Como os jornalistas podem usar dados para melhorar suas reportagens

Manual de Jornalismo de Dados

Editado por Jonathan Gray, Liliana Bounegru e Lucy Chambers

Introdução

- O que é o jornalismo de dados?
- Por que jornalistas devem usar dados?
- Por que o Jornalismo de Dados é importante?
- Alguns exemplos selecionados
- Jornalismo de dados em perspectiva
- O jornalismo guiado por dados numa perspectiva brasileira
- Existe jornalismo de dados e visualização no Brasil?

Na Redação

- O Jornalismo de dados da ABC (Australian Broadcasting Corporation)
- Jornalismo de Dados na BBC
- Como trabalha a equipe de aplicativos de notícias no Chicago Tribune
- Bastidores do Guardian Datablog
- Jornalismo de dados no Zeit Online
- Como contratar um hacker
- Aproveitando a expertise dos outros com Maratonas Hacker
- Seguindo o Dinheiro: Jornalismo de dados e Colaboração além das Fronteiras
- Nossas Histórias Vêm Como Código
- Kaas & Mulvad: Conteúdo pré-produzido para comunicação segmentada
- Modelos de Negócio para o Jornalismo de Dados

Estudos de Caso

- Basômetro: Passando o poder da narrativa para o usuário
- InfoAmazônia: o diálogo entre jornalismo e dados geográficos
- The Opportunity Gap: projeto de oportunidades em escolas
- Uma investigação de nove meses dos Fundos Estruturais Europeus
- A crise da Zona do Euro

- Cobrindo o gasto público com OpenSpending.org
- Eleições parlamentares finlandesas e financiamento de campanha
- Hack Eleitoral em tempo real (Hacks/Hackers Buenos Aires)
- Dados no Noticiário: WikiLeaks
- Hackatona Mapa76
- A cobertura dos protestos violentos no Reino Unido pelo The Guardian
- Boletins escolares de Illinois (EUA)
- Faturas de hospitais
- Care Home Crisis: A crise da empresas de saúde em domicílio
- O telefone conta tudo
- Quais modelos se saem pior na inspeção veicular britânica?
- Subsídios de ônibus na Argentina
- Jornalistas de dados cidadãos
- O Grande Quadro com o Resultado das Eleições
- Apurando o preço da água via crowdsourcing

Coletando dados

- Guia rápido para o trabalho de campo
- Seu Direito aos Dados
- Lei de Acesso à Informação no Brasil: Um longo caminho a percorrer
- Pedidos de informação funcionam. Vamos usá-los!
- Ultrapassando Obstáculos para obter Informação
- A Web como uma Fonte de dados
- O Crowdsourcing no Guardian Datablog
- Como o Datablog usou crowdsourcing para cobrir a compra de ingressos na Olimpíada
- Usando e compartilhando dados: a letra da lei, a letra miúda e a realidade

Entendendo os Dados

- Familiarizando-se com os dados em três passos
- Dicas para Trabalhar com Números
- Primeiros passos para trabalhar com dados
- O pão de 32 libras
- Comece com os dados e termine com uma reportagem
- Contando histórias com dados
- Jornalistas de dados comentam suas ferramentas preferidas
- Usando a visualização de dados para encontrar ideias

Comunicando os dados

- Apresentando os dados ao público
- Como construir um aplicativo jornalístico
- Aplicativos jornalísticos no ProPublica
- A visualização como carro-chefe do jornalismo de dados
- Usando visualização para contar histórias
- Gráficos diferentes contam histórias diferentes
- O faça-você-mesmo da visualização de dados: nossas ferramentas favoritas
- Como mostramos os dados no Verdens Gang
- Dados públicos viram sociais
- Engajando pessoas nos seus dados

O que é este livro (e o que ele não é)

A intenção deste livro é ser uma fonte útil para qualquer um que possa estar interessado em se tornar um jornalista de dados, ou em aventurar-se no jornalismo de dados.

Muitas pessoas contribuíram na sua composição, e, através do nosso esforço editorial, tentamos deixar essas diferentes vozes e visões brilharem. Nós esperamos que ele seja lido como uma conversa rica e informativa sobre o que é jornalismo de dados, por que ele é importante, e como fazê-lo.

Infelizmente, ler este livro não vai te dar um repertório completo de todo o conhecimento e habilidade necessários para se tornar um jornalista de dados. Para isso, seria necessária uma vasta biblioteca de informações composta por centenas de experts capazes de responder questões sobre centenas de tópicos. Felizmente, essa biblioteca existe: a internet. Ainda assim, nós esperamos que este livro possa te dar a noção de como começar e de onde procurar se você quiser ir além. Exemplos e tutoriais servem para serem ilustrativos e não exaustivos.

Nós nos consideramos muito sortudos por termos tido tanto tempo, energia, e paciência de todos os nossos voluntários, e fizemos o melhor para tentar usar isso com sabedoria. Esperamos que, além de ser uma fonte de referência útil, o livro sirva também para documentar a paixão e o entusiasmo, a visão e a energia de um movimento que está nascendo. O livro é uma tentativa entender o que acontece nos bastidores dessa cena de jornalismo de dados.

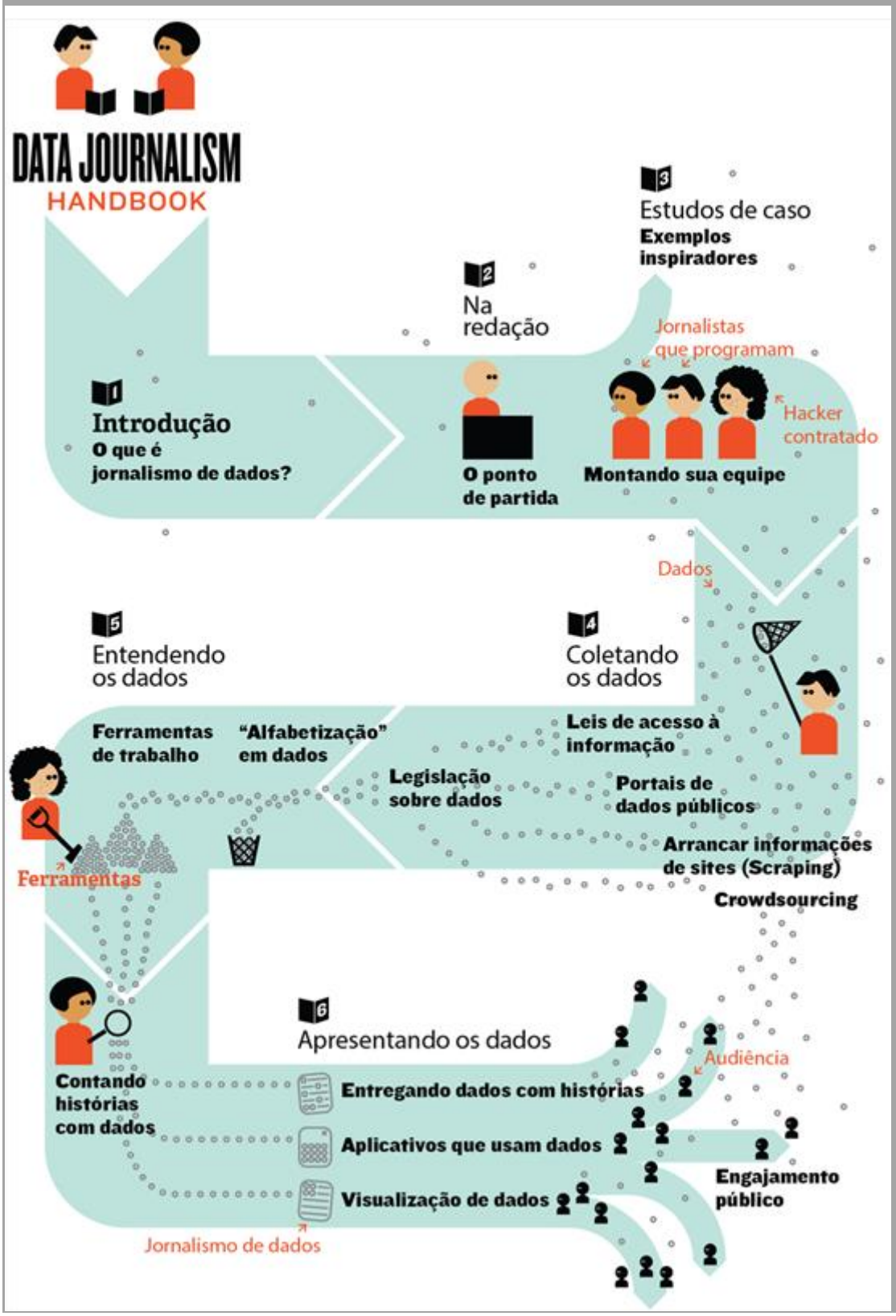
O Data Journalism Handbook é um trabalho em curso. Se você acha que há qualquer coisa que precisa ser corrigida ou está ausente, por favor nos avise para que ela seja incluída na próxima versão. Ele também está disponível de maneira gratuita em uma licença Creative Commons de Atribuição + Compartilhamento, e nós encorajamos fortemente a compartilhá-lo com qualquer um que possa estar interessado.

Liliana Bounegru (@bb_liliana)
Lucy Chambers (@lucyfedia)
Jonathan Gray (@jwyg)
Março de 2012

- See more at: http://datajournalismhandbook.org/pt/0_paginas_preliminares_3.html#sthash.CkLoMKly.dpuf

Visão Geral do Livro

A designer de infográficos Lulu Pinney criou este lindo pôster, que dá um panorama geral do conteúdo do Data Journalism Handbook.



Introdução

O que é o jornalismo de dados? Qual é o seu potencial? Quais são seus limites? De onde ele vem? Nesta seção iremos explicar o que é o jornalismo de dados e o que ele pode significar para as organizações jornalísticas. Paul Bradshaw (Birmingham City University) e Mirko Lorenz (Deutsche Welle) discorrem um pouco sobre o que há de diferente nesse tipo de reportagem. Jornalistas de dados de destaque nos contam por que o consideram importante e quais são seus exemplos favoritos. Finalmente, Liliana Bounegru (Centro Europeu de Jornalismo) coloca o jornalismo de dados em seu contexto histórico mais amplo.

O que há neste capítulo?

O que é o jornalismo de dados?

Eu poderia responder, simplesmente, que é um jornalismo feito com dados. Mas isso não ajuda muito.

Ambos, "dados" e "jornalismo", são termos problemáticos. Algumas pessoas pensam em "dados" como qualquer grupo de números, normalmente reunidos numa planilha. Há 20 anos, este era praticamente o único tipo de dado com o qual os jornalistas lidavam. Mas nós vivemos num mundo digital agora, um mundo em que quase tudo pode ser (e quase tudo é) descrito com números.

A sua carreira, 300 mil documentos confidenciais, todos dentro do seu círculo de amizades; tudo isso pode ser (e é) descrito com apenas dois números: zeros e uns. Fotos, vídeos e áudio são todos descritos com os mesmos dois números: zeros e uns. Assassinatos, doenças, votos, corrupção e mentiras: zeros e uns.

O que faz o jornalismo de dados diferente do restante do jornalismo? Talvez sejam as novas possibilidades que se abrem quando se combina o tradicional "faro jornalístico" e a habilidade de contar uma história envolvente com a escala e o alcance absolutos da informação digital agora disponível.

Estas possibilidades aperecem em qualquer estágio do processo, seja usando programas para automatizar o trabalho de combinar informação do governo local, polícia e outras fontes civis, como Adrian Holovaty fez no ChicagoCrime e depois no EveryBlock; seja usando um softtware para achar conexões entre centenas de milhares de documentos, como o The Telegraph fez com o MPs' expenses.

Imagem 1. Chamado para ajudar a investigar os gastos dos Membros do Parlamento (MPs) - (the Guardian)

Jornalismo de dados pode ajudar um jornalista a formular uma reportagem complexa através de infográficos envolventes. Por exemplo, as palestras espetaculares de Hans Rosling para visualizar a pobreza no mundo com o Gapminder atraíram milhões de visualizações em todo mundo. E o trabalho popular de David McCandless em destrinchar grandes números — como colocar gastos públicos dentro de contexto, ou a poluição gerada e evitada pelo vulcão islandês — mostra a importância de um design claro, como o doInformation is Beautiful.

Ou ainda o jornalismo de dados pode ajudar a explicar como uma reportagem se relaciona com um indivíduo, como a BBC e o Financial Times costumam fazem com seus orçamentos interativos (em que se pode descobrir como o orçamento público afeta especificamente você, em vez de saber como afeta uma "pessoa comum"). Ele pode também revelar o processo de construção das notícias, como o Guardian fez de maneira tão bem-sucedida compartilhando dados, contextos e questões com o Datablog.

Os dados podem ser a fonte do jornalismo de dados, ou podem ser as ferramentas com as quais uma notícia é contada — ou ambos. Como qualquer fonte, devem ser tratados com ceticismo; e como qualquer ferramenta, temos de ser conscientes sobre como eles podem moldar e restringir as reportagens que nós criamos com eles.

— Paul Bradshaw, Birmingham City University

Por que jornalistas devem usar dados?

O jornalismo está sitiado. No passado, nós, como uma indústria, contávamos com o fato de sermos os únicos a operar a tecnologia para multiplicar e distribuir o que havia acontecido de um dia para o outro. A imprensa servia como um portão: se alguém quisesse impactar as pessoas de uma cidade ou região na manhã seguinte, deveria procurar os jornais. Isso acabou.

Hoje as notícias estão fluindo na medida em que acontecem, a partir de múltiplas fontes, testemunhas oculares, blogs, e o que aconteceu é filtrado por uma vasta rede de conexões sociais, sendo classificado, comentado e, muito frequentemente, ignorado.

Esta é a razão pela qual o jornalismo de dados é tão importante. Juntar informações, filtrar e visualizar o que está acontecendo além do que os olhos podem ver tem um valor crescente. O suco de laranja que você bebe de manhã, o café que você prepara: na economia global de hoje existem conexões invisíveis entre estes produtos, as pessoas e você. A linguagem desta rede são os dados: pequenos pontos de informação que muitas vezes não são relevantes em uma primeira instância, mas que são extraordinariamente importantes quando vistos do ângulo certo.

Agora mesmo, alguns jornalistas pioneiros já demonstram como os dados podem ser usados para criar uma percepção mais profunda sobre o que está acontecendo ao nosso redor e como isto pode nos afetar.

A análise dos dados pode revelar "o formato de uma história" (Sarah Cohen), ou nos fornecer uma "nova câmera" (David McCandless). Usando os dados, o principal foco do trabalho de jornalistas deixa de ser a corrida pelo furo e passa a ser dizer o que um certo fato pode realmente significar. O leque de temas é abrangente: a próxima crise financeira em formação, a economia por trás dos produtos que usamos, o uso indevido de recursos ou os tropeços políticos. Tudo isso pode ser apresentado em uma visualização de dados convincente que deixe pouco espaço para discussão.

Exatamente por isso jornalistas deveriam ver nos dados uma oportunidade. Eles podem, por exemplo, revelar como alguma ameaça abstrata, como o desemprego, afeta as pessoas com base em sua idade, sexo ou educação. Usar

dados transforma algo abstrato em algo que todos podem entender e se relacionar.

Eles podem criar calculadoras personalizadas para ajudar as pessoas a tomarem decisões, seja comprar um carro, uma casa, decidir um rumo educacional ou profissional ou ainda verificar os custos de se manter sem dívidas.

Eles podem analisar a dinâmica de uma situação complexa, como protestos ou debates políticos, mostrar falácias e ajudar todos a verem as possíveis soluções para problemas complexos.

Ter conhecimento sobre busca, limpeza e visualização de dados é transformador também para o exercício da reportagem. Jornalistas que dominam estas habilidades vão perceber que construir artigos a partir de fatos e ideias é um alívio. Menos adivinhação, menos busca por citações; em vez disso, um jornalista pode construir uma posição forte apoiada por dados, o que pode afetar consideravelmente o papel do jornalismo.

Além disso, ingressar no jornalismo de dados oferece perspectivas de futuro. Hoje, quando redações cortam suas equipes, a maioria dos jornalistas espera se transferir para um emprego em relações públicas ou assessoria de imprensa. Jornalistas de dados e cientistas de dados, contudo, já são um grupo procurado de funcionários, não só nos meios de comunicação. As empresas e instituições ao redor do mundo estão buscando "intérpretes" e profissionais que saibam entrar fundo nos dados e transformá-los em algo tangível.

Há uma promessa de futuro nos dados e isso é o que o excita as redações, fazendo-as procurar por um novo tipo de repórter. Para freelancers, a proficiência com dados fornece um caminho para novas ofertas e remuneração estável também. Veja deste modo: em vez de contratar jornalistas para preencher rapidamente as páginas e os sites com conteúdo de baixo valor, a utilização dos dados poderia criar demanda para pacotes interativos, nos quais passar uma semana resolvendo uma questão é a única maneira de fazê-los. Esta é uma mudança bem-vinda em muitas partes da mídia.

