremos neste capítulo que isso também pode ser muito simples. Aqui vamos apresentar conceitos-chave, ferramentas e exemplos concretos do uso de ETL para publicação de dados.

O que é ETL e para que serve?

ETL é um processo de copiar dados de uma ou mais fontes (*extract*) para um ou mais destinos (*load*) e fazer alterações como limpeza ou validações durante o processo (*transform*).

Podemos usar um processo de fabricação como analogia. Imagine que você tem uma fábrica na qual entram matérias-primas. Elas são transportadas, processadas e passam por um controle de qualidade, entre outras etapas, até chegar a um produto final.

Em um processo de ETL, sua matéria-prima são dados, gerados por várias pessoas e sistemas. Eles podem estar em diversos formatos, como arquivos de texto, planilhas, bancos de dados, ou até imagens. Um ETL poderá automaticamente extrair esses dados, processá-los, e carregá-los em um banco de dados dedicado à análise e à visualização. Esse processo poderá acontecer toda semana, todo dia, ou toda hora, dependendo da frequência com que os dados mudem e das necessidades de análise.



ETL como um processo de fabricação. AUTORIA DOS ÍCONES: Justin Blake, U; DinosoftLab; Sophia; monkik; Atif Arshad, AE; Counloucon (Noun Project). Licença: CC-BY

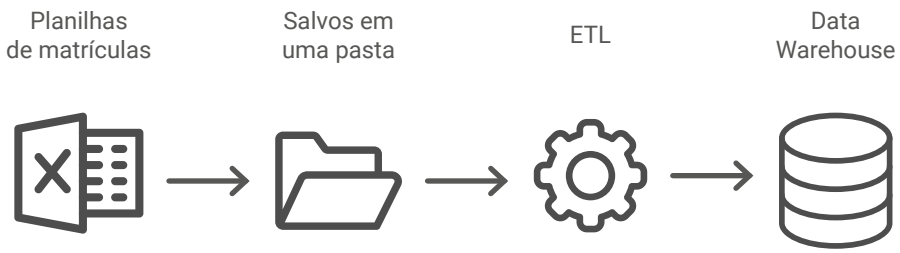
Um exemplo de ETL

Vamos demonstrar com um exemplo concreto. Imagine que estamos em um órgão de Educação. Nesse órgão, são feitas matrículas em escolas toda semana, e essas matrículas ficam registradas em uma série de planilhas.

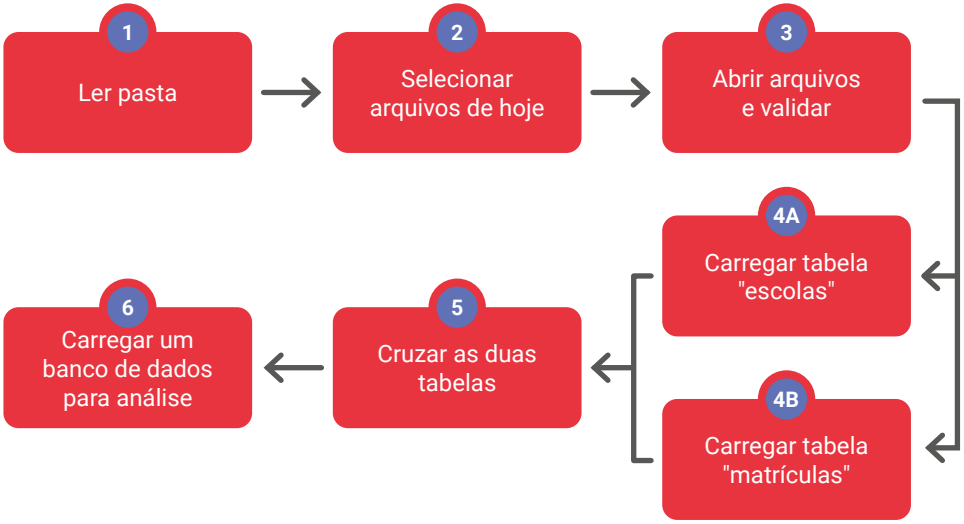
Alguém da equipe de gestão deseja saber a média de matrículas na semana passada por tipo de escola, mas as informações sobre as escolas ficam em outra planilha. Extrair essa informação demandaria trabalho manual significativo, e o pior: na próxima semana, esse trabalho precisaria ser refeito para se ter o número atualizado.

É nesse tipo de situação que um ETL pode economizar tempo e recursos. No diagrama abaixo, exemplificamos um ETL para automatizar o processo descrito acima. Esse ETL poderia ser programado para rodar automaticamente toda semana. Assim, o pessoal da gestão poderia acessar o dado atualizado semana a semana, simplesmente consultando o banco de dados de análise que fica na ponta final do processo.

FLUXO DE DADOS



ETAPAS DO ETL



CRIANDO UM ETL NO SEU COMPUTADOR

Muitas vezes, processos de ETL podem intimidar devido a sua complexidade. Por isso, é bom lembrar que é possível também criar ETLs mais simples, utilizando apenas seu próprio computador e alguns arquivos de dados.

Neste repositório[1], apresento um ETL simples em Airflow, utilizando Python para tratar dados do Sistema de Atendimento ao Cidadão da Prefeitura de São Paulo. Este outro tutorial[2] também utiliza o Airflow e ensina a dar os primeiros passos para rodá-lo no seu próprio computador.

Alertas, validações e publicação

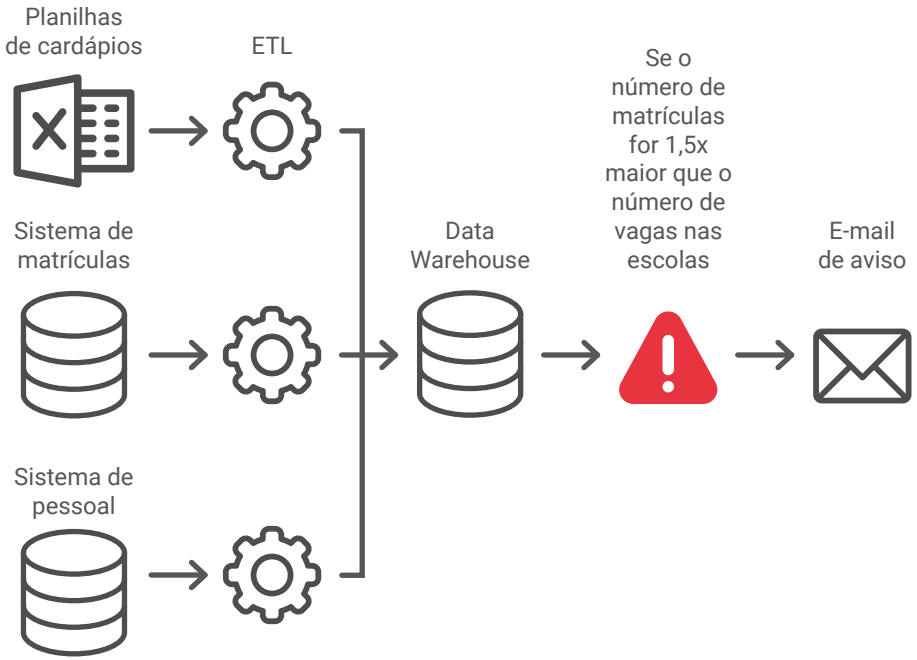
Processos de ETL não servem apenas para extrair dados de uma fonte e carregá-los em outro local. Um ETL também pode ser utilizado para executar verificações de rotina em dados, além de alertar as pessoas usuárias quando houver um erro.