Há uma barreira impedindo os jornalistas de usarem este potencial: treinamento para aprender como trabalhar com dados passo-a-passo, da primeira questão até um furo obtido pelo trabalho com os dados.

Trabalhar com dados é como pisar em um vasto e desconhecido território. À primeira vista, os dados brutos são intrigantes aos olhos e à mente. Esses dados

> são complicados. É bastante difícil moldá-los corretamente para a visualização. Isto requer jornalistas experientes, que têm energia para olhar aqueles dados brutos, por vezes confusos, por vezes chatos, e enxergar as histórias escondidas lá dentro.
>
> *— Mirko Lorenz, Deutsche Welle*

A Pesquisa

O Centro Europeu de Jornalismo realizou uma pesquisa para saber mais sobre as necessidades de formação dos jornalistas. Descobrimos que há uma grande vontade de sair da zona de conforto do jornalismo tradicional e investir tempo em dominar novas habilidades. Os resultados da pesquisa nos mostraram que os jornalistas veem a oportunidade, mas precisam de um pouco de apoio para acabar com os problemas iniciais que os impedem de trabalhar com dados. Existe uma confiança de que se o jornalismo de dados for adotado mais universalmente, os fluxos de trabalho, ferramentas e os resultados vão melhorar muito rapidamente. Pioneiros como The Guardian, The New York Times, Texas Tribune, e Die Zeit continuam a elevar o nível com suas histórias baseadas em dados.

Será que o jornalismo de dados permanecerá restrito a um pequeno grupo de pioneiros, ou será que cada organização de notícias em breve vai ter sua própria equipe dedicada ao jornalismo de dados? Esperamos que este manual ajude mais jornalistas e redações a tirar proveito deste campo emergente.

Imagem 2. Pesquisa do Centro Europeu de Jornalismo sobre necessidades de treinamento.

Por que o Jornalismo de Dados é importante?

Perguntamos a alguns dos principais profissionais da área por que eles acham que o o jornalismo de dados é um avanço importante. Aqui está o que disseram.

Filtrando o Fluxo de Dados

Quando a informação era escassa, a maior parte de nossos esforços estavam voltados à caçar e reunir dados. Agora que a informação é abundante, processá-la tornou-se mais importante. O processamento acontece em dois níveis: 1) análise para entender e estruturar um fluxo infinito de dados e 2) apresentação para fazer com que os dados mais importantes e relevantes cheguem ao consumidor. Como acontece na ciência, o jornalismo de dados revela seus métodos e apresenta seus resultados de uma forma que possam ser replicados.

— Philip Meyer, Professor Emérito da Universidade da Carolina do Norte, em Chapel Hill

Novas abordagens para a narrativa

O jornalismo de dados é um termo que, ao meu ver, engloba um conjunto cada vez maior de ferramentas, técnicas e abordagens para contar histórias. Pode

incluir desde a Reportagem com o Auxílio do Computador (RAC, que usa dados como uma "fonte") até as mais avançadas visualizações de dados e aplicativos de notícias. O objetivo em comum é jornalístico: proporcionar informação e análise para ajudar a nos informar melhor sobre as questões importantes do dia.

— Aron Pilhofer, New York Times

Como o fotojornalismo, só que com laptop

O jornalismo de dados só se diferencia do "jornalismo de palavras" porque usamos ferramentas distintas. Ambos trabalham buscando a notícia, fazendo reportagem e contando histórias. É como o fotojornalismo; só que substitui a câmera pelo laptop.

— Brian Boyer, Chicago Tribune

O Jornalismo de Dados é o Futuro

O jornalismo movido por dados é o futuro. Os jornalistas precisam ser conhecedores dos dados. Costumava-se conseguir novas reportagens conversando com pessoas em bares; e pode ser que, às vezes, você continue fazendo isso. Mas agora isso também é possível se debruçando sobre dados e se equipando com as ferramentas corretas para analisá-los e identificar o que há de interessante ali. Tendo isso em perspectiva, é possível ajudar as pessoas a descobrir como todas essas informações se encaixam e o que está acontecendo no país.

— Tim Berners-Lee, fundador da World Wide Web (WWW)

O processamento de dados encontra o a lapidação do texto

O jornalismo de dados está diminuindo a distância entre os técnicos estatísticos e os mestres da palavra. Faz isso ao localizar informações que fogem ao padrão e identificar tendências que não são apenas relevantes de um ponto de vista estatístico, mas também relevantes para decodificar a complexidade do mundo de hoje.

— David Anderton, jornalista freelancer

Atualizando o Seu Conjunto de Competências

O jornalismo de dados é um novo conjunto de competências para buscar, entender e visualizar fontes digitais em um momento em que os conhecimentos básicos do jornalismo tradicional já não são suficientes. Não se trata da substituição do jornalismo tradicional, mas de um acréscimo a ele.

Em um momento em que as fontes estão se tornando digitais, os jornalistas podem e devem estar perto dessas fontes. A internet abriu um mundo de possibilidades além da nossa compreensão atual. O jornalismo de dados é apenas o começo do processo de evolução de práticas antigas para se adaptar ao mundo online.

O jornalismo de dados cumpre dois objetivos importantes para as organizações de mídia: encontrar notícias únicas (que não sejam de agências), e executar a função fiscalização do poder. Especialmente em tempos de perigo financeiro, essas metas são bastante importantes para os jornais.

Do ponto de vista de um jornal local, o jornalismo de dados é crucial. Existe um ditado que diz que "uma telha solta na frente da sua porta é mais importante que uma revolta em um país distante". O fato que se coloca diante de você e provoca impacto direto na sua vida. Ao mesmo tempo, a digitalização está em todos os lugares. Porque jornais locais têm esse impacto direto na região em que são distribuídos e as fontes tornam-se cada vez mais digitais, um jornalista precisa saber como encontrar, analisar e visualizar histórias usando dados como matéria-prima.

— Jerry Vermanen, NU.nl

Um remédio para a assimetria da informação

A assimetria da informação — não a falta de informação, mas a incapacidade de absorvê-la e processá-la na velocidade e no volume com que chega até nós --, é um dos problemas mais significativos enfrentados pelos cidadãos ao fazer escolhas sobre como viver suas vidas. Informações obtidas pela imprensa e a mídia influenciam escolhas e ações dos cidadãos. O bom jornalismo de dados ajuda a combater a assimetria da informação.

— Tom Fries, Fundação Bertelsmann

Uma resposta para o uso de dados por assessorias de imprensa

A disponibilidade de ferramentas de medição e a diminuição de seus preços — em uma combinação autossustentável com foco na performance e na eficiência em todos os aspectos da sociedade — levaram tomadores de decisão a quantificar os progressos de suas políticas, monitorar tendências e identificar oportunidades.

As empresas continuam adotando novas métricas mostrando quão boa são as suas performances no mercado. Os políticos adoram se gabar sobre reduções

dos níveis de desemprego e aumentos do PIB. A falta de visão jornalística em temas como os escândalos da Enron, Worldcom, Madoff ou Solyndra é a prova da falta de habilidade dos jornalistas para ver através e além dos números. É mais fácil aceitar o valor de face dos números do que o de outros fatos, já que carregam uma aura de seriedade mesmo quando são complemente fabricados.

A fluência no uso de dados ajuda os jornalistas a analisar os números com senso crítico, e certamente os ajudará a ganhar terreno em seus contatos com assessorias de imprensa.

— Nicolas Kayser-Bril, Journalism++

Oferecendo interpretações independentes de informações oficiais

Após o terremoto devastador e o consequente desastre na usina nuclear de Fukushima, em 2011, o jornalismo de dados foi ganhando corpo e importância entre membros da mídia no Japão, país geralmente atrasado com relação ao jornalismo digital.

Estávamos perdidos quando o governo e especialistas não tinham dados confiáveis sobre os danos provocados. Quando os oficiais esconderam do público informações do sistema SPEEDI (rede de sensores japoneses que deve prever a propagação de radiação entre outras coisas), não estávamos preparados para decodificar os dados, mesmo que tivessem vazado. Voluntários começaram a coletar dados sobre radiação usando seus próprios dispositivos, mas nós não estávamos armados com o conhecimento de estatística, interpolação e visualização desses dados, entre outras coisas. Jornalistas precisam ter acesso aos dados brutos, e aprender a não confiar apenas nas interpretações oficiais deles.

— Isao Matsunami, Tokyo Shimbun

Lidar com o dilúvio informacional

Os desafios e oportunidades trazidos pela revolução digital continuam disruptivos para o jornalismo. Numa era de abundância de informação, jornalistas e cidadãos precisam de ferramentas melhores, seja quando estivermos fazendo a curadoria de material proibido por governos do Oriente Médio, processando dados surgidos de última hora, ou buscando a melhor maneira de visualizar a qualidade da água para uma nação de consumidores. À medida que lutamos contra os desafios do consumo apresentados por esse dilúvio de informações, novas plataformas de publicação também permitem a

qualquer pessoa ter o poder de reunir e compartilhar dados digitalmente, transformando-os em informação. Embora repórteres e editores têm sido os tradicionais vetores para coletar e disseminar informação, no ambiente informacional de hoje as notícias mais quentes aparecem antes na internet, e não nas editorias de jornais.

Ao redor do mundo o vínculo entre os dados e o jornalismo está em forte ascensão. Na era do big data, a crescente importância do jornalismo de dados reside na capacidade de seus praticantes de fornecer contexto, clareza e, talvez o mais importante, encontrar a verdade em meio à expansão de conteúdo digital no mundo. Isso não significa que as organizações de mídia de hoje não tenham um papel crucial. Longe disso. Na era da informação, jornalistas são mais necessários que nunca para fazer a curadoria, verificar, analisar e sintetizar a imensidão de dados. Neste contexto, o jornalismo de dados tem uma importância profunda para a sociedade.

Hoje, entender um grande volume de dados ("big data"), particularmente dados não estruturados, é um objetivo central para cientistas de dados ao redor do mundo, estejam eles em redações, em Wall Street ou no Vale do Silício. Um conjunto crescente de ferramentas comuns, quer empregadas por técnicos governamentais de Chicago, técnicos de saúde ou desenvolvedores de redações, fornece ajuda substancial para atingir esse objetivo.

— Alex Howard, O'Reilly Media

Nossas vidas são dados

Fazer bom jornalismo de dados é difícil porque o bom jornalismo é difícil. Significa descobrir como obter os dados, entendê-los e encontrar a história. Às vezes há becos sem saída e não há uma grande reportagem. Afinal, se fosse apenas uma questão de pressionar um botão certo, não seria jornalismo. Mas é isso o que faz o jornalismo de dados valer à pena e, em um mundo onde nossas vidas estão cada vez mais compostas por dados, a área torna-se essencial para uma sociedade justa e livre.

— Chris Taggart, OpenCorporates

Uma forma de economizar tempo

Jornalistas não têm tempo para gastar na transcrição de documentos ou tentando obter dados de PDFs, de modo que aprender um pouco de

programação (ou saber onde buscar pessoas que podem ajudar) é incrivelmente valioso.

Um repórter da Folha de S.Paulo estava trabalhando com um orçamento local e me chamou para agradecer o fato de termos colocado online as contas da cidade de São Paulo (dois dias de trabalho para um único hacker!). Ele disse que vinha transcrevendo essas informações manualmente ao longo de três meses, tentando construir uma reportagem. Eu também lembro de ter solucionado uma questão ligada a um PDF para o *Contas Abertas*, uma organização de notícias de monitoramento parlamentar: 15 minutos e 15 linhas de código conseguiram o mesmo resultado que um mês de trabalho.

— Pedro Markun, Transparência Hacker

Uma parte essencial do pacote de ferramentas dos jornalistas

É importante ressaltar a parte jornalística ou o lado da reportagem do jornalismo de dados. O exercício não deve ser o de analisar e visualizar por si só, mas também de usar os dados como uma ferramenta para se aproximar da verdade e do que está acontecendo no mundo. Vejo a capacidade de analisar e interpretá-los como parte essencial do kit atual de ferramentas jornalísticas, mais do que uma disciplina à parte. Por fim, trata-se de fazer boas reportagens e contar histórias da forma mais apropriada.

Esse novo jornalismo é outro meio de analisar o mundo e fazer com que os governantes prestem contas. Com uma quantidade cada vez maior de dados, é mais importante que nunca que os jornalistas estejam conscientes dessas técnicas. Isso deveria estar no arsenal de técnicas de reportagem de qualquer jornalista, seja aprender diretamente a trabalhar com os dados ou colaborar com alguém que cumpra esse papel.

O real poder do jornalismo de dados é ajudar a obter e provar informações quando, por outros meios, seria muito difícil. Um bom exemplo disso é uma reportagem de Steve Doig que analisava os danos provocados pelo furacão Andrew. Ele juntou dois conjuntos diferentes de dados: um mapeava o nível de destruição causado pelo furação, e o outro mostrava a velocidade dos ventos. Isso permitiu identificar áreas onde construções enfraquecidas e práticas de construção não confiáveis contribuíram para aumentar o impacto do desastre. O trabalho ganhou um Prêmio Pulitzer em 1993 e continua sendo um grande exemplo do potencial do jornalismo de dados.

Idealmente, usa-se dados para identificar fatos que fogem ao padrão, áreas de interesse ou coisas que são surpreendentes. Neste sentido, eles podem agir como um norte ou como pistas. Os números podem ser interessantes, mas apenas escrever sobre eles não é suficiente. Você ainda vai precisar fazer reportagem para explicar o que eles significam.

— Cynthia O'Murchu, Financial Times

Adaptação a Mudanças no nosso ambiente informacional

Novas tecnologias digitais trazem novas formas de produzir e disseminar conhecimento na sociedade. O jornalismo de dados pode ser entendido como uma tentativa da mídia de se adaptar às mudanças e responder a elas em um ambiente repleto de informação, incluindo o relato de histórias mais interativas e multidimensionais que permitem aos leitores explorar as fontes subjacentes às notícias e incentivá-los a participar da criação e avaliação de reportagens.

— César Viana, Universidade de Goiás

Um jeito de ver coisas que você não enxergaria de outra forma

Algumas histórias podem apenas ser entendidas e explicadas por meio da análise — e às vezes da visualização — de dados. Conexões entre pessoas ou entidades poderosas continuariam ocultas, mortes causadas por políticas contra drogas seguiriam escondidas, políticas ambientais que destroem a natureza seguiriam inabaláveis. Mas cada ponto acima não permaneceu nessa situação devido a dados que os jornalistas obtiveram, analisaram e ofereceram aos leitores. Os dados podem ser tão simples como uma planilha básica ou um registro de chamadas de celular, ou tão complexos como notas de avaliações de escolas ou informações sobre infecção hospitalar. No fundo, porém, todas essas histórias são temas que merecem ser contados.

— Cheryl Phillips, The Seattle Times

Uma forma de contar histórias mais ricas

Podemos pintar histórias de toda a nossa vida por meio de nossos rastros digitais. Do que consumimos e pesquisamos a onde e quando viajamos, nossas preferências musicais, nossos primeiros amores, as realizações de nossos filhos, e até os nossos últimos desejos, tudo isso pode ser monitorado, digitalizado, armazenado na nuvem e disseminado. Esse universo de informações pode vir à tona para contar histórias, responder a questões e oferecer uma compreensão da

vida de uma maneira que atualmente supera até mesmo a reconstrução mais rigorosa e cuidadosa de anedotas.

— Sarah Slobin, Wall Street Journal

Você não precisa de dados novos para dar um furo

Às vezes, os dados já são públicos e estão disponíveis, mas ninguém olhou para eles com cuidado. No caso do relatório da Associated Press sobre 4.500 páginas de documentos revelados que descrevem ações de empresas de segurança privada contratadas durante a guerra do Iraque, o material foi obtido por um jornalista independente ao longo de vários anos. Ele fez diversos pedidos, por meio da lei de acesso à informação dos EUA (Freedom of Information Act) ao Departamento de Estado dos Estados Unidos. Eles escanearam os documentos em papel e os subiram no site DocumentCloud, o que tornou possível fazer uma análise abrangente da situação.

— Jonathan Stray, The Overview Project

Alguns exemplos selecionados

Nós pedimos a alguns de nossos voluntários que dessem seus exemplos favoritos de jornalismo de dados e dissessem o que gostavam neles. Aqui estão:

"Do no Harm", do Las Vegas Sun

Meu exemplo favorito é o a série Do No Harm de 2010 do Las Vegas Sun sobre serviço hospitalar. O The Sun analisou mais de 2,9 milhões de registros financeiros de hospitais, que revelaram mais de 3.600 lesões, infecções e erros médicos que poderiam ter sido prevenidos. Eles obtiveram as informações por meio de uma requisição de dados públicos e identificaram mais de 300 casos nos quais pacientes morreram por conta de erros que poderiam ter sido evitados. A reportagem possui diferentes elementos, que incluem: um gráfico interativo que permite ao leitor ver, por hospital, onde lesões decorrentes de cirurgia aconteceram mais que o esperado; um mapa e uma linha do tempo que mostra infecções se alastrando hospital por hospital e um gráfico interativo que permite aos usuários ordenar os dados por lesões evitáveis ou por hospital para ver onde as pessoas estão se machucando. Gosto deste trabalho porque é muito fácil de entender e navegar. Os usuários podem explorar os dados de uma maneira muito intuitiva.

Além disso, a iniciativa causou um impacto real: o legislativo de Nevada reagiu com seis projetos de lei. Os jornalistas envolvidos trabalharam arduamente para obter e limpar os dados. Um dos jornalistas, Alex Richards, mandou as informações de volta aos hospitais e para o Estado no mínimo uma dúzia de vezes para que as falhas fossem corrigidas.

— Angélica Peralta Ramos, La Nación (Argentina)

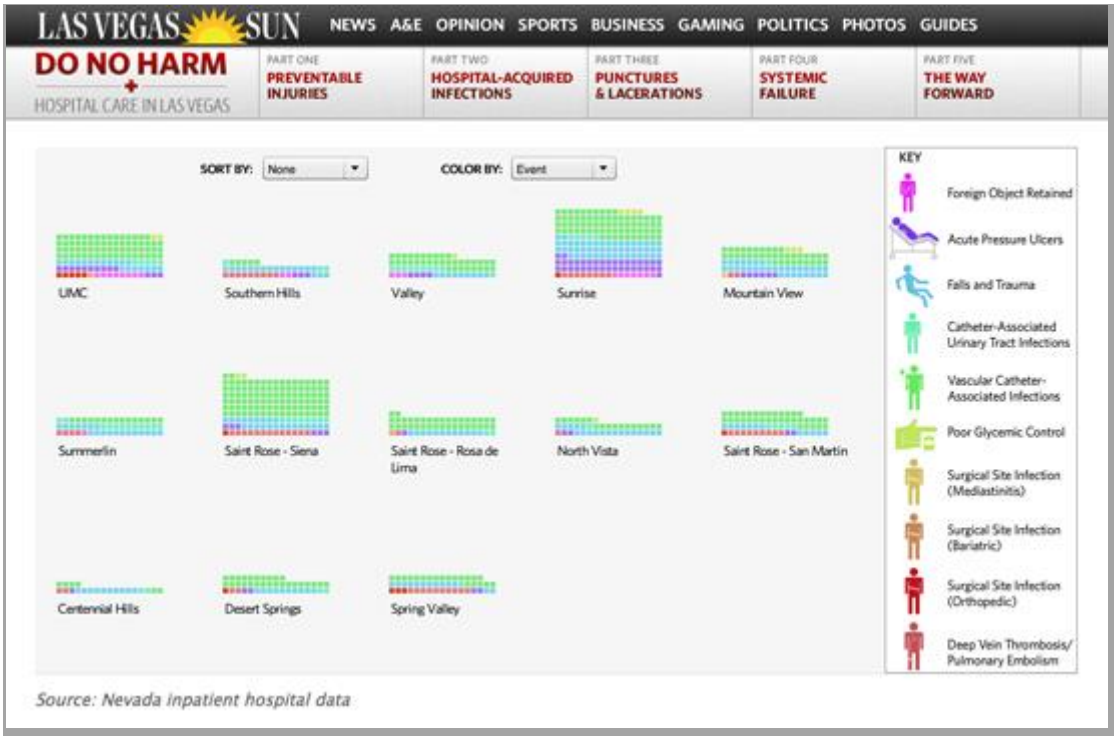


Imagem 3. Do No Harm (The Las Vegas Sun)

Banco de dados da Folha de Pagamento do Governo

Eu adoro o trabalho que organizações pequenas e independentes estão desempenhando todo dia, tais como a ProPublica ou o Texas Tribune que têm em Ryan Murphy um grande repórter de dados. Se eu tivesse que escolher, elegeria o projeto de Banco de Dados dos salários de empregados do governo do Texas Tribune. Este projeto coleta 660 mil salários de empregados públicos em um banco de dados para usuários procurarem e ajudarem a gerar matérias a partir dele. Você pode procurar por agência, nome ou salário. É simples, informativo e está tornando pública uma informação antes inacessível. É fácil de usar e automaticamente gera matérias. É um grande exemplo de por que o Texas Tribune consegue a maioria de seu tráfego das páginas de dados.

— Simon Rogers, the Guardian

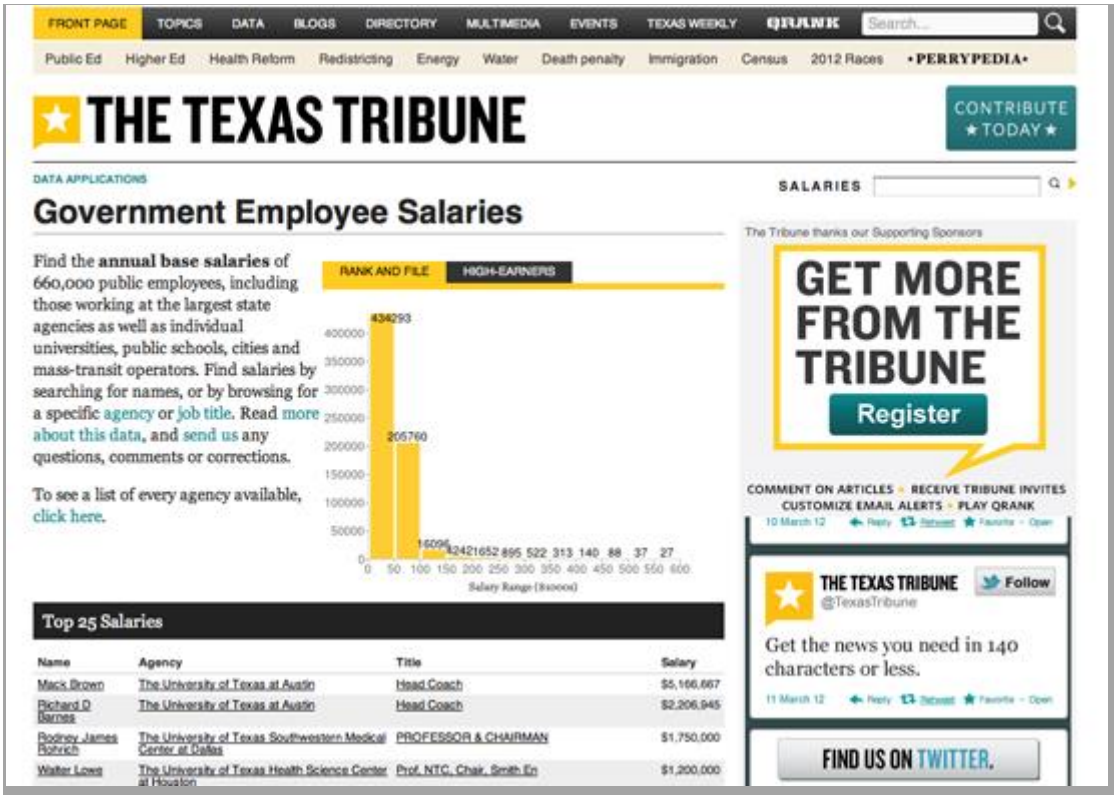


Imagem 4. Salários dos empregados do Governo (The Texas Tribune)

Visualização integral dos Registros da Guerra do Iraque, Associated Press

O trabalho de Jonathan Stray e Julian Burgess em cima dos Registros de Guerra do Iraqueé uma iniciativa inspiradora na análise e visualização de textos utilizando técnicas experimentais para ganhar profundidade em temas que valem a pena serem explorados dentro de um grande conjunto de dados textuais.

Por meio de técnicas de análise de texto e algoritmos, Jonathan e Julian criaram um método que mostrava blocos de palavras-chave contidas nos milhares de relatórios do governo americano sobre Guerra do Iraque vazados pelo Wikileaks, tudo de uma forma visual.

Embora haja restrições aos métodos apresentados e a abordagem seja experimental, o trabalho mostra um enfoque inovador. Em vez de tentar ler todos os arquivos e revirar os registros de guerra com uma noção preconcebida do que poderia ser achado com determinadas palavras-chaves, esta técnica calcula e visualiza tópicos e termos-chave de particular relevância.

Com a crescente quantidade de informação textual (emails, relatórios, etc) e numérica vindo ao domínio público, achar maneiras de identificar áreas vitais

de interesse será mais e mais importante – é um subcampo excitante do jornalismo de dados.

— Cynthia O'Murchu, Financial Times

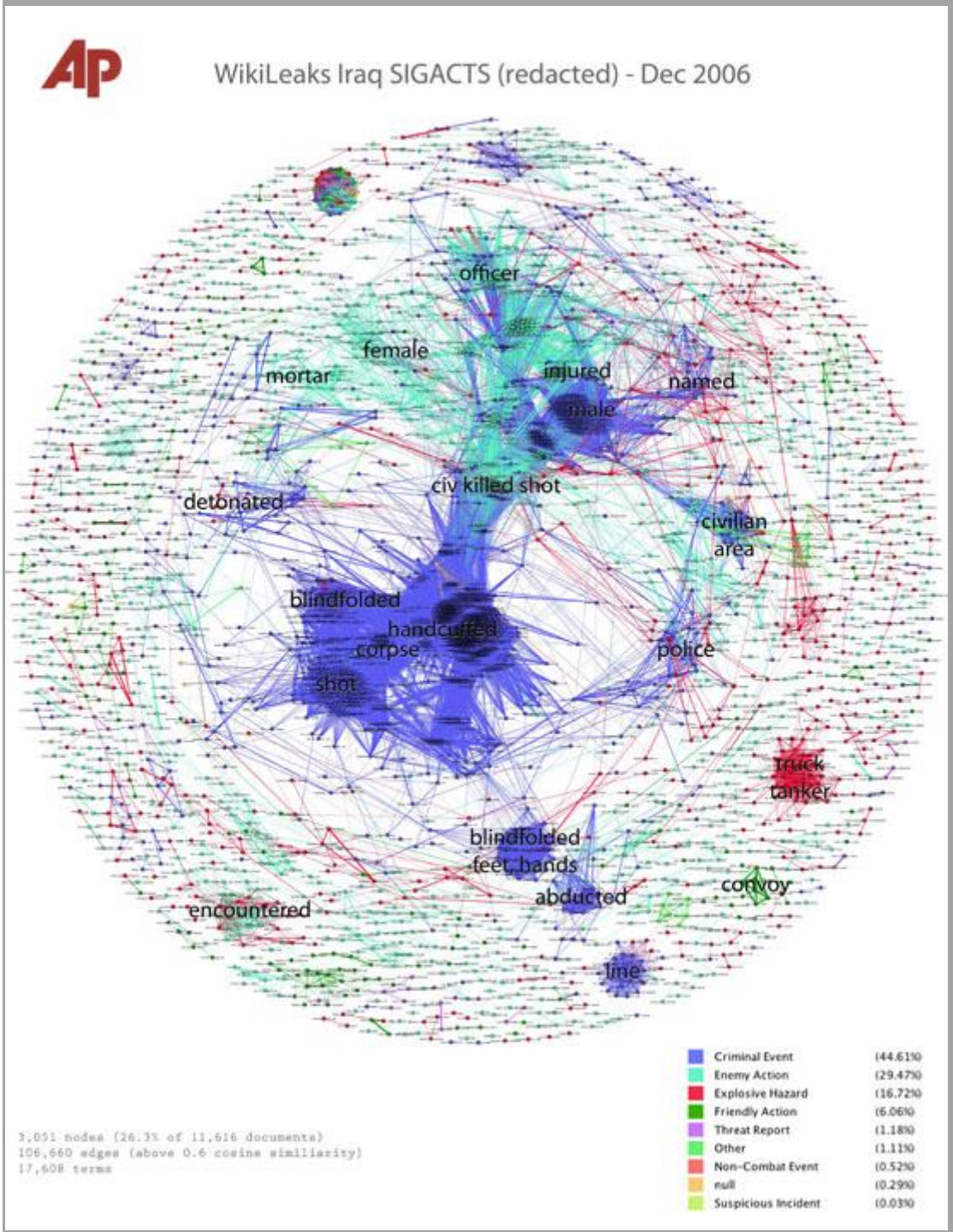


Imagem 5. Analizando os Registros de Combate (Associated Press)

Murder Mysteries

Uma das minhas obras favoritas de jornalismo de dados é o projeto *Murder Mysteries* de Tom Hardgrove do Scripps Howard News Service. Ele construiu um banco de dados detalhado de mais de 185 mil assassinatos não resolvidos a partir de dados governamentais e da requisição de registros públicos. A partir disso, ele desenvolveu um algoritmo que procura por padrões sugerindo a possível presença de serial killers. Este projeto é completo: trabalho árduo montando uma base de dados melhor que a do próprio governo, análise inteligente usando técnicas de ciências sociais e apresentação interativa dos dados online de modo que os leitores possam eles mesmos explorarem.

— Steve Doig, Walter Cronkite School of Journalism, Arizona State University

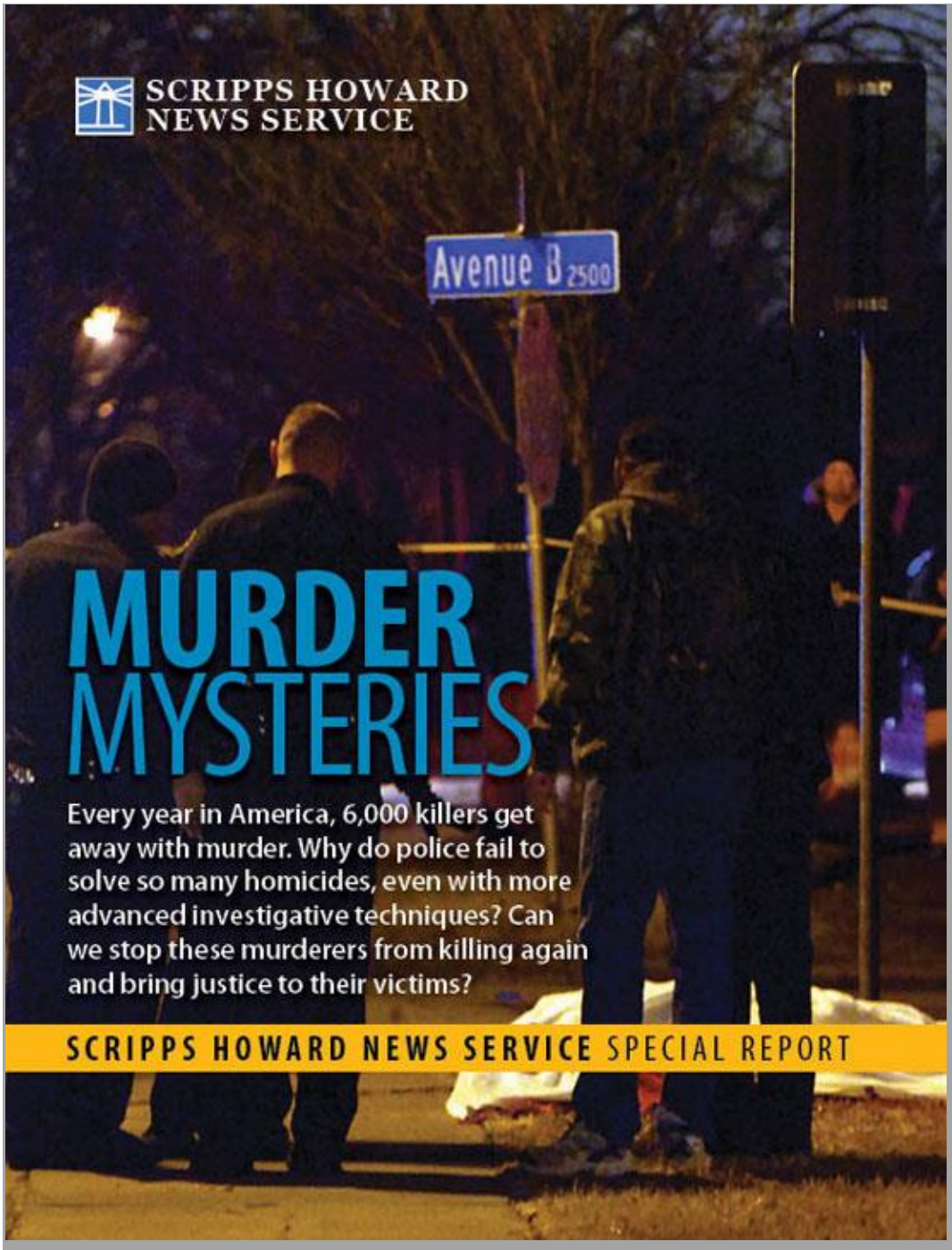


Imagem 6. Murder Mysteries (Scripps Howard News Service)

Message Machine

Eu adoro a reportagem e a postagem nerd do blog Message Machine da ProPublica. Tudo começou quando alguns tuiteiros mostraram curiosidade sobre terem recebido diferentes emails da campanha presidencial de Obama. Os colegas da ProPublica notaram e pediram para seu público encaminhar qualquer email que tivesse recebido da campanha. A forma como mostram os dados é elegante, uma apresentação visual da diferença entre muitos emails

distintos que foram enviados naquela noite. É extraordinário porque eles coletaram os próprios dados (reconhecidamente uma pequena amostra, mas grande o suficiente para montar uma reportagem). Mas é ainda mais incrível porque eles estão contando a história de um fenômeno emergente: big data usado em campanhas políticas para disparar mensagens especificamente preparadas para cada pessoa. Isso é só um gostinho das coisas por vir.