Por exemplo, essas são algumas das ocorrências que podem ser programadas para gerar alertas de rotina:

1 Acesse: http://bit.ly/sacpmsp

2 Acesse: http://bit.ly/etl-airflow

- Quando há **problemas de carregamentos dos dados.** Por exemplo, se um ETL busca em uma pasta por algumas planilhas, um erro humano pode ter gerado uma planilha mal formatada, ou salva no formato errado, o que pode ativar o alerta.
- **Se um sistema não está acessível**, como um banco de dados que está offline ou uma pasta com arquivos faltando em determinado servidor.
- Quando **algo muda inesperadamente nos dados,** como uma mudança retroativa nos dados de um período anterior.
- Quando **um dado passa ou cai abaixo de um certo valor.** Imagine uma tabela com o número de matrículas e vagas de um órgão de Educação. Se o número de matrículas for 1,5 vez maior do que o de vagas, por exemplo, podemos imaginar que há algum erro, e assim disparar um alerta.



Além disso, alertas não precisam ser disparados apenas em caso de erro. Podemos desejar, por exemplo, que algumas pessoas recebam um relatório no começo de toda semana contendo os dados da semana anterior. A produção e o envio desse relatório podem ser automatizados em um processo de ETL.

Por fim, processos de abertura de dados podem ser incluídos também em um ETL. É possível automatizar a geração de planilhas de dados e seu upload para portais de dados abertos utilizando portais que dispõem de APIs para acesso, como o CKAN (repositório para publicação de dados sobre o qual você verá mais adiante, no capítulo 8).

Algumas ferramentas de ETL

Existe uma grande disponibilidade de ferramentas de ETL, muitas das quais são de código aberto e gratuitas. Abaixo listamos algumas dessas ferramentas e quais as principais diferenças entre elas.

Uma das diferenças mais importantes diz respeito à infraestrutura. Muitas empresas fornecem ferramentas que vêm necessariamente acopladas à infraestrutura, oferecidas como um serviço. Isso significa que a empresa cuida também da hospedagem.

Embora isso possa trazer praticidade, por dispensar a necessidade de a organização configurar sua própria infraestrutura, tal modelo traz o risco de *lock-in*, ou seja, de se ficar preso ao serviço daquela empresa, uma vez que será difícil mudar seu ETL para outra empresa ou para a própria infraestrutura. Já outras ferramentas deixam a infraestrutura nas mãos de quem a usa, o que permite que se mude o provedor dela com maior facilidade, apesar de isso gerar um custo maior de configuração dessa infraestrutura inicialmente.

FERRAMENTA	CÓDIGO ABERTO	CUSTO	CONFIGURAÇÃO	INFRA-ESTRUTURA
Apache NiFi	Sim	Gratuito (exceto custo de infra)	Interface gráfica	Da organização
Apache Airflow	Sim	Gratuito (exceto custo de infra)	Python	Da organização
AWS Glue	Não	Variável com uso	Python e Scala	Parte do serviço
Google Dataflow	Não	Variável com uso	Java e Python	Parte do serviço

IMPORTANTE

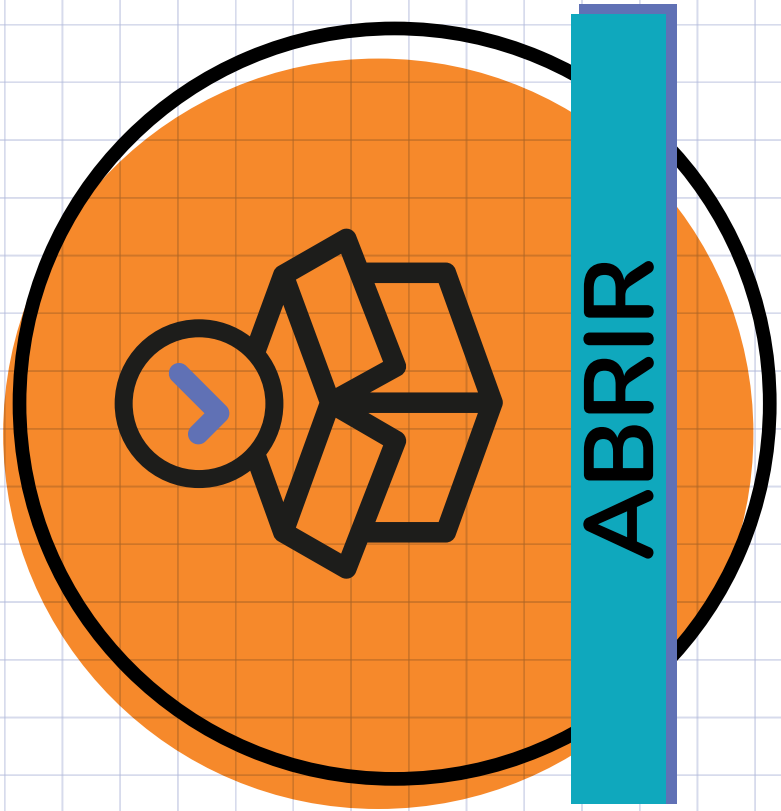
O campo da engenharia de dados, especializado em processos de ETL, tem mudado significativamente nos últimos anos. A introdução de volumes de dados maiores e de dados mais variados, a redução do custo de armazenagem e o aumento do poder computacional têm causado uma revolução na área. Por esse motivo, é importante ter em mente que nenhuma ferramenta irá, por enquanto, apresentar sozinha uma solução definitiva aos problemas de ETL.

PARA SABER MAIS

POST "ETL: o que é e qual sua importância?". Acesse: http://bit.ly/etl-imp

POST (em inglês): "ETL: Understanding It and Effectively Using It", de Punit Pathak. Acesse: http://bit.ly/etl-hp

POST (em inglês): "Functional Data Engineering: A Modern Paradigm For Batch Data Processing, de Maxime Beauchemin. Acesse: http://bit.ly/data-enge


ABRIR

CAPÍTULO VII

APRESENTANDO OS DADOS: DATAVIZ E DASHBOARDS

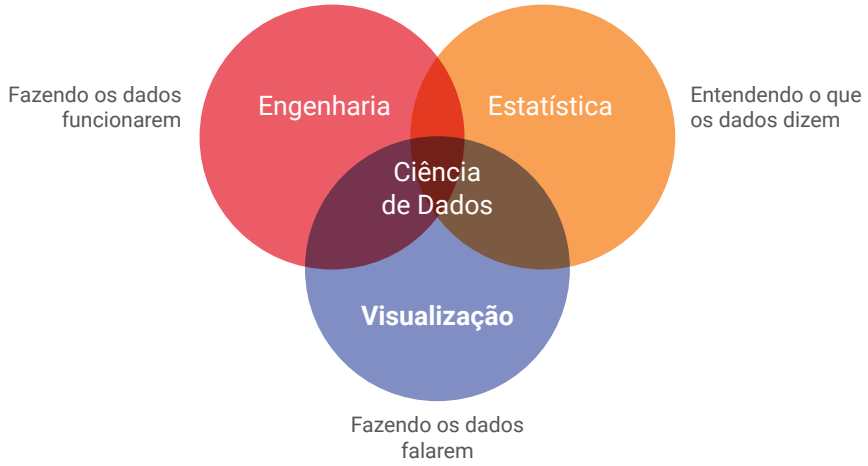
Bernardo Loureiro

A esta altura do guia, já avançamos bastante em nossa jornada de publicação dos dados. A etapa que veremos agora é, muitas vezes, considerada como um adendo, uma cereja do bolo ao final do processo. No entanto, saber apresentar bem os dados é um passo fundamental dessa jornada. **Pouco adianta obter os dados, organizá-los, limpá-los e prepará-los se não conseguirmos comunicar informações relevantes a partir deles.** E esse papel cabe justamente à visualização de dados.