— Brian Boyer, Chicago Tribune

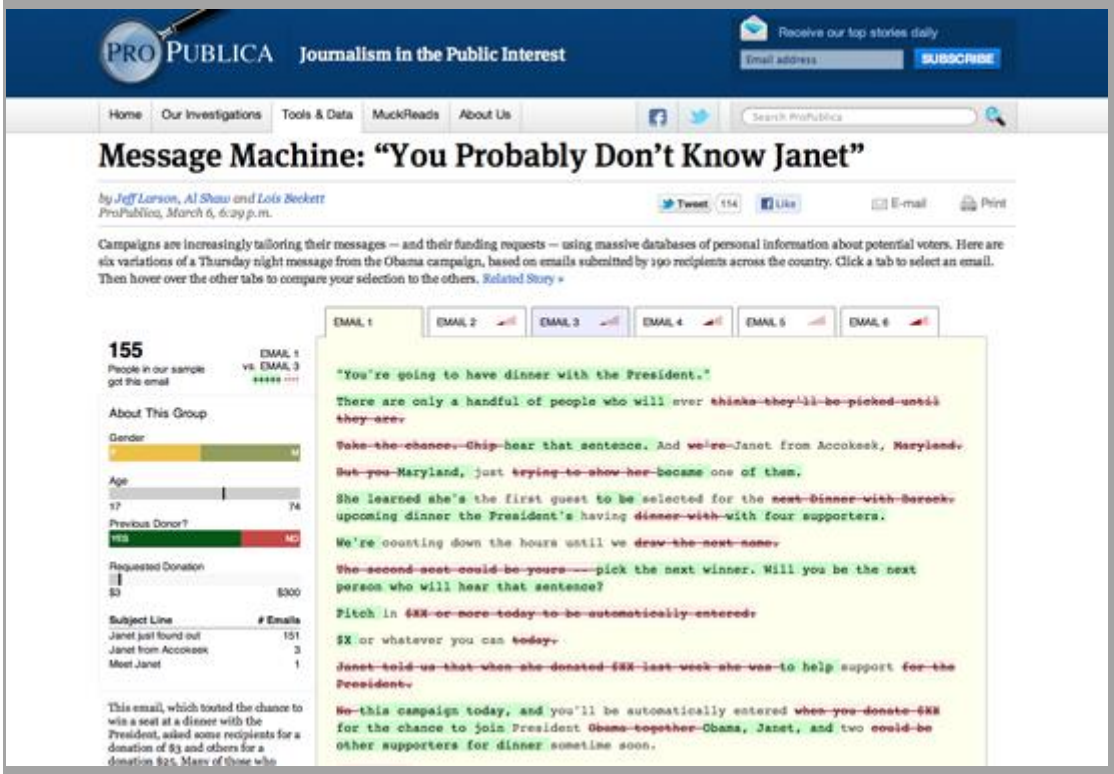


Imagem 7. Message Machine (ProPublica)

Chartball

Um dos meus projetos de jornalismo de dados favoritos é o trabalho de Andrew Garcia Phillips no Chartball. Andrew é um grande fã de esportes com um apetite voraz por dados, um olho espetacular para design e capacidade de programar. Com o Chartball ele visualiza não apenas a história, mas detalha os sucessos e fracassos de cada um dos jogadores e dos times de beisebol. Ele coloca em contexto, cria um gráfico atraente e seu trabalho é profundo, divertido e interessante. E olha que eu nem me importo tanto com esportes.

— Sarah Slobin, Wall Street Journal

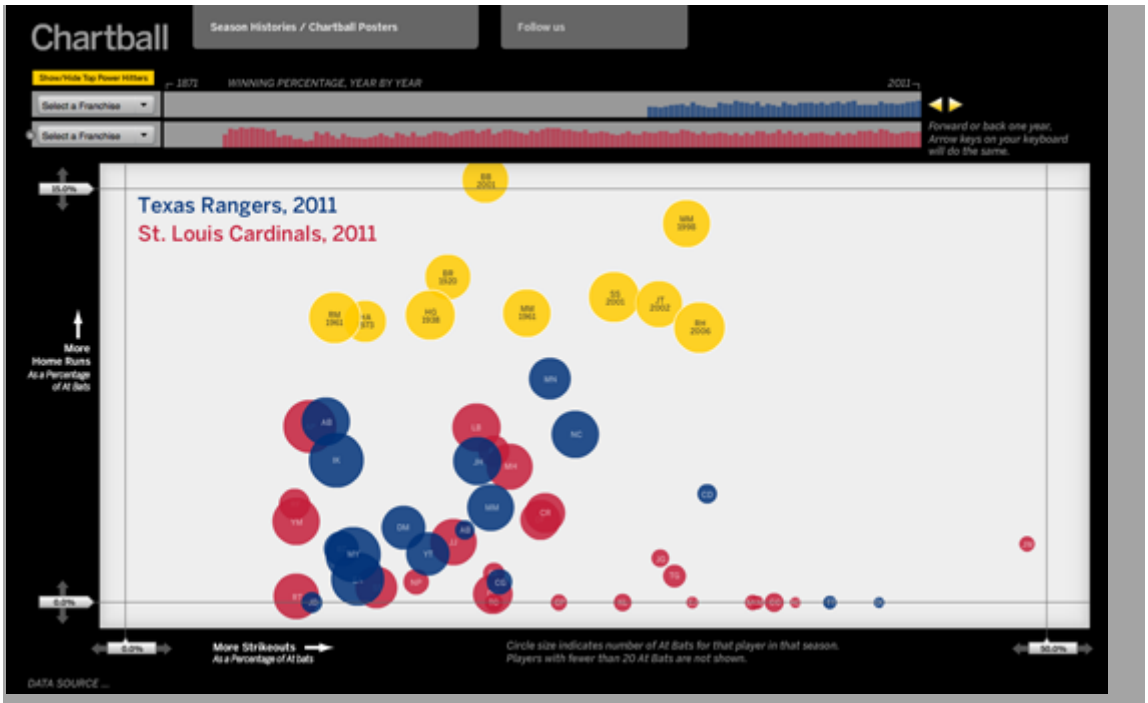


Imagem 8. Vitórias e derrotas em tabelas (Chartball)

Jornalismo de dados em perspectiva

Em agosto de 2010, eu e alguns colegas do Centro Europeu de Jornalismo organizamos o que acreditamos ser uma das primeiras conferências internacionais sobre jornalismo de dados, realizada em Amsterdã. Naquele momento, não havia muitas discussões sobre o tema e poucas organizações eram amplamente reconhecidas por trabalhar na área.

Um dos mais importantes passos para dar visibilidade ao termo foi a forma como grupos de mídia como The Guardian e The New York Times lidaram com a imensa quantidade de dados divulgados pelo WikiLeaks. Nesse período, o termo passou a ser usado de maneira mais ampla (ao lado de Reportagem com Auxílio do Computador, ou RAC) para descrever como jornalistas estavam usando dados para melhorar suas reportagens e para aprofundar investigações sobre um tema.

Ao conversar com jornalistas de dados experientes e teóricos do Jornalismo no Twitter, me parece que uma das primeiras definições do que hoje reconhecemos como jornalismo de dados foi feita em 2006, por Adrian Holovaty, fundador do EveryBlock, um serviço de informação que permite ao usuários descobrir o que está acontecendo na sua região, no seu quarteirão. No seu pequeno ensaio "Uma maneira fundamental na qual sites de jornais têm que mudar", ele defende que jornalistas devem publicar dados estruturados, compreensíveis por máquinas, ao lado do tradicional "grande borrão de texto":

Por exemplo, digamos que um jornal publicou uma notícia sobre um incêndio próximo. Ler essa história num celular é bacana e elegante. Viva a tecnologia! Mas o que realmente quero é ser capaz de explorar os dados brutos dessa história, um a um, com diferentes camadas. Ter a infraestrutura para comparar detalhes deste incêndio com os detalhes dos anteriores: data, horário, local, vítimas, distância para o quartel do Corpo de Bombeiros, nomes e anos de experiência dos bombeiros que foram ao local, tempo que levaram para chegar, e incêndios subsequentes, quando vierem a ocorrer.

Mas o que torna essa forma peculiar diferente de outros modelos de jornalismo que usam banco de dados ou computadores? Como e em que extensão o jornalismo de dados é diferente das vertentes de jornalismo do passado?

Reportagem com Auxílio do Computador (RAC) e o Jornalismo de Precisão

Há uma longa história de uso de dados para aprofundamento da reportagem e distribuição de informação estruturada (mesmo que não legível por máquinas). Talvez o mais relevante para o que hoje chamamos de jornalismo de dados é a Reportagem com Auxílio do Computador (RAC) que foi a primeira tentativa organizada e sistemática de utilizar computadores para coletar e analisar dados para aprimorar a notícia.

A RAC foi usada pela primeira vez em 1952 pela rede de TV americana CBS, para prever o resultado da eleição presidencial daquele ano. Desde a década de 60, jornalistas (principalmente os investigativos, principalmente nos Estados Unidos) têm analisado bases de dados públicas com métodos científicos para fiscalizar o poder de forma independente. Também chamado de "jornalismo de interesse público", defensores dessa técnicas baseadas no auxílio do computador têm procurado revelar tendências, contrariar o senso comum e desnudar injustiças perpetradas por autoridades e corporações. Por exemplo, Philip Meyer tentou desmontar a percepção de que apenas os sulistas menos educados participaram do quebra-quebra nas manifestações de 1967 em Detroit. As reportagens da série "A cor do dinheiro", publicadas nos anos 80 por Bill Dedman, revelaram preconceito racial sistemático nas políticas de empréstimo dos principais bancos. No seu artigo "O que deu errado", Steve Doig procurou analisar os padrões de destruição do Furacão Andrew no início dos anos 90, para entender as consequências das políticas e práticas falhas de desenvolvimento urbano. Reportagens movidas por dados prestaram valiosos serviços públicos e deram prêmios cobiçados aos autores.

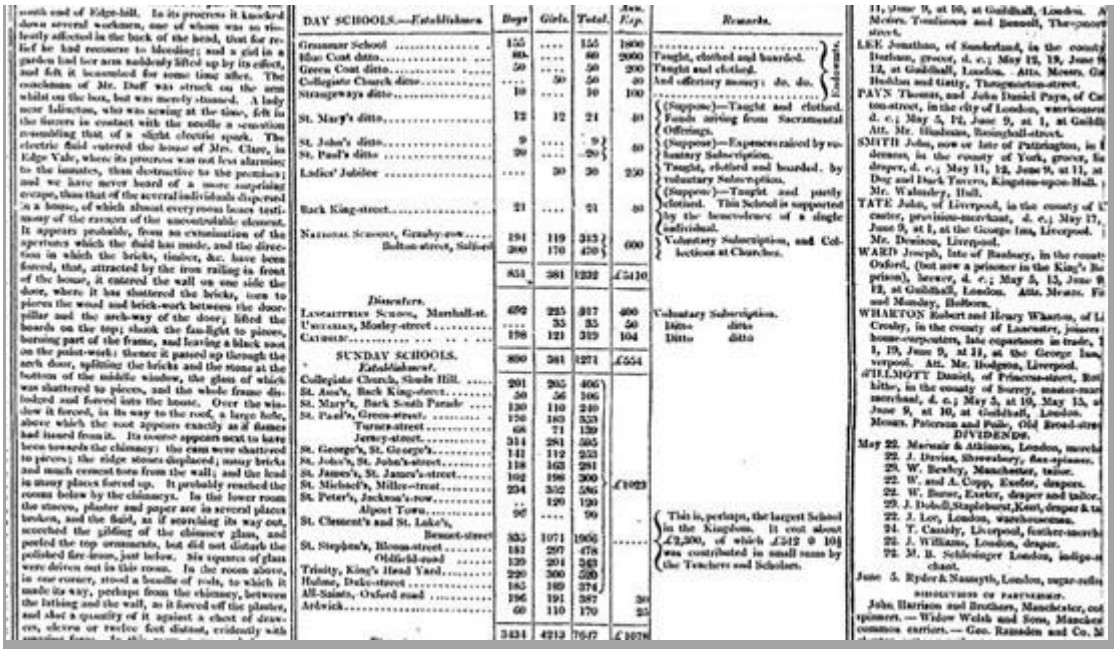
No início dos anos 70, o termo *jornalismo de precisão* foi cunhado para descrever esse tipo de apuração jornalística: "o emprego de métodos de pesquisa das ciências sociais e comportamentais na prática jornalística" (em *The New Precision Journalism* de Philip Meyer). O jornalismo de precisão foi proposto para ser praticado nas instituições jornalísticas convencionais por profissionais formados em jornalismo e em ciências sociais. Nasceu como resposta ao "New Journalism", que aplicava técnicas de ficção à reportagem. Meyer defendia que eram necessários métodos científicos para coleta e análise de dados, em vez de técnicas literárias, para permitir que o jornalismo alcançasse sua busca pela objetividade e verdade.

O jornalismo de precisão pode ser entendido como reação a algumas das inadequações e fraquezas do jornalismo normalmente citadas: dependência dos releases de assessorias (mais tarde descrito como "churnalism" ou "jornalismo de batedeira"), predisposição em acatar as versões oficiais, e por aí vai. Estas são decorrentes, na visão de Meyer, da não aplicação de técnicas e métodos científicos como pesquisas de opinião e consulta a registros públicos. Como feito nos anos 60, o jornalismo de precisão serviu para retratar grupos marginais e suas histórias. De acordo com Meyer:

O jornalismo de precisão foi uma forma de expandir o arsenal de ferramentas do repórter para tornar temas antes inacessíveis, ou parcialmente acessíveis, em objeto de exame minucioso. Foi especialmente eficiente para dar voz à minoria e grupos dissidentes que estavam lutando para se verem representados.

Um artigo influente publicado nos anos 80 sobre a relação entre o jornalismo e as ciências sociais ecoa o discurso atual em torno do jornalismo de dados. Os autores, dois professores de jornalismo americanos, sugerem que nas décadas de 70 e 80, a compreensão do público sobre o que é notícia se amplia de uma concepção mais direta de "fatos noticiosos" para "reportagens de comportamento" (ou reportagens sobre tendências sociais). Por exemplo, ao acessar os bancos de dados do Censo ou de outras pesquisas, os jornalistas conseguem "extrapolar o relato de eventos isolados e oferecer contexto que dá sentido ao fatos específicos".

Como podíamos esperar, a prática do uso de dados para incrementar a reportagem é tão antiga quanto a própria existência dos dados. Como Simon Rogers aponta, o primeiro exemplo de jornalismo de dados no The Guardian remonta a 1821. Foi uma lista, obtida de fonte não oficial, que relacionava as escolas da cidade de Manchester ao número de alunos e aos custos de cada uma. De acordo com Rogers, a lista ajudou a mostrar o verdadeiro número de alunos que recebiam educação gratuita, muito maior do que os números oficiais revelavam.