Neste capítulo, vamos explicar alguns conceitos básicos sobre a visualização de dados. Aliados a exemplos práticos, esses conceitos irão ajudar a entender o que faz uma visualização ser bem ou malsucedida, e a tomar decisões na hora de fazer suas visualizações.

Qual o papel da visualização de dados?

O campo da ciência de dados envolve uma série de habilidades, uma das quais é a visualização. Se a engenharia de dados trata de fazer os dados funcionarem (disponibilidade, limpeza, formatação) e a estatística permite entender o que está contido nos dados (correlações, amostragem, testes), a visualização de dados trata de **fazer os dados falarem.**



FONTE: Adaptado do post "The Data Science Venn Diagram", de Drew Conway[1].

[1] Acesse: http://bit.ly/dsvdiagram

Em outras palavras, ao produzir visualizações, nossa responsabilidade está em permitir que pessoas consigam compreender algo a partir dos dados, produzindo conhecimento e, muitas vezes, possibilitando decisões e tomada de ação.

O público que irá consumir nossas visualizações pode variar bastante (se possível, tenha em mente o seu público em específico). Mas, de modo geral, há algumas características que podem nos ajudar a produzir melhores visualizações.

Características de uma pessoa usuária "típica" de visualizações de dados

1. **Tem pouco tempo disponível** para entender o que está sendo mostrado. Um motivo para sermos efetivos e certeiros nas nossas visualizações.
2. Pode ser especialista no assunto que está sendo tratado, mas **muitas vezes não conhece bem os dados ou nunca teve acesso a eles**. Por isso, é importante lembrar que certos aspectos dos dados, que podem ser óbvios para você, não o são para o seu público, e talvez precisem ser reforçados.
3. **Tem seus próprios vieses e pré-concepções**, tanto sobre o assunto da visualização quanto sobre visualizações em si. Por exemplo, alguém pode acreditar que já entende as causas de um determinado fenômeno, ou então pode tender a associar cores verdes com valores "bons" e cores vermelhas com valores "ruins".

Se você tiver esses pontos em mente, e seguir os princípios apresentados a seguir, terá boas chances de produzir visualizações bem-sucedidas.

O QUE FAZ UMA VISUALIZAÇÃO SER BEM-SUCEDIDA? VEJA UM EXEMPLO

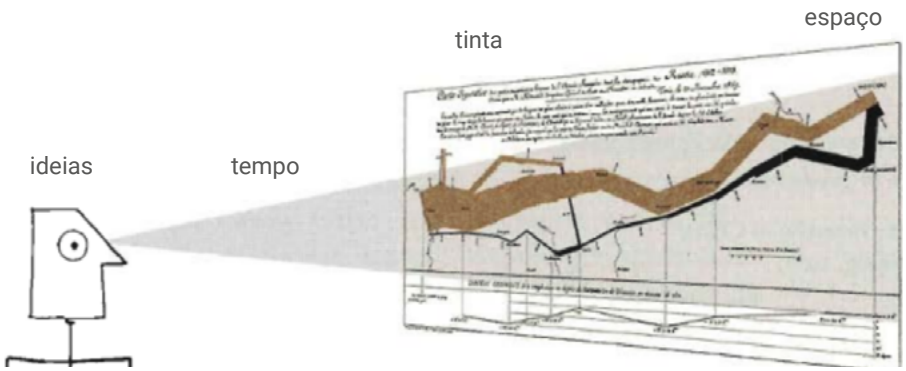
Como você contaria 200 anos de história de 200 países em apenas 4 minutos? É isso que Hans Rosling busca fazer em um vídeo[2] produzido para a BBC Four. Sua aliada? Uma ótima visualização de dados, que utiliza conscientemente muitos dos recursos gráficos disponíveis, como posições, tamanhos, cores e animações.

Princípios da visualização: a excelência gráfica

No seu livro "The Visual Display of Quantitative Information", o estatístico Edward Tufte apresenta um conceito muito útil para pensar visualizações: o **princípio da excelência gráfica**.

Para Tufte, excelência gráfica trata de comunicar **"o maior número de ideias, no tempo mais curto, com o mínimo de tinta e no menor espaço".**

[2] Acesse: http://bit.ly/rosling200



FONTE: *Adaptado do livro "The Visual Display of Quantitative Information", de Edward Tufte (2001).*

Em outras palavras, o princípio trata de maximizar o "retorno" das visualizações com uma maior economia de recursos. Isso é importante pois nos força a lembrar que todos os elementos gráficos de uma visualização são importantes: desde as cores, fontes, tamanhos e posições, até o espaço em branco que deixamos.

Sempre que estamos produzindo alguma visualização, devemos parar e nos fazer as seguintes perguntas: Será que estou oferecendo um maior número de ideias ao meu público? Será que este elemento gráfico, esta cor, ou este texto está adicionando algo? Alguma coisa poderia ser removida ou simplificada? Essas perguntas irão auxiliar a refinar nosso trabalho.

Tipos diferentes de gráficos para cada situação

A tabela abaixo contém alguns tipos de gráficos mais comuns, e qual a situação mais recomendada para se usar cada um deles.

Vale ressaltar que há poucas regras absolutas na visualização de dados, do tipo "se X, faça Y", ou então "nunca faça Z". Por esse motivo, trate a tabela a seguir como uma série de recomendações e, quando em dúvida, sempre se apoie nos princípios da visualização para tomar uma decisão.

TIPO DE GRÁFICO	USO E RECOMENDAÇÕES
Números e tabelas	• Destacar números importantes • Quando houver muitos números, usar cores de fundo neles para mostrar padrões (“table heatmap”)
Barras	• Horizontais para comparar categorias • Verticais para comparar variáveis ordinais • Não cortar o eixo vertical
Linhas	• Mostrar valores ao longo do tempo • Pode-se cortar o eixo vertical para enfatizar uma diferença
Pizza	• Comparar, no máximo, 4 ou 5 categorias de um total
Dispersão	• Comparar a relação entre duas variáveis

Outra maneira muito importante de melhorar suas habilidades de visualização de dados é ampliando o seu repertório. Ao final deste capítulo, listamos algumas referências sobre o tópico; procure utilizá-las como inspiração na hora de produzir suas visualizações.

Produzindo painéis interativos, ou “dashboards”

Painéis interativos de dados, ou “dashboards”, são uma maneira muito popular de visualização. Infelizmente, é muito comum encontrar dashboards que não são efetivos, muitas vezes devido à falta de concisão e de objetividade.

Uma definição de dashboards é a de “interfaces gráficas que exibem **indicadores essenciais** sobre algum **assunto**”. É frequente ver dashboards que apresentam indicadores demais ou que não têm um assunto bem definido e, assim, acabam produzindo mais confusão do que informando.

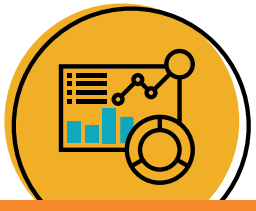
Na tabela abaixo, listamos algumas ferramentas populares usadas no desenvolvimento de dashboards.

FERRAMENTA	CÓDIGO ABERTO	CUSTO	CONFIGURAÇÃO E USO	INFRA-ESTRUTURA	PRÓS
Metabase	Sim	Gratuito	Consultas via interface gráfica ou SQL	Do usuário (possui versão paga em nuvem)	Facilidade de uso, custo, comunidade
Apache Superset	Sim	Gratuito	Consultas via interface gráfica ou SQL	Do usuário	Custo, interatividade dos painéis
Looker	Não	Sob demanda	Consultas via interface gráfica; modelagem de dados em LookML	Parte do serviço	Visualizações sofisticadas, camada de modelagem
Power BI	Não	Dez dólares por mês por usuário (plano Pro)	Consultas via interface gráfica ou SQL; modelagem de dados em DAX	Parte do serviço	Camada de modelagem, preço relativamente baixo
Google Data Studio	Não	Gratuito	Consultas via interface gráfica; possibilidade de integração com várias fontes de dados	Parte do serviço	Custo, facilidade de uso

É útil aqui relembrar um exemplo emblemático de dashboard: o painel de instrumentos. Se observarmos um painel de carro atentamente, veremos que ele tem apenas os indicadores essenciais e que auxiliam a conduzir o veículo com segurança. Além disso, esses indicadores possuem uma hierarquia clara, construída a partir de elementos gráficos como cores, tamanhos e posições.

Ao construir dashboards, lembre-se de se ater aos princípios da visualização, de apresentar apenas os indicadores essenciais e de ter um assunto claro e bem definido.

FONTE: Wikipédia, usuário Soupeurfaive.



UM DASHBOARD BEM-SUCEDIDO TEM UM TEMA CLARO

O site[3] contendo os dados oficiais do tráfego na web de centenas de páginas do governo dos EUA é um bom exemplo de dashboard. Observe o uso consciente de cores, tamanhos, e textos. Veja quantas ideias você consegue absorver dele em um curto período.

[3] Acesse: http://bit.ly/analyt-usa

PARA SABER MAIS

BLOG Categoria “Dataviz” do datavizbr no Medium. Acesse: http://bit.ly/datavizbr

BLOG (em inglês): Categoria “Dataviz” do Nightingale no Medium. Acesse: http://bit.ly/dvnightingale

SITE (em inglês): Designer Lisa Charlotte Rost. Acesse: http://bit.ly/lisarost

POST “Planning Your Dashboard”, da Gooddata. Acesse: http://bit.ly/plann-dashb

LIVRO (em inglês): “The Visual Display of Quantitative Information”, de Edward Tufte (2001)

CAPÍTULO VIII

DOCUMENTANDO E CONECTANDO DADOS

Vitor Baptista

Até aqui já preparamos os dados e criamos visualizações para que as pessoas usuárias possam explorá-los. Mas e os dados brutos, qual a melhor forma de publicá-los? Este capítulo fala sobre as duas principais formas de distribuição de dados brutos: download e API. E também falaremos sobre repositórios de dados e documentação.

Formas de distribuição de dados

As formas mais comuns de distribuição de dados são a disponibilização para download e APIs. São formas complementares, com características e objetivos diferentes.

O **download de dados** é a forma mais simples, mas também a mais importante, pois muitas análises comuns, como o cálculo de médias e desvios-padrão, necessitam do acesso aos dados completos.

A maioria dos dados deve ser disponibilizada por download, mesmo os que também possuem APIs. As exceções são conjuntos de dados com um volume muito grande e/ou com atualização muito frequente. Mesmo nesses casos, pense em formas de viabilizar o download. Por exemplo, usando torrents para distribuir arquivos muito grandes, ou publicando o histórico dos dados de conjuntos que são atualizados muito frequentemente (como a localização dos ônibus).

Já a **API**, sigla para **Application Programming Interface**, é uma interface de comunicação. As APIs permitem que uma pessoa usuária de um sistema ("cliente") acesse diretamente dados ou serviços de um órgão ("servidor"). Existem dois tipos de APIs: as de consulta e as de modificação de dados.

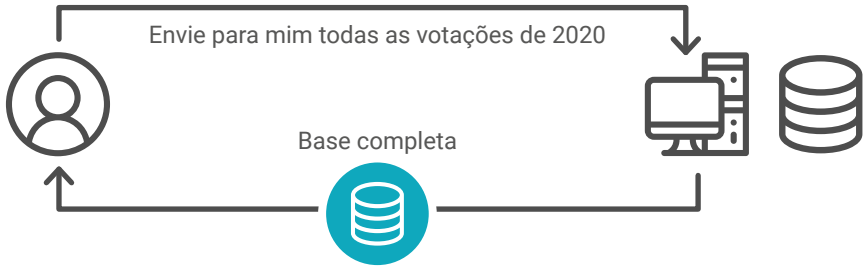
As **APIs de consulta** permitem que seja acessado um dado específico. Por exemplo, a Câmara dos Deputados disponibiliza uma API que retorna o resultado da votação de um projeto de Lei.[1]

[1] Acesse: http://bit.ly/api-camara

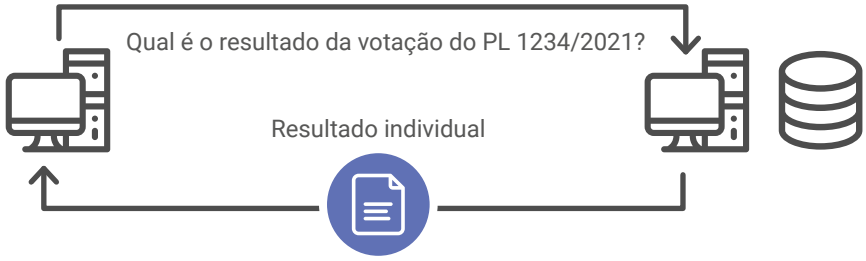
Já as **APIs de modificação de dados** proveem acesso a serviços de determinado órgão. Por exemplo, a Polícia Federal oferece o agendamento de emissão de passaportes por meio de seu website. Esse serviço também poderia ser oferecido através de uma API. Assim, seria possível criar um software que, usando essa API, oferecesse esse serviço de agendamento.

As APIs de consulta são comparáveis ao download dos dados. A forma do acesso aos dados é diferente, mas o objetivo final é o mesmo: consultar os dados. Já as APIs de modificação oferecem acesso a serviços, o que não é possível por meio de um simples download.