DAY SCHOOLS.—*Establishment.*

	Boys	Girls	Total	Ann. Exp.	Remarks
Grammar School	155	....	155	1800	Taught, clothed and boarded.
Blue Coat ditto	60	....	60	2000	Taught and clothed.
Green Coat ditto	50	....	50	200	And offertory money: do. do.
Collegiate Church ditto	....	50	50	20	Endowments.
Strangeways ditto	10	....	10	100	
St. Mary's ditto	12	12	24	40	(Suppose)—Taught and clothed. Funds arising from Sacramental Offerings.
St. John's ditto	9	....	9	60	(Suppose)—Expences raised by voluntary Subscription.
St. Paul's ditto	20	....	20		Taught, clothed and boarded. by voluntary Subscription.
Ladies' Jubilee	....	30	30	250	(Suppose)—Taught and partly clothed. This School is supported by the benevolence of a single individual.
Back King-street	21	....	21	80	Voluntary Subscription, and Collection at Churches.
National School, Granby-row	194	119	313	600	
Bolton-street, Salford	200	170	470		
	851	381	1232	£5410	
Dissenters.					
Lancasterian School, Marshall-st.	492	225	917	400	Voluntary Subscription.
Unitarian, Mosley-street	....	35	35	50	Ditto ditto
Catholic	198	121	319	104	Ditto ditto
	690	381	1271	£554	
SUNDAY SCHOOLS. *Establishment.*					
Collegiate Church, Shude Hill	201	205	406		
St. Ann's, Back King-street	50	56	106		
St. Mary's, Back South Parade	130	110	240		
St. Paul's, Green-street	176	163	353		
Turner-street	68	71	139		
Jersey-street	314	281	595		
St. George's, St. George's	141	112	253		
St. John's, St. John's-street	118	163	281		
St. James's, St. James's-street	102	198	300	£1023	
St. Michael's, Miller-street	234	352	586		
St. Peter's, Jackson's-row	..	120	120		
Algiers Town	96	....	96		
St. Clement's and St. Luke's, Bennet-street	335	1071	1906		This is, perhaps, the largest School in the Kingdom. It cost about £2,300, of which £342 0 10½ was contributed in small sums by the Teachers and Scholars.
St. Stephen's, Bloom-street	181	297	478		
Oldfield-road	139	204	343		
Trinity, King's Head Yard	220	300	520		
Hulme, Duke-street	185	189	374		
All-Saints, Oxford road	196	191	387	30	
Ardwick	60	110	170	25	
	3434	4213	7647	£3078	

Imagem 9. Jornalismo de dados no The Guardian em 1821 (the Guardian)

Outro exemplo seminal na Europa é Florence Nightingale e seu relato fundamental,"Mortalidade no Exército Britânico", publicado em 1858. No seu relato ao Parlamento inglês, ela usou gráficos para defender o aperfeiçoamento do serviço de saúde do exército britânico. O mais famoso é o seu gráfico *crista de galo*, uma espiral de seções em que cada uma representa as mortes a cada mês, que destacava que a imensa maioria das mortes foi consequência de doenças preveníveis em vez de tiros.

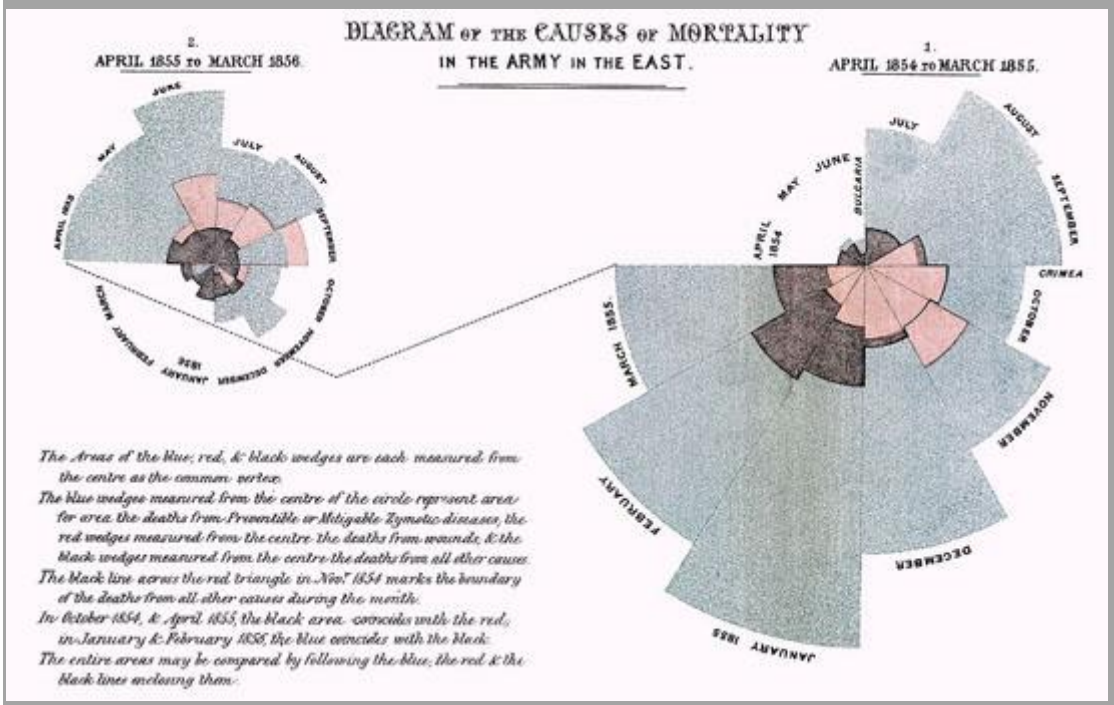


Imagem 10. Mortalidade do exército britânico por Florence Nightingale (imagem da Wikipedia)

Jornalismo de dados e a Reportagem com Auxílio do Computador

Atualmente há um debate sobre "continuidade e mudança" em torno do rótulo "jornalismo de dados" e sua relação com vertentes jornalísticas anteriores que empregaram técnicas computacionais para analisar conjuntos de dados.

Alguns defendem que há diferença entre RAC e jornalismo de dados. Defendem que RAC é uma técnica para apurar e analisar dados de forma a aprimorar uma reportagem (normalmente investigativa), enquanto o jornalismo de dados se concentra na maneira como os dados permeiam todo o processo de produção jornalístico. Nesse sentido, o jornalismo de dados dedica tanta — às vezes, até mais — atenção aos dados propriamente ditos em vez de apenas empregá-los como forma de descobrir ou melhorar uma reportagem. Por isso, vemos o Datablog do The Guardian e o jornal Texas Tribune publicando conjunto de dados lado a lado com as notícias - ou até mesmo apenas os dados sozinhos — para as pessoas analisarem ou explorá-los.

Outra diferença é que, no passado, jornalistas investigativos enfrentariam escassez de informações em relação a questão que estavam tentando responder ou ponto que buscavam esclarecer. Embora, evidentemente, isso continua a acontecer, há ao mesmo tempo uma abundância de informações que os jornalistas não necessariamente sabem como manipular. Não sabem como extrair valor dos dados. Um exemplo recente é o Combined Online Information System, maior banco de dados de gastos públicos do Reino Unido. Este banco de dados foi por muito tempo cobrado pelos defensores da transparência mas, quando foi lançado, deixou jornalistas perplexos e confusos. Como Philip Meyer escreveu recentemente para mim: "Quando a informação era escassa, a maior parte dos nossos esforços eram dedicados à caça e à obtenção de informação. Agora que é abundante, o processamento dessa informação é mais importante."

Por outro lado, alguns ponderam que não há diferença significativa entre o jornalismo de dados e a Reportagem com Auxílio do Computador. Já é senso comum que mesmo as mais modernas técnicas jornalísticas tem um histórico e, ao mesmo tempo, algo de novo. Em vez de debater se o jornalismo de dados é uma novidade completa ou não, uma posição mais produtiva seria considerá-lo parte de longa tradição, mas que agora responde a novas circunstâncias e condições. Mesmo que não haja uma diferença entre objetivos e técnicas, o surgimento do termo "jornalismo de dados" no início do século indica nova fase em que o absoluto volume de dados que estão disponíveis online — combinado

com sofisticadas ferramentas centradas no usuário, plataformas de crowdsourcing e de publicação automática --permitem que mais pessoas trabalhem com mais dados mais facilmente do que em qualquer momento anterior da história.

Jornalismo de dados significa alfabetização de dados do público

A internet e as tecnologias digitais estão alterando fundamentalmente a forma como a informação é publicada. O jornalismo de dados é uma parte do ecossistema de práticas e ferramentas que surgiram em torno dos serviços e sites de dados. Citar e compartilhar fontes e referências faz parte da natureza da estrutura de links da internet, é a forma como estamos acostumados a navegar pela informação hoje em dia. Voltando um pouco no tempo, o princípio na base da fundação da estrutura de links da web é o mesmo princípio de citação usado nos trabalhos acadêmicos. Citar e compartilhar as fontes e dados por trás da notícia é uma das maneiras mais básicas em que o jornalismo de dados pode aperfeiçoar o jornalismo, aquilo que o fundador da WikiLeaks, Julian Assange, chama de "jornalismo científico".

Ao permitir que cada um mergulhe com atenção nas fontes de dados e descubra informação relevante para si mesmo, ao mesmo tempo que checa afirmações e desafia suposições comumente aceitas, o jornalismo de dados efetivamente representa a democratização de recursos, ferramentas, técnicas e métodos antes restritos aos especialistas; seja repórteres investigativos, cientistas sociais, estatísticos, analistas ou outros especialistas. Ao mesmo tempo em que citar e oferecer links para as fontes de dados é característica do jornalismo de dados, estamos caminhando para um mundo em que os dados estão perfeitamente integrados ao tecido da mídia. Jornalistas de dados têm papel importante ao ajudar a diminuir as barreiras para compreensão e imersão nos dados, e aumentar a alfabetização de dados dos seus leitores em grande escala.

No momento, a comunidade de pessoas que se auto-denominam jornalistas de dados é bastante diferente da comunidade mais madura da RAC. Tomara que, no futuro, vejamos laços mais fortes entre essas duas comunidades, da mesma forma que vemos novas organizações não governamentais e organizações de mídia cidadã como a ProPublica e o Bureau de Jornalismo Investigativo trabalharem de mãos dadas com redações tradicionais em investigações. Ao mesmo tempo em que a comunidade de jornalismo de dados possa ter formas inovadoras para entregar dados e apresentar notícias, a abordagem profundamente analítica e crítica da comunidade da RAC tem muito a ensinar ao jornalismo de dados.

— Liliana Bounegru, Centro Europeu de Jornalismo

O jornalismo guiado por dados numa perspectiva brasileira

A partir do final dos anos 2000, as práticas de Jornalismo Guiado por Dados (JGD) não apenas estavam em vias de se estabelecer nas redações da América do Norte e Europa, como também haviam se tornado a principal estratégia de grande parte da imprensa para a recuperação da audiência, que vem caindo há décadas. Pode-se dizer que, hoje, o jornalismo guiado por dados "está na moda". Além da popularização das ferramentas e do apelo comercial de visualizações e outros produtos relacionados ao JGD, foi importante para isso a adoção de políticas de acesso à informação e transparência por governos de todo o mundo. Conhecidos como políticas de "dados abertos" (open data) ou "transparência pública" (open government), estes mecanismos inundaram a Internet com bases de dados antes muito difíceis de se obter. Os jornalistas, portanto, têm hoje o material e as ferramentas para o o JGD ao alcance das mãos.

Serviços online, como Google Drive, Infogr.am, DocumentCloud e CartoDB, apenas para citar alguns, permitem construir, organizar e analisar bancos de dados, bastando um computador e habilidade com a língua inglesa para usá-los. Em maio de 2012, a Presidência da República sancionou a Lei nº 12.527, conhecida como Lei de Acesso à Informação, que obriga todos os órgãos públicos brasileiros a divulgar dados administrativos e a atender a solicitações de informação qualquer cidadão. Estes dois fatores reavivaram o interesse da imprensa brasileira pela aplicação de técnicas computacionais na produção de notícias.

São os próprios repórteres, individualmente, os principais disseminadores dos conceitos de JGD no cenário mundial. Você pode encontrar aqui uma lista de quase cem referências com links para esses trabalhos. No Brasil, existem cada vez mais jornalistas se preparando para atuar nesta especialidade, além dos veteranos da Reportagem Assistida por Computador (RAC) dos anos 1990. Um dos principais indícios deste interesse foi a criação de uma equipe dedicada apenas ao jornalismo guiado por dados na redação de O Estado de São Paulo, pioneira no Brasil, no ano de 2012. Em maio daquele ano, a equipe do Estadão Dadoslançou o Basômetro, um dos primeiros aplicativos jornalísticos brasileiros. Em agosto do mesmo ano, a Folha de S. Paulo passou a hospedar o blog FolhaSPDados, cujo objetivo é criar visualizações gráficas e mapas relacionados às reportagens publicadas no veículo impresso e no site da

empresa. A mesma Folha passou a hospedar o blog Afinal de Contas, dedicado a analisar o noticiário a partir de análises de dados. Outros veículos, como a Gazeta do Povo, do Paraná, têm usado a experiência da redação com jornalismo investigativo na produção de grandes reportagens baseadas em dados. Já o gaúcho Zero Hora, por exemplo, vem se dedicando ao tema do jornalismo guiado por dados e transparência pública através de reportagens e do blog Livre Acesso, inaugurado em 2012 para acompanhar a aplicação da Lei de Acesso à Informação no país.

No campo do jornalismo independente, o principal exemplo é o InfoAmazônia, criado em 2012 pelo Knight Fellow Gustavo Faleiros, em parceria com o webjornal O Eco e a Internews. Em 2013, O Eco criou o Ecolab, um Laboratório de Inovação em Jornalismo Ambiental. A Agência Pública é outra redação independente a aplicar técnicas de JGD, embora o faça esparsamente. Apesar disso, foi responsável por uma das principais contribuições ao JGD no Brasil, por meio de uma parceria com o Wikileaks, para oferecer a biblioteca de documentos diplomáticos PlusD, entre outras bases de dados.

Estes exemplos sugerem estarmos vivenciando os primeiros passos de um movimento de institucionalização das práticas de jornalismo guiado por dados nas redações brasileiras. As bases do sucesso do JGD no país, entretanto, foram lançadas nos anos 1990.

Breve histórico do Jornalismo Guiado Por Dados no Brasil

Ainda durante o governo de Fernando Collor de Mello como presidente do Brasil, o jornalista Mário Rosa, então empregado no Jornal do Brasil, usou o Sistema Integrado de Administração Financeira do Governo Federal (Siafi) para verificar o superfaturamento na compra de leite em pó pela Legião Brasileira de Assistência, então presidida pela primeira-dama, Rosane Collor. Lúcio Vaz relata o caso no livro “A ética da malandragem”:

Assinada pelo jornalista Mário Rosa, a matéria estava completa, com dados jamais vistos, como números de ordens bancárias (Obs.) e de empenhos (reservas feitas no Orçamento da União). Mário havia descoberto o Sistema Integrado de Administração Financeira (Siafi), uma expressão que se tornaria muito conhecida de jornalistas e políticos nos anos seguintes. O acesso a esse sistema, que registra os gastos do governo federal, possibilita fazer uma completa radiografia de todos os pagamentos feitos a empreiteiras, fornecedores, Estados e municípios. Uma mina de diamante para os repórteres.

O jornalismo ganhava uma nova e importante fonte de informação, mais técnica, quase científica. Estavam superados os métodos mais arcaicos de apuração, que envolviam, eventualmente, o enfrentamento com jagunços.

Na época, o acesso a este tipo de base de dados governamental era vedado a cidadãos e jornalistas. O próprio autor da reportagem, Mário Rosa, só pôde realizar pesquisas no Siafi porque o então senador Eduardo Suplicy (PT-SP) lhe emprestou a senha a que tinha direito no desempenho de suas atividades parlamentares. A partir desta e de outras reportagens, o Governo Federal decidiu permitir oficialmente o acesso de jornalistas ao Siafi, tornando-o uma das primeiras bases de dados públicas a serem franqueadas a repórteres no Brasil.

Ascânio Seleme, hoje diretor de redação de O Globo, foi outro repórter que, ainda nos anos 1990, usou a senha de um parlamentar para realizar pesquisas no Siafi, em colaboração com o analista econômico Gil Castelo Branco, diretor da Organização Não-GovernamentalContas Abertas. Estes dois casos são, provavelmente, os primeiros exemplos de JGD na história do jornalismo brasileiro.

Ao longo dos anos 1990, repórteres como Fernando Rodrigues e José Roberto de Toledo, da Folha de S. Paulo, começam a usar técnicas de RAC. A partir de cursos ministrados na redação por tutores do National Institute for Computer-Assisted Reporting dos Estados Unidos, uma subdivisão da associação Investigative Reporters and Editors (IRE/NICAR), estas técnicas foram disseminadas na redação e depois passaram a integrar o currículo do programa de trainees da Folha. A partir de 1998, Fernando Rodrigues começou a construir o banco de dados Políticos do Brasil, lançado na Web e em livro. Em 2002, José Roberto de Toledo se torna um dos sócios-fundadores e vice-presidente da Associação Brasileira de Jornalismo Investigativo (Abraji), entidade fundamental na disseminação dos conceitos e técnicas da RAC no Brasil, tendo treinado mais de quatro mil jornalistas.

A estruturação da Abraji se deu a partir de um seminário promovido pelo Centro Knight para o Jornalismo nas Américas em dezembro de 2002, cujos principais palestrantes foram Brant Houston, autor de um manual de RAC e então diretor do IRE, e Pedro Armendares, da organização mexicana Periodistas de Investigación, que era um dos tutores dos cursos de RAC organizados pela Folha de S. Paulo.