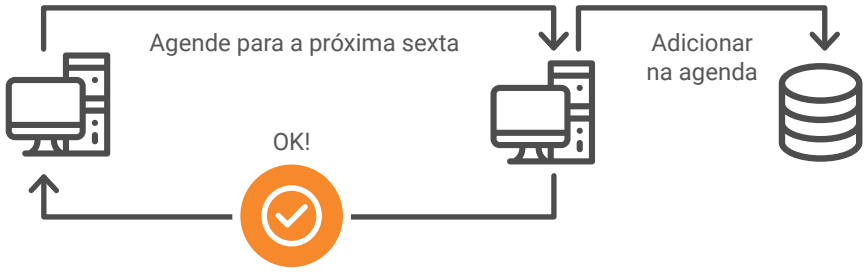
DOWNLOAD



API CONSULTA



MODIFICAÇÃO



Quando criar APIs?

No geral, os conjuntos de **dados mais propícios a serem disponibilizados via APIs** têm uma ou mais das seguintes características:

- **Grande volume de dados**, pois assim é possível consultar só a parte dos dados que interessa. Exemplo: imagens de satélite.
- **Alta taxa de atualização**. Como a consulta é feita diretamente ao servidor, os dados sempre estarão atualizados. Exemplo: a localização atual dos ônibus.

As APIs também podem oferecer funcionalidades mais complexas, como encontrar a rota do ponto A ao B usando transporte público, ou ainda permitir que serviços de um órgão, como o agendamento de consultas, sejam acessados por softwares.

Esses critérios podem te guiar, mas o mais importante é escutar as pessoas que usam os dados. Existe demanda pela API que planeja desenvolver? Além dos custos de criação, APIs também têm custos de manutenção. Por isso, conheça a demanda para aumentar a chance de que o investimento feito tenha bom retorno.

Repositórios de dados

Repositórios são sistemas para publicação, organização, busca e distribuição de dados. Eles auxiliam os publicadores a documentar e a organizar os dados, e auxiliam também quem os usa, oferecendo ferramentas de busca e pré-visualização.

O **CKAN** é o sistema de repositórios de dados mais usado mundialmente. Ele é gratuito e tem seu código livre, o que significa que, caso necessário, sua equipe poderá modificá-lo sem depender de terceiros, garantindo sua soberania tecnológica. Ele foi criado em 2006 pela Open Knowledge Internacional, e é usado em mais de 40 países, inclusive o Brasil, Estados Unidos e Reino Unido. Há quase 200 instâncias do CKAN mapeadas pelo mundo, sendo usadas por diversos tipos de organizações[2].

Quer conhecer melhor esse sistema e seu funcionamento a partir da visão de um publicador de dados? Assista a duas videoaulas sobre o tema, do curso Publicadores de Dados[3].

Documentação

Independente de como os dados sejam distribuídos, as pessoas que os usam precisarão compreendê-los. Para isso, uma boa documentação é primordial. Ela deve descrever o conteúdo dos dados, sua abrangência, metodologia de coleta e limitações. Essas informações que servem para descrever os dados são chamadas de **metadados**.

Podemos imaginar as partes da documentação como um funil, partindo do mais geral ao mais específico. Em primeiro lugar, descreva o conjunto de dados de modo geral, com atributos como:

[2] Conheça aqui os exemplos de repositórios pelo mundo que utilizam o CKAN: http://bit.ly/ckan-inst

[3] Aula: "O que é o CKAN?". Acesse: http://bit.ly/ckan-1
Aula: "Usando o CKAN". Acesse: http://bit.ly/ckan-2

Exemplo:

Título	Unidades Básicas de Saúde em João Pessoa - PB
Descrição	Lista das Unidades Básicas de Saúde (UBS) de João Pessoa incluindo localização, número de leitos e horários de funcionamento
Fonte	Secretaria de Saúde de João Pessoa
Responsável técnico	Joana Silva <j.silva@saude.joaopessoa.pb.gov.br>
Última atualização	15/12/2020
Frequência de atualização	Semestral
Versão	1.3
Licença	Creative Commons Zero

Essas informações permitem que se decida rapidamente se o conjunto de dados é relevante ou não para seus objetivos. Depois disso, descreva cada arquivo, com os atributos:

Título	Unidades Básicas de Saúde
Descrição	Localização e informações sobre as UBS
Última atualização	10/12/2020

E então descreva a estrutura do conteúdo dos arquivos. Nesse passo, o que deverá ser descrito depende do tipo de arquivo. Por exemplo, um mapa exige informações diferentes de uma tabela. Já para documentar dados tabulares, como CSVs e arquivos do LibreOffice ou Excel, crie um dicionário de dados.

Dicionários de dados

Dicionários de dados são uma das partes mais importantes da documentação, descrevendo a estrutura de uma tabela. Para cada coluna presente em sua base de dados, descreva:

Título	codigo_ibge
Descrição	Código IBGE do município
Tipo	Texto, número, data etc.
Restrições	• O campo é obrigatório ou opcional? • Se for uma data, qual o formato? (exemplo: 31/12/2020 ou 2020/12/31) • Se houver códigos, quais seus significados? (exemplo: “U - Leito de UTI”, “E - Leito de Enfermaria”)

Defina também quais são os **identificadores** – conjunto de uma ou mais colunas que identifica unicamente uma entidade na tabela. Por exemplo, o identificador de uma lista de servidores públicos pode ser o CPF (ou parte dele, para evitar publicar dados pessoais).

Os identificadores devem ser:

- Únicos;
- Não-vazios; e
- Imutáveis.

Além dos identificadores das entidades do próprio conjunto de dados, descreva também aqueles que referenciam outras tabelas. Por exemplo, uma lista das Unidades Básicas de Saúde pode conter uma coluna com o código IBGE do município, permitindo que esses dados sejam cruzados com outros conjuntos que usem esse mesmo código.

Materiais suplementares

Por fim, inclua documentos como:

- Glossário;
- Conjuntos de dados relacionados;
- Metodologia de coleta e codificação;
- Limitações conhecidas;
- Exemplos de uso; e
- Dúvidas frequentes.

E quaisquer outros materiais que auxiliem na compreensão dos dados.

Formatos de documentação

A documentação pode ser escrita em diversos formatos, como:

- Arquivos de texto;
- Repositórios de dados;
- Documentação legível por máquina:
  - Data Packages;
  - OpenAPI / Swagger.

O mais simples é documentar em arquivos de texto, pois eles só exigem um editor de texto. A desvantagem dessa forma é a ausência de um padrão global, com cada órgão podendo documentar de um modo diferente. Caso escolha esse formato, crie um padrão e publique em formatos abertos, como PDF e TXT.

Já os repositórios de dados oferecem mais estrutura, por definirem formulários com as informações necessárias para publicação. Esses formulários guiam o trabalho de documentação, definindo campos obrigatórios e opcionais. Além disso, como o repositório conhece o conteúdo de cada campo (título, descrição, autor etc.), ele pode permitir a busca baseada em um desses atributos, como encontrar os conjuntos de dados de um determinado autor.

Por último, as documentações legíveis por máquina definem um formato padrão para a documentação. Assim, um software que entenda esse formato pode prover funcionalidades baseadas nela. Por exemplo, se a documentação define que uma coluna contém datas no formato Dia/Mês/Ano, um software pode validar os dados confirmando que isto é verdade.

Para conjuntos de dados, use o padrão Data Package. Já para APIs, use o OpenAPI (também conhecido como Swagger).

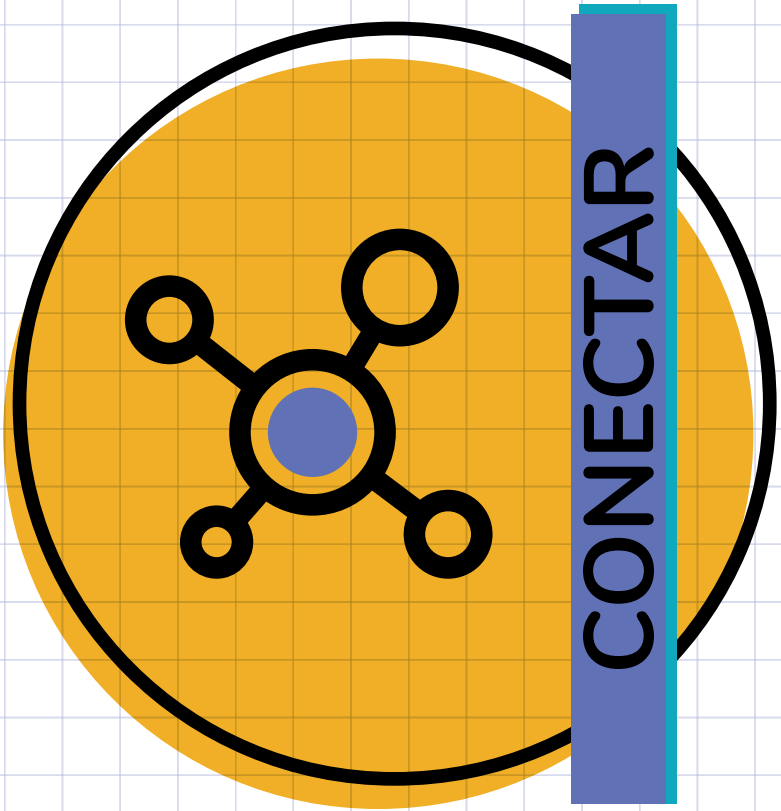
Como vimos neste capítulo, a disponibilização e documentação de dados são processos iterativos. Primeiro, publique os dados para download. Depois, dependendo da demanda, planeje publicar também via APIs. Aproveite as dúvidas e erros relatados para melhorar a documentação e o formato de publicação dos dados.

PARA SABER MAIS

PADRÃO Data Packages (em inglês), da Open Knowledge Foundation.
Acesse: http://bit.ly/data-pack

PROJETO Frictionless Data (em inglês), da Open Knowledge Foundation.
Acesse: http://bit.ly/frict-data

GUIA "Modelo de Maturidade de Dados Abertos", do Open Data Institute.
Acesse: http://bit.ly/maturidade-guia


CONECTAR

CAPÍTULO IX

PARTICIPANDO DO ECOSSISTEMA DE DADOS ABERTOS

Haydée Svab

Parabéns, publicadoras e publicadores, por chegarem à etapa final da nossa trilha! Depois de pensarmos e exemplificarmos o planejamento, o desenvolvimento e a abertura dos dados, chegou a hora de mergulharmos no ecossistema dos dados abertos, estabelecendo pontos de conexão sólidos com a sociedade.

Ao longo deste guia, vimos que simplesmente disponibilizar dados ao público muitas vezes pode não ser suficiente para torná-los úteis. Cidadãs e cidadãos estão interessados em serviços que facilitem o seu cotidiano, e que podem ser construídos ou melhorados com dados abertos e/ou informações decorrentes deles. Com isso, é necessário estabelecer e ativar, desde o início de seu projeto, um ecossistema de dados abertos que facilite a interação e a comunicação entre as partes interessadas.

Pensando em desenvolver esse ecossistema, as seguintes **estratégias** podem ser empregadas:

1. mapear pessoas e organizações que atuam com o tema (partes interessadas);
2. compreender as relações entre elas;
3. identificar os recursos necessários para cada uma se engajar;
4. estabelecer e acompanhar indicadores que monitoram o ecossistema como um todo.

Podemos enxergar as partes interessadas de várias formas, como considerando sua natureza e como cada uma lida com os dados abertos:

- **Setor público**: o Estado e seu funcionalismo, responsável por fazer a "máquina girar" e manter os serviços públicos disponíveis. Também engloba o governo, responsável pela implementação das diretrizes políticas. São grandes produtores de dados abertos e também de políticas regulatórias sobre os dados e seus graus de privacidade, transparência e abertura.
- **Setor privado**: empresas e negócios, em geral. São grandes consumidores de dados abertos. Mas também podem ser produtores de dados abertos, como no caso das indústrias farmacêuticas que

produzem vacinas para a Covid-19 e que podem disponibilizar seus dados de pesquisa anonimizados, por exemplo.

- **Imprensa e mídia**: veículos tradicionais (jornais, revistas, TV) e não tradicionais (mídias sociais e alternativas). Muitas vezes, assumem papel de intermediários entre quem produz e quem consome o dado na ponta. São fundamentais no movimento de combate à desinformação e para conferir maior credibilidade e impacto ao ecossistema.
- **Terceiro setor**: organizações não-governamentais, institutos, associações e movimentos sociais (sociedade civil organizada). São principalmente consumidores de dados, embora, cada vez mais, também produzam dados.
- **Escolas e universidades**: pessoas que estudam e pesquisam nessas instituições. Consomem dados abertos como insumo para geração de conhecimento, disponibilização de informação e avanço da ciência, e assim também geram dados para o ecossistema. Além disso, nas atividades de ensino-aprendizagem, também atuam como intermediários entre oferta e demanda de dados.

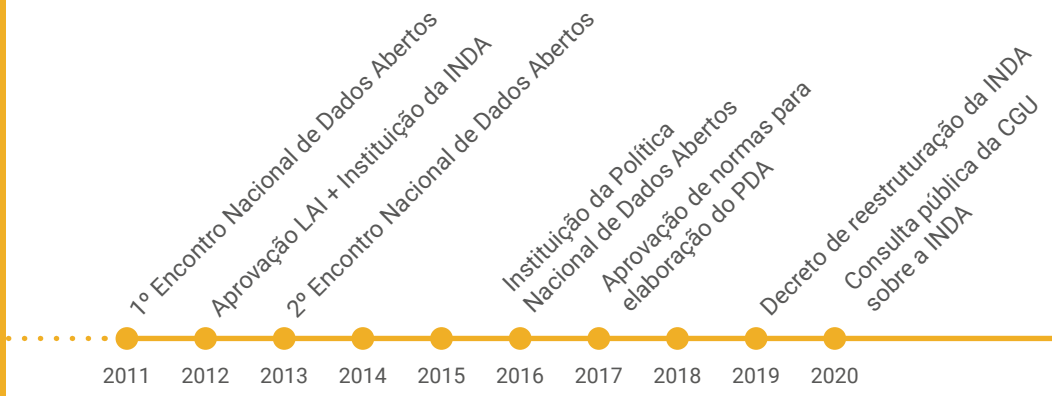
INFRAESTRUTURA DE DADOS ABERTOS

FONTE: Adaptado do artigo (em inglês) "Characterizing Data Ecosystems to Support Official Statistics With Open Mapping Data for Reporting Sustainable Development Goals"[1], de Marc van den Homberg e Iryna Susha (2018).