Embora seja uma associação voltada ao jornalismo investigativo em geral, a Abraji atuou na última década principalmente na divulgação da RAC e na defesa do acesso à informação, como uma das entidades integrantes do Fórum de Direito de Acesso a Informações Públicas, criado em 2003, e através de cursos e palestras – dois fatores fundamentais para a emergência do jornalismo guiado por dados ao longo da década de 2000. Duas outras entidades tiveram um papel importante no estabelecimento destas práticas nas redações brasileiras: as organizações não-governamentais Transparência Brasil e Contas Abertas.

A primeira foi criada em 2000 com o objetivo de construir e manter bases de dados sobre financiamento eleitoral, histórico de vida pública e processos sofridos por parlamentares em nível municipal, estadual e federal, notícias sobre corrupção publicadas nos principais jornais brasileiros e sobre o desempenho dos juízes membros do Supremo Tribunal Federal. A segunda entidade, criada em 2005, acompanha o processo de execução orçamentária e financeira da União, através de monitoramento do Siafi, e promove o treinamento de jornalistas para fiscalizar gastos públicos. As bases de dados mantidas pela Transparência Brasil e Contas Abertas permitiram a repórteres realizar reportagens investigativas ao longo da década, quando o acesso às informações do Estado dependia de gestão caso-a-caso junto a órgãos do governo e às redações não investiam neste tipo de recurso.

Um indício da crescente importância das bases de dados para as redações ao longo da década de 2000 está na lista de vencedores do Prêmio Esso de Melhor Contribuição à Imprensa, vencido em 2002 e 2006 por Fernando Rodrigues, pelo arquivo de declarações de bens de políticos brasileiros Controle Público e pelo livro “Políticos do Brasil”, respectivamente; pela Transparência Brasil, em 2006, e pela Contas Abertas, em 2007. Em 2010, a reportagem vencedora do Prêmio Esso, o mais importante do jornalismo brasileiro, foi a série “Díários Secretos”, publicada pela Gazeta do Povo, do Paraná. Para elucidar os movimentos de contratação de funcionários na Assembleia Legislativa do Paraná, os repórteres construíram um banco de dados com todas as nomeações realizadas pela casa entre 2006 e 2010, a partir de diários oficiais impressos. Cruzando os dados no software para criação de planilhas Microsoft Excel, puderam descobrir casos de contratação de funcionários-fantasmas e nepotismo.

Dados são a tábua de salvação da imprensa?

Esse breve histórico sugere que o jornalismo guiado por dados não foi assimilado pelas redações brasileiras através da divulgação promovida por associações profissionais internacionais, imprensa e jornalistas, que tem se intensificado desde 2010, mas vem sendo constituído como prática na cultura jornalística brasileira em paralelo com o processo de informatização. Todavia, pode-se inferir que o interesse crescente de empresas e profissionais do mundo inteiro pelo jornalismo guiado por dados alimenta e incentiva o interesse pelo tema nas redações do Brasil. Números da ferramenta de buscas Google mostram que, a partir de 2010, há um volume crescente de procura por páginas relacionadas ao jornalismo guiado por dados, como pode ser verificado na figura abaixo.

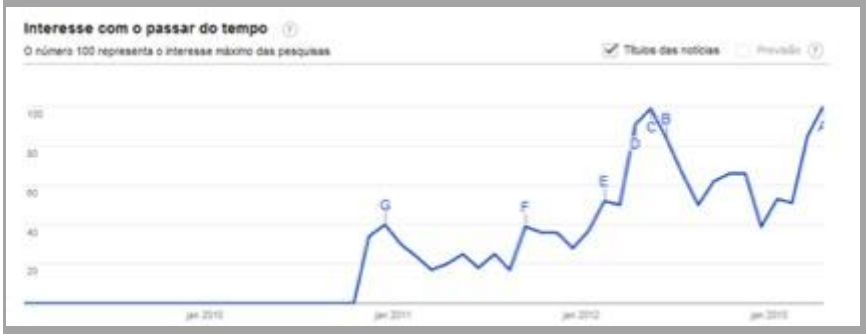


Imagem 11. Volume de buscas por “data journalism” entre janeiro de 2010 e agosto de 2013 (Google Trends, 18 set. 2013)

O primeiro ponto de inflexão na curva de interesse pelo termo “data journalism” (jornalismo de dados) no Google coincide com a criação de uma seção dedicada ao tema, o DataBlog, pelo jornal britânico The Guardian, no final de 2010, e atinge seus dois maiores picos em maio de 2012, quando o jornal americano Seattle Times ganha o prêmio de melhor reportagem em jornalismo guiado por dados da associação Global Editors Network, e em abril de 2013, quando o The Guardian publica no repositório de vídeos YouTube um documentário sobre a história do jornalismo guiado por dados na redação do veículo britânico.

O interesse da imprensa pelo jornalismo guiado por dados, porém, já era evidente dois anos. No dia 11 de janeiro de 2009, a New York Magazine, editada pelo grupo controlador do New York Times, trazia na capa a manchete “O novo jornalismo” e uma foto de duas páginas de cinco membros dos setores de Tecnologias para Redação Interativa, gráficos e multimídia da empresa, acompanhada do subtítulo “O que estes cybergeeks renegados estão fazendo no New York Times? Talvez o salvando”. A matéria conta a história da formação do

grupo de Tecnologias para Redação Interativa dentro da organização, cujos membros, liderados por Aron Pilhofer, são classificados na reportagem como “nerds”, “desenvolvedores/repórteres ou repórteres/desenvolvedores” e “cybergeeks”.

O New York Times é uma das maiores e mais respeitadas empresas de jornalismo do mundo e, para além do sucesso mercadológico, pode ser considerada a própria encarnação da cultura e da mitologia da profissão. O interesse das redações brasileiras e mundiais pelas práticas de jornalismo guiado por dados não está ligado apenas a seus benefícios para as rotinas produtivas e o atendimento do interesse público, mas também à esperança de salvar uma indústria em decadência justamente por efeito das tecnologias digitais.

— Marcelo Träsel, Pontifícia Universidade Católica do Rio Grande do Sul

Existe jornalismo de dados e visualização no Brasil?

Existe jornalismo de dados e de visualização no Brasil? Existe. Está crescendo? Quero acreditar que está, mas não de jeito sistemático e organizado, e não na grande mídia. Sendo honesto, tenho pouca esperança de que estas técnicas e ferramentas vão criar raízes profundas nela — com algumas exceções notáveis —, pelo menos até que não aconteçam algumas mudanças profundas. Aqui estão alguns dos principais motivos:

1. **A alergia ao pensamento lógico, racional, e quantitativo**: Tenha em conta só os seguintes *fatos*: Alguns dos principais jornais do país continuam a publicar horóscopos sem pudor nenhum; as TVs nacionais cobrem aparições de virgens e santos como se fossem fatos, e não ilusões; a principal revista semanal de informação geral é uma fonte substancial de exemplos de grosseira falta de critério estatístico e visual. Estes são só sintomas de um fenômeno subjacente que pode gerar um clima pouco propício para o desenvolvimento da profissão.
2. **A falta de conhecimento dos rudimentos de métodos de pesquisa**: O jornalista brasileiro, como muitos outros de tradição mediterrânea (não se esqueçam que sou espanhol) é, em geral, um escritor-humanista, não um pesquisador-cientista. Como ter os dois perfis é fundamental em qualquer redação, a mídia brasileira precisa hoje menos do primeiro e mais do segundo. Em algumas palestras no país, enquanto comentava exemplos de gráficos ou histórias que poderiam ser melhoradas, falei casualmente: “Aqui podem ver um caso claro de quando é melhor usar a mediana e não a média”, só para ficar chocado pelos olhares de confusão de uma parte da audiência. Se nós não sabemos algo tão básico como o que é uma mediana, o que dizer de desvio padrão, análises de regressão, valor-p, ou métodos bayesianos, tão em moda hoje graças ao sucesso de Nate Silver no The New York Times?
3. **O ensino universitário do jornalismo**: A falta de sabedoria científica e tecnológica é culpa, em grande parte, de um sistema de educação que não tem se adaptado às necessidades dos jornalistas de hoje. Em um mundo em que os dados são cada vez mais acessíveis, em que empresas e governos contratam especialistas para manipular dados antes de apresentá-los ao público, o corpo profissional, que na teoria teria que servir de filtro, carece das habilidades necessárias para cumprir com seu trabalho adequadamente.

Pior, por culpa do próximo ponto que descrevo, também está se blindando contra colegas que possam ajudar nessa tarefa.

4. **A obrigatoriedade do diploma**: A decisão néscia de fazer o diploma universitário de jornalismo obrigatório para o exercício da profissão pode dificultar o emprego de gente com perfil diverso para as redações a não ser em posições de segunda categoria. Alem disso, a exigência do diploma servirá também como desculpa para que os departamentos de Jornalismo não sintam a necessidade de se renovarem para oferecer aos estudantes um melhor treinamento em habilidades conceituais e tecnológicas.

Por que isto é um grande desafio? Hoje é muito difícil achar jornalistas diplomados que, ao mesmo tempo, tenham conhecimentos científicos ou técnicos profundos. Não é só que o jornalista médio não saiba mexer com dados; é que não sabe nem ler uma tabela de números, colocar eles em contexto, e extrair histórias, o que é muito mais importante. Como consequência, a grande mídia precisa contar com especialistas (cientistas, economistas, sociólogos, etc.) como repórteres e editores, e também com profissionais de ciências da computação para colaborar na análise profunda e na gestão de dados.

Me permitam fazer um parêntese neste ponto, e ser muito claro. Um hacker que desenvolve ferramentas para que os cidadãos acessem dados públicos, e que segue as regras éticas próprias da profissão, é tão jornalista quanto o repórter que escreve sobre o último escândalo do Governo, gostem os partidários do diploma obrigatório ou não. Se for contratado por um meio de comunicação, deve ser na posição de jornalista ou, pelo menos, com salário e poder de decisão equivalentes aos de um repórter ou editor no mesmo nível.

Eu leciono infografia e visualização numa escola de Comunicação e Jornalismo. Não conheço nenhum caso de ex-estudante que tenha mostrado o seu diploma para um empregador durante uma entrevista. Os jovens jornalistas são avaliados pelas suas habilidades e conhecimentos.

Por que ter esperança

Na situação atual, portanto, é impensável que mesmo os melhores jornais do país reproduzam o que grandes meios de comunicação dos Estados Unidos — The New York Times, The Washington Post, The Boston Globe, LA Times, ProPublica, The Texas Tribune — estão conseguindo: juntar equipes

multidisciplinares que sistematicamente criam complexos e profundos projetos de jornalismo de dados, visualizações e infográficos interativos.

Essas publicações não consideram o jornalismo de dados acessório ou enfeite, mas elemento central das suas coberturas que não só dão prestígio, mas também atraem leitores. Em recentes palestras, Jill Abramson, diretora executiva do The New York Times, se referiu aos seus departamentos de "news applications" (aplicativos interativos de notícia), multimídia e infografia como pilares essenciais do jornal e do seu rumo futuro. Um dos exemplos mais citados por ela é Snow Fall, uma cobertura multimídia, que de forma muito orgânica mistura texto com imagem, animações e infografias.

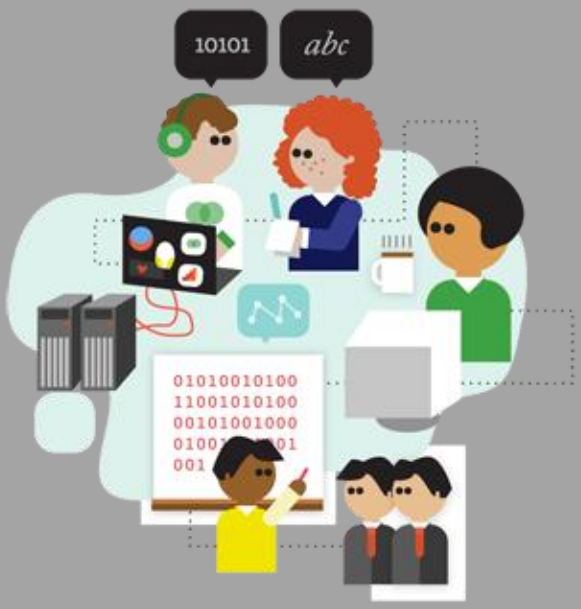
Tendo em conta este panorama desolador, porque acho que o jornalismo de dados e a visualização podem crescer e, por sinal, estão crescendo no Brasil? No que é que baseio minha esperança?

Em primeiro lugar, em corajosas iniciativas dentro dos grandes veículos jornalísticos. São produto geralmente do esforço — não suficientemente reconhecido e sustentado — de pequenos grupos de profissionais com vontade e energia. A equipe do Estadão Dados e o blog Afinal de contas, de Marcelo Soares na Folha de S. Paulo são bons exemplos. São ainda só sementes de um fenômeno que teria que florescer nos próximos anos, mas pelo menos existem. Tem também projetos isolados, esporádicos, feitos por outros veículos da mídia, como as revistas Época e Veja, e jornais como o Correio, na Bahia, o Estadão, e a Folha. Porém, falta dar continuidade a estes casos notáveis.

Em segundo lugar, indivíduos e organizações além da mídia tradicional estão mostrando uma criatividade invejável. Não tenho intenção de ser exaustivo na listagem de projetos que tem chamado a minha atenção nos últimos tempos, mas gostaria de destacar alguns que combinam os dados com um interessante trabalho de design e visualização:InfoAmazonia e sua impressionante combinação de bancos de dados e representação cartográfica; o Radar Parlamentar, que analisa matematicamente os padrões de voto dos congressistas; as propostas resultantes do W3C, como o Retrato da Violência Contra a Mulher no RS e Para Onde vai Meu Dinheiro; e o projeto Escola que queremos.

Quem sabe, talvez sejam estes hackers, desenvolvedores, designers, jornalistas independentes, organizações não governamentais, e fundações os que ocupem um espaço hoje quase vazio, e os que cumpram uma parte importante da tarefa de informação pública que, em tempos anteriores, correspondeu à mídia tradicional. O futuro promete, em qualquer caso. — *Alberto Cairo, Universidade de Miami*

Na Redação



Como o jornalismo de dados encontra espaço em redações pelo mundo? Como os pioneiros do jornalismo de dados convenceram seus colegas de que era uma boa ideia publicar bases de dados ou lançar aplicativos baseados em dados? Os jornalistas devem aprender a programar ou trabalhar em conjunto com desenvolvedores talentosos? Nesta seção olharemos para o papel do jornalismo de dados na Australian Broadcasting Corporation, BBC, Chicago Tribune, Guardian, Texas Tribune e Zeit Online. Aprenderemos como identificar e contratar bons desenvolvedores, como fazer com que as pessoas se comprometam com um tema através de hackatonas (maratonas hackers) e outros eventos, como colaborar além das fronteiras e modelos de negócio para jornalismo de dados.

O que há neste capítulo?

- Aproveitando a expertise dos outros com Maratonas Hacker
- Seguindo o Dinheiro: Jornalismo de dados e Colaboração além das Fronteiras
- Nossas Histórias Vêm Como Código
- Kaas & Mulvad: Conteúdo pré-produzido para comunicação segmentada
- Modelos de Negócio para o Jornalismo de Dados

O Jornalismo de dados da ABC (Australian Broadcasting Corporation)

A Australian Broadcasting Corporation é a empresa pública de radiofusão na Austrália. O orçamento anual gira em torno de um 1 bilhão de dólares australianos, que abastece sete redes de rádio, 60 estações locais de rádio, 3 serviços digitais de televisão, um novo serviço internacional de televisão e uma plataforma online para transmitir a oferta cada vez maior de conteúdo gerado pelo usuário. Na última contagem, havia mais de 4.500 funcionários em tempo equivalente a integral e quase 70% deles produzem conteúdo.

Nós somos uma radiofusora nacional intensamente orgulhosa de nossa independência; embora sejamos financiados pelo governo, temos autonomia garantida por lei. Nossa tradiçao é o serviço público independente de jornalismo. A ABC é reconhecida como a empresa de mídia mais confiável no país.

Estes são tempos estimulantes; sob a gestão de um diretor administrativo (o ex-executivo de jornal Mark Scott), os produtores de conteúdo da ABC foram encorajados a ser "ágeis", como diz o mantra corporativo.

É claro que é mais fácil falar do que fazer.

Mas uma inciativa recente para incentivar essa produção foram competições nas quais os funcionários faziam rápidas apresentações (pitchs) de projetos multiplataforma que gostariam de desenvolver - as ideias vencedoras recebiam o financiamento da empresa. Assim foi concebido o primeiro projeto de jornalismo de dados da ABC.

No começo de 2010, entrei em um desses pitchs para mostrar minha proposta para três dos avaliadores. Eu estava remoendo esta ideia há algum tempo, ambicionando algo como o jornalismo de dados que o, agora legendário, Guardian Datablog estava oferecendo. E isso foi só o começo.