[1] Acesse: http://bit.ly/data-ecos

Pontos de Conexão – governo

Para que o ecossistema de dados seja ativado e não esmoreça, é preciso que haja canais de comunicação e pontos de conexão que promovam o diálogo e troca de experiências não apenas entre as partes interessadas, mas também internamente. Assim nasceu o **Encontro Nacional de Dados Abertos (ENDA)**, no qual sempre houve participação não só de órgãos de governo, mas também de fornecedores de tecnologias e de entes da sociedade civil.



A partir da linha do tempo acima, observa-se que, entre o 1º ENDA, em 2011, e o 2º ENDA, em 2013, dois fatos importantes ocorreram: a publicação da LAI (Lei de Acesso à Informação Pública) e a instituição da Infraestrutura Nacional de Dados abertos (INDA)[2], bem como do seu comitê gestor. É notável que a existência de espaços de troca como o ENDA não apenas estabeleceu diálogos no setor público e deste com as demais partes interessadas, como também fortaleceu o ecossistema de dados

[2] A INDA é um conjunto de padrões, tecnologias, procedimentos e mecanismos que dão condições para disseminação e compartilhamento de dados e informações públicas, segundo modelos de dados abertos e padrões de interoperabilidade. Para saber mais, acesse: http://bit.ly/inda-br e http://bit.ly/wiki-inda

abertos e a articulação de políticas e planos, como a **Política Nacional de Dados Abertos**[3]. Vale ressaltar que não é mandatório que esses pontos de conexão sejam de âmbito nacional, já que é possível estabelecer fóruns em diversos níveis: organizacional, municipal, estadual ou regional.

Pontos de Conexão – sociedade

Além de organizar e abrir os dados, é importante mostrar os benefícios disso e encorajar a adoção de políticas de dados abertos em governos, em empresas e na sociedade civil em geral. Nesse sentido, é necessário haver pontos de conexão entre essas partes. Eventos como o Open Data Day e Hackathons ajudam não apenas a divulgar os dados, mas fazem com que as pessoas entendam melhor o processo e se engajem na sua utilização.

- **Open Data Day (ODD)**
  É uma celebração anual em nível mundial que se apoia em dois pilares: dados abertos e nível local de engajamento. É uma oportunidade para mostrar os benefícios dos dados abertos e encorajar a adoção de políticas de dados abertos em governos, em empresas e na sociedade civil. Qualquer pessoa ou organização pode participar de ou promover um evento do ODD na sua cidade.

- **Hackathons**
  São maratonas que podem durar poucos dias ou até semanas, e que reúnem as pessoas hackers (programadores, designers, desenvolvedores e inventores) para criar projetos e resolver problemas. Para que sejam bem-sucedidos, precisam envolver governos, empresas e sociedade civil. Temos como potencialidades dos hackathons: gerar inovações, promover participação social, aumentar a transparência e aproximar os atores entre si. Como principal fragilidade, tem-se a baixa sustentabilidade das soluções desenvolvidas.

[3] Instituída em 2016, a Política define regras para disponibilização de dados abertos governamentais. Acesse: http://bit.ly/pnda-br

A **inovação digital** não é nem pode ser exclusividade das empresas. Ela também precisa existir no setor público e no terceiro setor! Isso é o que nos mostra casos como o do LabHacker[4] da Câmara dos Deputados, que nasceu com o impulso de um hackathon em 2013 e existe até hoje; o do MobiLab[5], que em São Paulo incubou startups de mobilidade, estimulando a criação de soluções abertas para cidades inteligentes; ou o do Porto Digital[6] em Recife, um parque tecnológico que vem viabilizando formas de contratação de inovação digital com muito sucesso.

Canais de Comunicação

Para além dos encontros pontuais, ainda que periódicos, é preciso que haja canais operando em fluxo contínuo para estabelecer conexões e cooperações no ecossistema dos dados abertos. Para isso, existem canais entre cidadãos e a administração pública, entre os quais destacamos:

- **Pedidos de acesso à informação**
  A LAI estabelece diretrizes que possibilitam a qualquer pessoa solicitar informações públicas sem necessidade de justificativa, bastando fazer um pedido de informação pelo SIC (Serviço de Informação ao Cidadão), por meio eletrônico ou presencial.

- **Coleta de feedback**
  Considerando que já se estão publicando bases de dados e promovendo eventos de divulgação, como saber se essas ações estão sendo efetivas? Será que a informação e as formas de entregá-la estão sendo úteis para quem está recebendo? Para saber isso, é importante contar com canais de feedback, como o próprio SIC ou canais exclusivos para isso. O Portal de Dados do governo federal[7],

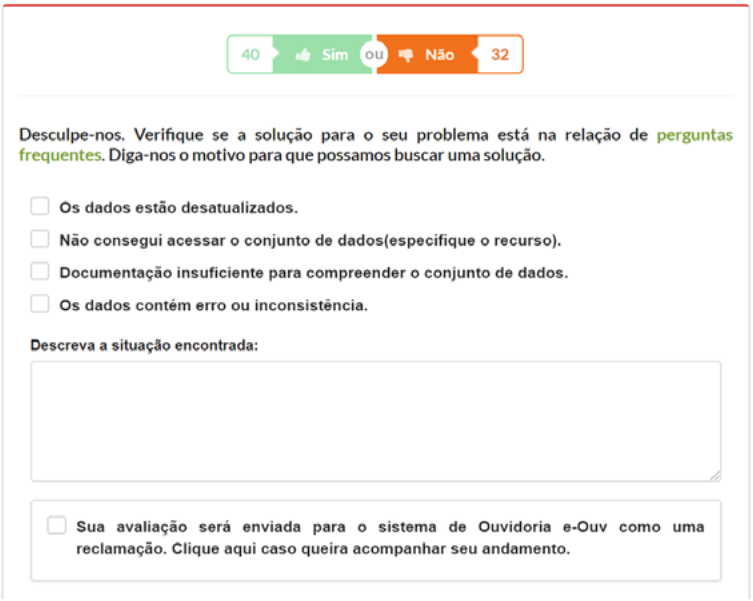
[4] Acesse: http://bit.ly/labhacker-cf

[5] Acesse: http://bit.ly/mobilab-sp

[6] Acesse: http://bit.ly/portodigital-pe

[7] Acesse: http://bit.ly/dados-br

por exemplo, tem 3 formas interessantes de captar (e já classificar) o feedback: (1) dar retorno específico sobre uma base de dados; (2) sugerir novos dados; e (3) dar sugestões mais genéricas. Neste caso, já existem pré-classificações que facilitam a tabulação e uso dessa informação para melhorar a qualidade dos serviços.



Exemplo de feedback específico a partir do acesso a uma base de dados do Portal de Dados Abertos do governo federal

- **Ouvidoria**
  A partir das manifestações das pessoas usuárias, a ouvidoria recebe, analisa, orienta e encaminha questões às áreas responsáveis para que seja feita a apuração e o encaminhamento dos casos.