Meu raciocínio era de que, sem dúvida, dentro de 5 anos a ABC teria sua própria divisão de jornalismo de dados. Era inivitável, opinei. Mas a questão era como chegaríamos lá e quem começaria.

Para os leitores que desconhecem a ABC, pensem em uma grande burocracia construída ao longo de 70 anos. Seus carros-chefes sempre foram rádio e televisão. Com o advento do website, na última década a oferta de conteúdo desenvolveu-se em texto, fotos e num grau de interatividade inimaginável no

passado. O espaço virtual estava forçando a ABC a repensar os modos de obter lucro e o seu conteúdo. É claro que é um trabalho contínuo.

Mas algo mais estava acontecendo com o jornalismo de dados. O governo 2.0 (que, como descobrimos, é largamente ignorado na Austrália) estava começando a oferecer novas maneiras de contar histórias até então limitadas a zeros e uns.

Eu disse tudo isso para as pessoas durante minha rápida apresentação. Também disse que precisávamos identificar novos conjuntos de habilidades e treinar jornalistas em novas ferramentas. Precisávamos de um projeto para começar.

E eles me deram o dinheiro para isso.

Em 24 de Novembro de 2011, o projeto multiplataforma online de notícias da ABC foi lançado com Coal Seam Gas by the Numbers (Gás Metano de Carvão em Números).

(Nota da tradução: O gás metano retirado do carvão é um tipo de gás natural usado como combustível. Como foram descobertas grandes e valiosas reservas desse gás na Austrália, e sua exploração pode envolver problemas ambientais, ele se tornou um dos principais assuntos em discussão no país)

Imagem 1. Coal Seam Gas by the Numbers (ABC News Online)

Foi feito com cinco páginas de mapas interativos, visualização de dados e texto. Não era exclusivamente jornalismo de dados, mas um híbrido de diferentes formas de jornalismo nascido da mistura das pessoas na equipe e do tema, um dos assuntos mais quentes na Australia.

A "jóia da coroa" do projeto era um mapa interativo mostrando poços de metano e concessões de exploração na Austrália. Os usuários podem pesquisá-los por localização e alternar entre o layout que mostra a concessão ou os poços. Dando um zoom no mapa, podem acompanhar o responsável pela exploração, a condição do poço, e a sua data de perfuração. Outro mapa mostra onde há exploração do gás próxima a aquíferos australianos.

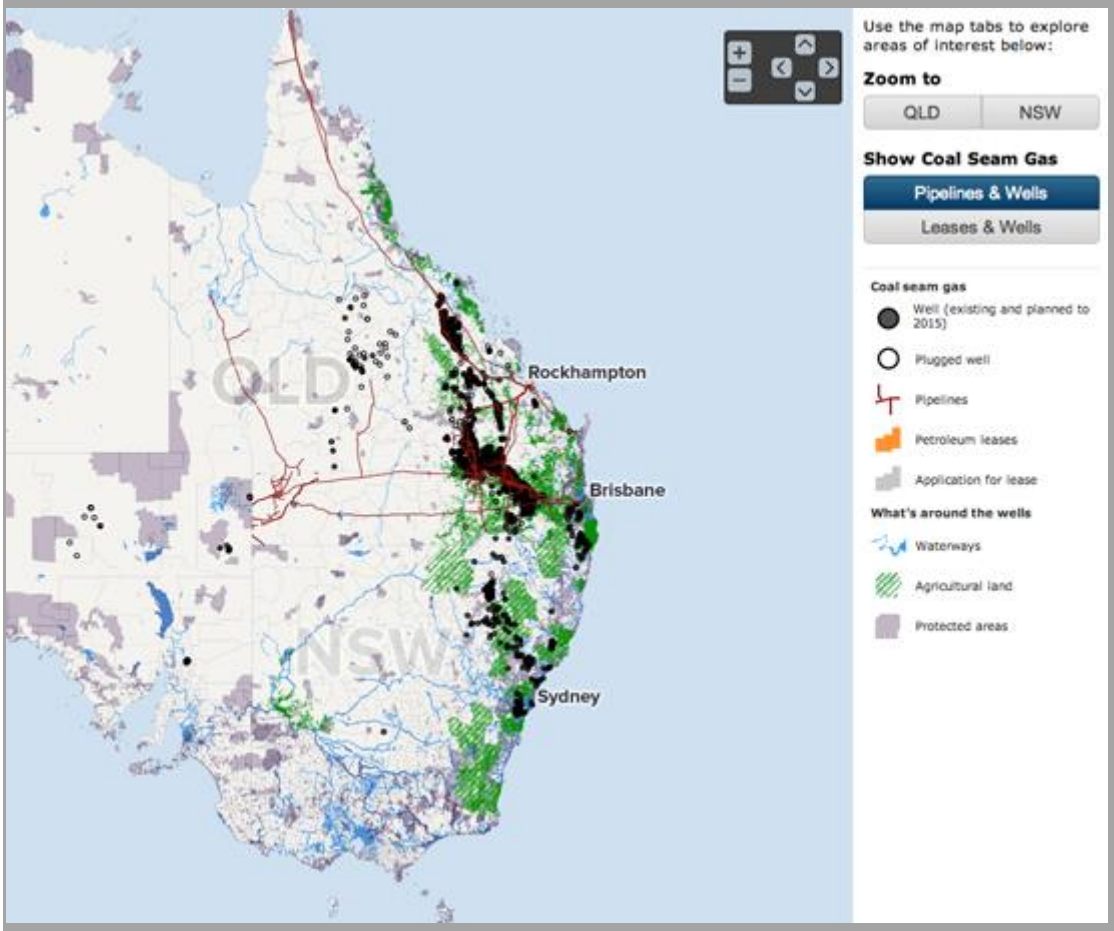


Imagem 2. Mapa interativo de poços de gás e concessões na Austrália (ABC News Online)

Nós fizemos visualizações de dados que trataram especificamente do problema da geração de um subproduto de água com grande concentração de sal. Outra parte do projeto investigou o despejo de produtos químicos numa bacia de rios.

Nosso Time

- Um desenvolvedor web e webdesigner
- Um jornalista que liderou o projeto
- Um pesquisador, trabalhando meio-período, com expertise em extração de dados, planilhas de Excel, e "limpeza" dos dados
- Um jornalista iniciante trabalhando meio período

- Um consultor de produção executiva
- Um consultor acadêmico, com experiência em mineração de dados, visualização de gráficos e habilidades avançadas de pesquisa
- Os serviços de um gerente de projetos e a assistência administrativa da unidade multiplataforma da ABC
- Importante destacar que nós também tivemos um grupo de jornalistas de referência e outras pessoas que íamos consultando conforme precisávamos

De Onde Conseguimos os Dados?

As informações para os mapas interativos foram retiradas de shapefiles (um tipo comum de dado geoespacial) baixados de sites do governo.

Outros dados sobre a água e o sal vieram de diferentes relatórios

As informações sobre os lançamentos de químicos vieram de licenças ambientais emitidas pelo governo.

O Que Aprendemos?

O projeto Coal Seam Gas by the Numbers foi ambicioso no conteúdo e na escala. O mais importante para mim foio que aprendemos e como poderíamos fazer isso de uma maneira diferente da próxima vez"

O projeto juntou um monte de pessoas que normalmente não se encontravam na ABC: em termos leigos, os hacks e os hackers. Muitos de nós não falávamos a mesma língua e nem mesmo acompanhávamos o trabalho do outro grupo. Jornalismo de dados é disruptivo!

Lições práticas:

- Estar num mesmo local é vital para a equipe. Nosso desenvolvedor e designer trabalhou fora da ABC e veio para as reuniões. Isso, definitivamente, não é o ideal! Coloque todos na mesma sala dos jornalistas.
- Nosso consultor de produção executiva também estava em outro andar do prédio. Precisávamos estar muito mais perto para que tivéssemos a possibilidade de "dar uma passada" rapidamente.
- Escolha uma história que é exclusivamente orientada pelos dados

Olhando o Contexto

Grandes organizações de mídia precisam se engajar na construção de capacidades para enfrentar os desafios do jornalismo de dados. Meu palpite é que há um monte de geeks e hackers se escondendo nos departamentos mais técnicos das empresas desesperados para sair. Então precisamos de workshops "hack e hacker" onde os geeks escondidos, jornalistas jovens, desenvolvedores web e webdesigners saiam para brincar com os jornalistas mais experientes e compartilhem habilidades e que sejam orientados.

Ipso facto, o jornalismo de dados é interdisciplinar. Equipes de jornalismo de dados são feitas de pessoas que não tenham trabalhado juntas antes. O espaço digital borrou as fronteiras.

Vivemos em um meio político fraturado e de desconfiança. O modelo de negócio que antes entregava jornalismo profissional independente — imperfeito como ele é — está à beira do colapso. Devemos nos perguntar, como muitos já estão fazendo, como o mundo se parecerá sem um "quarto poder" viável. O intelectual e jornalista norte-americano Walter Lippman observou em 1920 que "admite-se que uma opinião pública forte não pode existir sem o acesso a notícias." Essa declaração não é menos verdadeira agora. No século 21, todo mundo está na blogosfera. É difícil diferenciar mentirosos, dissimulados e grupos de interesse de jornalistas profissionais. Praticamente qualquer site ou fonte pode ser feito de forma a parecer ter credibilidade e ser honesto. As manchetes de confiança estão morrendo na vala. E, neste novo espaço de lixo jornalístico, links podem levar o leitor, infinitamente, a outras fontes mais inúteis, mas de aparência brilhante, que continuam linkando de volta ao salão de espelhos digitais. O termo técnico para isso é: bullshit baffles brains (besteira que confunde cérebros: expressão em inglês para indicar fraudes).

No meio digital, todo mundo é um contador de histórias, certo? Errado. Se o jornalismo profissional — e com isso quero dizer aquele que abraça uma narrativa ética, equilibrada e corajosa na busca da verdade — quiser sobreviver, o ofício deverá reafirmar-se no espaço digital. Jornalismo de dados é apenas mais uma ferramenta que nos permitirá navegar nesse espaço. É onde vamos mapear, remexer, classificar, filtrar, extrair e ver aparecer a história no meio de todos aqueles zeros e uns. No futuro trabalharemos lado a lado com os hackers, os desenvolvedores, os designers e os programadores. É uma transição que requer séria capacitação. Precisamos de gestores de notícias que "saquem" a conexão jornalismo/ meio digital para começar a investir nessa construção.

— Wendy Carlisle, Australian Broadcasting Corporation

Jornalismo de Dados na BBC

O termo "jornalismo de dados" pode abranger uma série de disciplinas e é usado de diversas formas em organizações jornalísticas. Por isso, pode ser útil definir o que entendemos por "jornalismo de dados" aqui na BBC. Em linhas gerais, o termo abrange projetos que utilizam dados para realizar uma ou mais das seguintes ações:

- Permitir que um leitor descubra informação pessoalmente relevante
- Revelar uma história extraordinária e até então desconhecida
- Ajudar o leitor a entender melhor uma questão complexa

Essas categorias podem se sobrepor e, num ambiente on-line, muitas vezes podem se beneficiar de algum nível de visualização.

Faça-o pessoal

No site da BBC News, utilizamos dados para fornecer serviços e ferramentas aos nossos usuários há mais de uma década.

O exemplo mais consistente, publicado primeiramente em 1999, são as nossas Tabelas da rede escolar, que utilizam dados publicados anualmente pelo governo. Os leitores podem encontrar escolas locais, inserindo um código postal, e compará-las de acordo com uma série de indicadores. Jornalistas de Educação também trabalham com a equipe de desenvolvimento para arrastar os dados às suas matérias antes da publicação.

Quando começamos a fazê-las, não havia site oficial que providenciasse uma maneira para o público explorar os dados. Mas agora que o Ministério da Educação tem o seu próprio serviço de comparativo, passamos a nos concentrar mais sobre as histórias que emergem a partir dos dados.

O desafio nesta área deve ser o de proporcionar o acesso aos dados nos quais há um claro interesse público. Um exemplo recente de um projeto que expôs um grande conjunto de dados, normalmente não disponíveis para o público, foi a reportagem especial Every Death on Every Road (Cada morte em Cada estrada). Nós fornecemos uma busca por código postal, permitindo que os usuários encontrem a localização de todas as fatalidades ocorridas nas estradas do Reino Unido na última década.

Nós fizemos visualizações de alguns dos principais fatos e números que emergem a partir dos dados da polícia e, para dar ao projeto uma sensação mais

dinâmica e uma face humana, fizemos uma parceria com a London Ambulance Association e a rádio e TV BBC de Londres para monitorar acidentes em toda a capital à medida que aconteciam. Isto foi relatado online e em tempo real, e também através do Twitter utilizando a hashtag #crash24, e as colisões foram mapeadas à medida que eram relatadas.

Ferramentas Simples

Além de proporcionar maneiras de explorar grandes conjuntos de dados, também tivemos sucesso ao criar ferramentas simples para usuários, que fornecem informações pessoalmente relevantes. Estas ferramentas interessam àqueles sem tempo disponível, que podem não querer uma longa análise. A capacidade de compartilhar facilmente um fato pessoal é algo que tornarmos padrão.

Um exemplo é a nossa ferramenta The world at 7 billion: What's your number (O mundo em 7 bilhões: Qual é o seu número?), publicada para coincidir com a data oficial em que a população mundial ultrapassou 7 bilhões. Ao inserir a data de nascimento, o usuário podia descobrir qual "número" ele era, em termos de população mundial, quando nasceu. Esse número podia ser compartilhado depois através do Twitter ou Facebook. O aplicativo usava dados fornecidos pelo fundo de desenvolvimento da população das Nações Unidas. Era muito popular, e tornou-se o link mais compartilhado em 2011 no Facebook do Reino Unido.

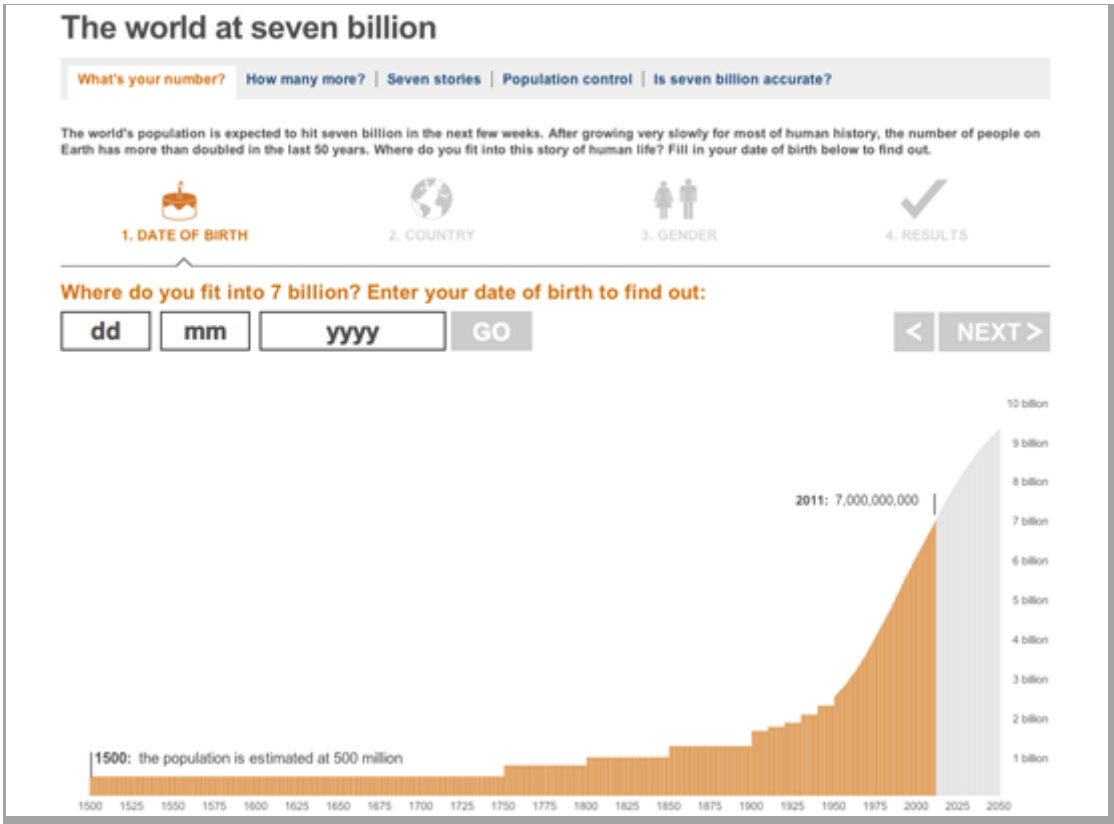


Imagem 3. O mundo em 7 bilhões (BBC)

Outro exemplo recente é o da calculadora do orçamento da BBC, que permitia aos usuários descobrirem quão melhor ou pior será para as suas contas quando a nova lei orçamentária do Reino Unido entrar em vigor — e compartilhar esse dado. Fizemos uma parceria com a empresa de contabilidade KPMG LLP, que nos forneceu cálculos com base no plano de orçamento anual do governo, e então trabalhamos arduamente para criar uma interface atraente que incentivasse os usuários a completarem a conta de quanto economizariam ou gastariam a mais com as novas regras.