WIKILEGIS: UMA FERRAMENTA DO PORTAL E-DEMOCRACIA PARA CONSTRUÇÃO COLABORATIVA DE PROJETOS DE LEI

Uma das formas de promover a colaboração é permitir o envio de comentários e sugestões a partir de um texto-base. Com essa ideia, a Câmara dos Deputados desenvolveu o Wikilegis, ferramenta que permite que pessoas contribuam para a construção e aperfeiçoamento de projetos de lei. Resumidamente, o processo conta com 3 momentos: (i) um parlamentar propõe um texto para consulta; (ii) em determinado período, quaisquer pessoas podem comentar trechos do texto, além de opinar sobre comentários de outras pessoas; (iii) o parlamentar pode observar e incorporar as sugestões feitas. A principal vantagem desse tipo de ferramenta é que ela permite que participantes interajam entre si. Dessa forma, uma ideia pode alimentar a outra, ativando assim a "inteligência coletiva" do ecossistema. Por fim, a plataforma exibe os resultados da consulta e permite o download das contribuições em formato CSV, uma boa prática de transparência para processos participativos. Como se trata de uma ferramenta disponível sob licença livre, qualquer organização pode adotá-la. Conheça: **http://bit.ly/wikilegis-br**

PARA SABER MAIS

ARTIGO “Dados Abertos: a retrospectiva de um comitê”, de Augusto Herrmann Batista (2020). Acesse: http://bit.ly/retrospec-inda

WIKI INDA, com documentos e outras referências. Acesse: http://bit.ly/wiki-inda

SITE “Open Data Day”. Acesse: http://bit.ly/odd-br

SOBRE AS AUTORAS E OS AUTORES

Bernardo Loureiro

Urbanista e programador, especializado em mapeamento, análise e visualização de dados. É criador do laboratório Medida SP, no qual realiza pesquisa, consultoria e desenvolvimento para setores público e privado. Foi consultor UNESCO de análise e visualização de dados no Pátio Digital, iniciativa da Secretaria de Educação da Prefeitura de São Paulo, onde ajudou a desenvolver projetos de transparência e abertura de dados. Já elaborou e lecionou cursos para centenas de alunos em instituições como SESC-SP e como Agente de Governo Aberto. Formado em Arquitetura e Urbanismo pela USP e mestre em Desenho Urbano pela Parsons School of Design.

Fernanda Campagnucci

Diretora-executiva da Open Knowledge Brasil. De 2013 a 2019, atuou como gestora pública na Prefeitura de São Paulo, tendo sido responsável pela política municipal de transparência, abertura de dados e integridade na Controladoria Geral do Município, além de ter liderado projetos de tecnologia, inovação e governo aberto na Secretaria Municipal de Educação. Graduada em Jornalismo e mestre em Educação pela Universidade de São Paulo, é doutoranda em Administração Pública e Governo na Fundação Getúlio Vargas (EAESP-FGV). Especialista em Transparência e Accountability pela Universidade do Chile (2014), foi fellow de Governo Aberto da Organização dos Estados Americanos (2015), Líder de Dados Abertos do Open Data Institute (2016) e fellow de governo da Unidade Operacional Governança Digital da Universidade das Nações Unidas, a UNU-EGOV (2018). É professora convidada do Insper nos cursos de Compliance e de Inovação no Setor Público.

Haydée Svab

Cientista de Dados e Pesquisadora em Mobilidade Urbana Atualmente, é CEO da ASK-AR (consultoria em análise de dados), membro do Conselho Deliberativo da AEAMESP (Associação dos Engenheiros e Arquitetos de Metrô) no triênio 2017/2019 e da comunidade Transparência Hacker. Foi consultora do Banco Mundial, cofundadora dos grupos RLadies – São Paulo, PoliGNU e PoliGen. É doutoranda em Smart Cities na Ciência de Computação (IME-USP), mestra em Engenharia e Planejamento de Transportes (Poli-USP), especialista em Democracia Participativa, Repúblicas e Movimentos Sociais (UFMG) e formada em Engenharia Civil/Arquitetura pela USP (Programa Poli-FAU).

Leandro Oliveira

Arquiteto de soluções de big data e analytics na Globo, com anos de experiência em projetos de dados, em especial dados de domínio público (esferas federal, estadual e municipal), sendo responsável por todo o ciclo de vida dos dados: Arquitetura, Engenharia e Ciência de Dados. Seus projetos de domínio público de maior relevância foram: Anda SP e Fora do Ponto – projetos de mobilidade urbana nas cidades do Rio de Janeiro e São Paulo; Eleições 2018 – Apoio na geração de pautas a partir de dados históricos das últimas eleições e apuração ao vivo dos dados; professor voluntário na academia de tecnologia interna, ensinando arquitetura da informação, big data e ciência de dados; e professor de informática básica para terceira idade. Formado em Ciência da Computação e pós-graduado em Engenharia de Software na UFRJ e em Arquitetura Empresarial e Sistemas Corporativos na PUC-RJ.

Natalia Langenegger

Doutoranda em Direito do Estado pela Faculdade de Direito da Universidade de São Paulo (USP); mestra em direito e desenvolvimento pela Escola de Direito de São Paulo da Fundação Getulio Vargas (FGV Direito SP); bacharela em Direito pela Pontifícia Universidade Católica de São Paulo (PUC-SP); e pesquisadora em temas relacionados ao tratamento de dados pessoais pelo poder público. Foi aceita como visiting fellow no Information Society Project, Yale Law School. Foi professora voluntária na Universidade de Brasília (UnB) e fez intercâmbio de mestrado na Universidade de Tilburg – Holanda. Foi pesquisadora na FGV Direito SP, no Internetlab e no Instituto de Pesquisas Econômicas Aplicadas (Ipea). É egressa e foi professora e orientadora na Escola de Formação da Sociedade Brasileira de Direito Público (SBDP).

Vitor Baptista

Mestre em Computação pela UFPB, onde analisou o comportamento de votação dos deputados federais. Trabalha com dados abertos desde 2012. Por 5 anos, foi desenvolvedor na ONG inglesa Open Knowledge Foundation. Nesse tempo, fez parte das equipes de desenvolvimento do CKAN e OpenSpending e, mais tarde, foi o líder técnico do OpenTrials, um repositório aberto de informações sobre testes clínicos, e do Frictionless Data, um conjunto de ferramentas para validação e especificação de conjuntos de dados. Antes disso, trabalhou como consultor na ThoughtWorks em Porto Alegre. Atualmente trabalha como cientista de dados na empresa estadunidense Royalty Exchange, sendo responsável por toda sua infraestrutura de banco de dados analíticos.

Siga em contato conosco na jornada de publicação de dados! Aqui estão alguns canais por meio dos quais você pode continuar em diálogo com outros publicadores de dados:

FÓRUM DA OPEN KNOWLEDGE BRASIL espaço para diálogo sobre dados abertos, licenças livres, open source, ciência aberta e outros temas ligados ao conhecimento livre. **https://discuss.okfn.org/c/local-groups/okbr/76**

FÓRUM DE JORNALISMO DE DADOS comunidade para a discussão de assuntos relacionados ao jornalismo de dados e dúvidas técnicas sobre o trabalho com dados. **https://forum.jornalismodedados.org/**

FÓRUM DADOS ABERTOS um espaço para a comunidade brasileira de dados abertos discutir e colaborar. **https://dadosabertos.social/**

SOBRE A OKBR

SITE INSTITUCIONAL assine as newsletters de sua preferência e siga a OKBR nas redes sociais. **https://www.ok.org.br/contato/**

ESCOLA DE DADOS torne-se um membro da escola de dados e tenha acesso exclusivo a materiais e descontos em cursos! Assine nossa newsletter e receba novidades em seu e-mail: **https://escoladedados.org/**


OPEN KNOWLEDGE
BRASIL


ESCOLA DE DADOS