Minerando os Dados

Mas onde está o jornalismo em tudo isso? Uma definição mais tradicional do jornalismo de dados é descobrir histórias a partir de dados. Existe informação exclusiva que se esconde na base de dados? Os números são precisos? Será que eles provam ou refutam um problema? Estas são questões que um jornalista de dados ou alguém que pratica Reportagem com Auxílio do Computador (RAC) deve se perguntar. Mas uma quantidade considerável de tempo pode ser gasta para se peneirar conjuntos gigantescos de dados na esperança de encontrar algo excepcional.

Nesta área, descobrimos que é mais produtivo fazer parceria com equipes de investigação ou com programas que têm experiência e tempo para investigar uma história. O programa Panorama da BBC, sobre temas cotidianos, levou meses trabalhando com o Centre for Investigative Journalism, coletando dados sobre os salários do setor público. O resultado foi um documentário televisivo e um relatório on-line especial, Public Sector pay: The numbers, (Salários do Setor Público: Os Números) onde todos os dados foram publicados e visualizados com análises feitas por setor.

Além da parceria com jornalistas investigativos, ter acesso a uma série de jornalistas com conhecimento especializado é essencial. Quando um colega da editoria de negócios analisou dados sobre cortes de gastos anunciados pelo governo do Reino Unido, chegou à conclusão de que o governo estava fazendo parecer com que os cortes fossem maiores do que realmente eram. O resultado foi uma reportagem exclusiva, Making sense of the datacomplementada por uma clara visualização, que ganhou um prêmio da Royal Statistical Society.

Entendendo um problema

Mas o jornalismo de dados não tem de ser apenas encontrar uma informação exclusiva que ninguém conseguiu enxergar antes. O trabalho da equipe de visualização de dados é combinar bom design com uma narrativa editorial clara, de modo a fornecer uma experiência atraente para o usuário. Produzir visualizações dos dados corretos pode ser útil para proporcionar uma melhor compreensão de um problema ou de uma história e nós frequentemente usamos essa abordagem em nossas narrativas na BBC. Uma técnica usada em nosso Rastreador de Pessoas em Busca de Emprego no Reino Unido é um mapa de calor mostrando onde há mais gente procurando emprego ao longo do tempo para fornecer uma visão clara de mudança.

A matéria com dados Eurozone debt web (Rede da dívida da Zona do Euro) explora o emaranhado de empréstimos entre países. Ela ajuda a explicar uma questão complicada de forma visual, usando cor e setas de tamanhos proporcionais às dívidas combinadas com um texto claro. É importante é incentivar o usuário a explorar o recurso ou a seguir uma narrativa, sem fazer com que ele se sinta oprimido pelos números.

Visão Geral da Equipe

A equipe que produz o jornalismo de dados para o site da BBC News é composta por cerca de 20 jornalistas, designers e desenvolvedores.

Além de projetos de dados e visualizações, a equipe produz todos os infográficos e recursos interativos multimídia no site de notícias. Juntos, eles formam um conjunto de técnicas narrativas que chamamos de *jornalismo visual*. Não temos pessoas especificamente identificadas como jornalistas de dados, mas toda a equipe editorial deve ser proficiente no uso de aplicativos de planilhas básicas, tais como Excel e Google Docs, para analisar dados.

Centrais para qualquer projeto de dados são as habilidades técnicas e conselhos dos nossos desenvolvedores e as habilidades de visualização dos nossos designers. Enquanto somos todos "primeiramente" jornalista, ou designer ou desenvolvedor, continuamos a trabalhar duro para aumentar a nossa compreensão e proficiência em cada uma das outras áreas.

Os produtos principais para explorar dados são Excel, Google Docs e Google Fusion Tables. A equipe tem usado também, mas em menor grau, MySQL, bancos de dados do Access e Solr para explorar conjuntos de dados maiores e usado RDF e SPARQL para começar a procurar formas em que podemos modelar eventos usando tecnologias vinculadas aos dados. Desenvolvedores também usam sua linguagem de programação preferida, seja ActionScript, Python ou Perl, para combinar, analisar, ou geralmente separar um conjunto de dados com o qual podem estar trabalhando. Perl é usado para algumas das publicações.

Para explorar e fazer visualização de dados geográficos usamos Google Maps e Bing Maps, além do Google Earth junto com ArcMAP da Esri.

Para gráficos, usamos o pacote Adobe, incluindo After Effects, Illustrator, Photoshop e Flash, embora raramente publicamos arquivos em Flash no site, já que o JavaScript — especialmente JQuery e outras bibliotecas JavaScript, tais como Highcharts, Raphael e D3 — cada vez mais atendem nossos requisitos de visualização de dados.

— Bella Hurrell and Andrew Leimdorfer, BBC

Como trabalha a equipe de aplicativos de notícias no Chicago Tribune

Como trabalha a equipe de aplicativos de notícias no Chicago Tribune

A equipe que produz aplicativos de notícias para o Chicago Tribune é um grupo de felizes hackers incorporados à redação. Trabalhamos próximos aos editores e repórteres para auxiliá-los em: 1) apuração e reportagem, 2) ilustração de matérias online e 3) construção de recursos de web sempre vivos para os leitores da região de Chicago.

É importante a nossa presença dentro da redação. Geralmente o trabalho aparece quando conversamos diretamente com os repórteres. Eles sabem que ficamos felizes em pensar em maneiras de retirar dados de um site governamental ruim, arrancar informações de uma pilha de PDFs, ou, posto de outra maneira, transformar "não-dados" em um material que você possa analisar. É uma espécie de estratégia do nosso grupo; com esse contato, descobrimos outros projetos de dados em potencial.

Diferentemente de outros grupos nesse ramo, nossa equipe foi fundada por gente vinda do ramo de tecnologia que viu no jornalismo uma mudança na carreira. Alguns de nós fizeram mestrado em Jornalismo depois de muitos anos vivendo de programação, outros vieram da comunidade open government.

Trabalhamos com agilidade. Para ter certeza de que estamos sempre na mesma página, toda manhã começa com um encontro de 5 minutos para atualizarmos, uns aos outros, sobre os avanços nos trabalhos. Frequentemente programamos em pares: dois desenvolvedores em um teclado são quase sempre mais produtivos do que dois desenvolvedores em dois teclados. A maioria dos projetos não leva mais de uma semana para ser finalizado, mas, nos trabalhos mais longos, apresentamos todas as semanas os resultado aos participantes do projeto (quase sempre repórteres e editores). "Erre rapidamente" é o nosso mantra. Se você está fazendo errado, é preciso que você saiba o mais rápido possível, especialmente se o trabalho tem um prazo de entrega.

Há um imenso lado positivo em hackear de maneira sistemática, sempre tendo em vista um deadline: estamos sempre atualizando o nosso kit de ferramentas. Toda semana, produzimos rapidamente um aplicativo ou dois e, depois, ao contrário dos trabalhos convencionais com software, podemos deixar o projeto

de lado e seguir para o próximo. É uma alegria que dividimos com os repórteres, e toda semana aprendemos algo novo.

Imagem 4. A equipe de aplicativos de notícias do Chicago Tribune (foto de Heather Billings)

Todas as ideias de aplicativos vêm dos repórteres e editores na redação. Isso, creio, nos diferencia de programadores de outras redações que frequentemente dão suas próprias sugestões. Construímos fortes relações pessoais e profissionais na redação e o pessoal sabe que, quando tem dados, pode vir até nós.

Muito do nosso trabalho na redação é dar suporte ao repórter. Ajudamos a cavar os dados, fazer com que informações em PDFs voltem a ser planilhas, extrair dados de telas de sites, etc. É um serviço que gostamos de prover porque faz com que saibamos com antecedência as reportagens que envolvem trabalhos de dados na redação. Parte desse trabalho vira aplicativo de notícias: um mapa, uma tabela ou, às vezes, um site maior.

Antes, direcionávamos o leitor ao aplicativo a partir da reportagem, o que não resultava em muito tráfego. Hoje os aplicativos ficam próximos ao topo do nosso site e são eles que levam o leitor ao texto, o que funciona bem para ambos: para o aplicativo e a reportagem. Existe uma seção do site para o nosso

trabalho, mas o link não recebe muitas visitas. Isso não nos surpreende. "Ei, hoje eu quero ver dados!" não é algo que todo mundo diz.

Adoramos ter pageviews e adoramos os elogios de nossos colegas, mas não é isso que faz valer o esforço. A motivação deve sempre ser o impacto: na vida das pessoas, na lei, no controle dos políticos, e por aí vai. O texto vai dialogar com as tendências e as humanizar com algumas histórias. Mas o que o leitor deve fazer quando termina a reportagem? Sua família está segura? Suas crianças estão sendo corretamente educadas? Ficamos felizes quando, com o nosso trabalho, ajudamos o leitor a encontrar sua própria história nos dados. Exemplos de trabalhos personalizados e impactantes incluem nossos aplicativos deRelatório de Segurança de Casas de Repouso e de Boletim Escolar.

— Brian Boyer, Chicago Tribune

Bastidores do Guardian Datablog

Quando nós lançamos o Datablog, não tínhamos nenhuma ideia sobre quem estaria interessado em dados brutos, estatísticas e visualizações. Como disse uma pessoa experiente no meu escritório, "por que alguém iria querer isto?"

O Guardian Datablog, que eu edito, era para ser um pequeno blog oferecendo as bases de dados completas por trás de nossas matérias. Agora ele consiste em uma página inicial; buscas de dados sobre países e desenvolvimento global; visualização de dados de artistas gráficos do Guardian e de outras partes da rede e ferramentas para exploração de dados sobre gastos públicos. Todos os dias, usamos as Planilhas do Google para compartilhar todos os dados por trás de nossos trabalho; nós visualizamos e analisamos esses dados e, então, os usamos para criar reportagens no jornal e no site.

Como editor de notícias e jornalista que trabalha com gráficos, o projeto seria um desdobramento lógico do trabalho que eu já estava fazendo, que consistia em acumular bases de dados e "brigar" com elas para tentar dar mais sentido às matérias do dia.

Para nós, a pergunta do começo deste texto hoje está respondida. Os últimos anos têm sido incríveis em relação aos dados públicos. Em seu primeiro dia de governo, o presidente Obama começou a liberar as caixas-pretas de dados governamentais dos Estados Unidos, e seu exemplo foi seguido, em pouco tempo, por outros sites de dados de governamentais ao redor do mundo: Austrália, Nova Zelândia e o site do governo britânico data.gov.uk.

Tivemos o escândalo dos gastos dos membros do parlamento britânico, a mais inesperada matéria de jornalismo de dados — o resultado foi que o governo do Reino Unido está agora comprometido a liberar uma enorme quantia de dados todos os anos.

Tivemos uma eleição geral em que cada um dos principais partidos políticos se comprometeu com a transparência de dados, abrindo o acesso aos nossos dados para o mundo. Jornais dedicaram valorosos espaços em suas colunas para a liberação da base de dados COINS (Combined Online Information System, que guarda milhões de informações sobre gastos públicos).

Ao mesmo tempo, enquanto a web bombardeia mais e mais dados, leitores ao redor do mundo estão mais interessados que nunca nas informações cruas por trás das notícias. Quando lançamos o Datablog, pensamos que a audiência seria

formada por desenvolvedores de aplicativos. Na verdade, ela é formada de pessoas que querem saber mais sobre emissão de carbono, imigração no Leste Europeu, o número de mortes no Afeganistão, ou até mesmo a quantidade de vezes que os Beatles usaram a palavra "amor" em suas canções (613).

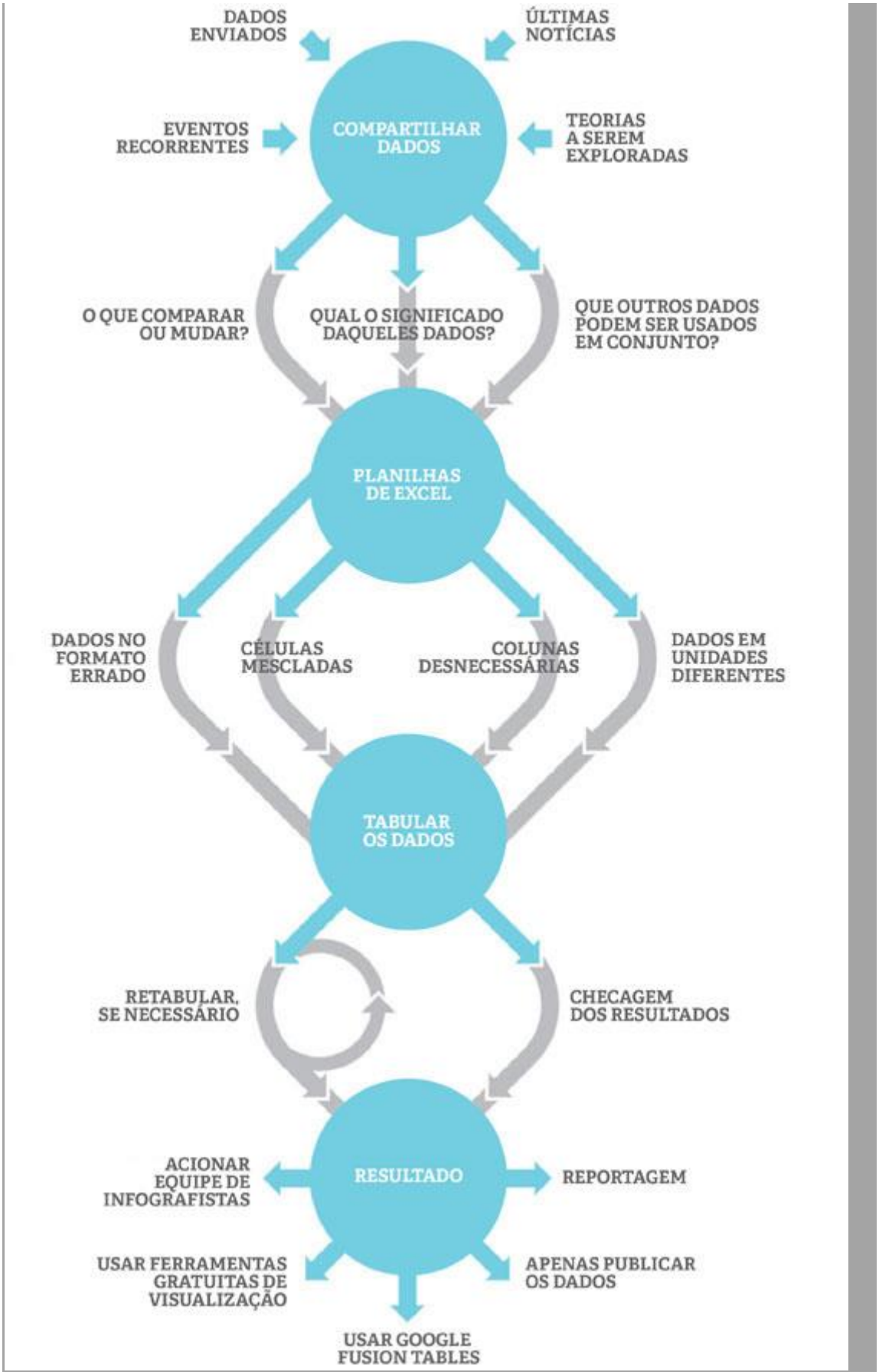


Imagem 5. O processo de produção do Datablog

Gradualmente, o trabalho do Datablog foi aparecendo nas histórias com que nos deparamos e as enriquecendo. Nós fizemos um crowdsourcing (disponibilizamos online) 458 mil documentos relativos aos gastos dos membros do parlamento inglês e analisamos em conjunto com os usuários os dados detalhados sobre as alegações dos parlamentares nos documentos. Ajudamos nossos usuários a explorar bancos de dados relativos a gastos públicos e publicamos os dados por trás das notícias.

Mas a grande mudança para o jornalismo de dados aconteceu na Primavera de 2010, começando com uma planilha: 92.201 linhas de dados, cada uma contendo detalhes de uma ação militar no Afeganistão. Este foi o War Logs (registros de guerra) liberado pelo WikiLeaks. Quer dizer, a primeira parte dele. Houve ainda dois outros episódios em seguida: o do Iraque e o dos cabos. O termo oficial utilizado para nomear o banco de dados das duas primeiras partes foi SIGACTS: Banco de Dados de Ações Significativas dos Estados Unidos (Significant Actions Database).

A organização das notícias está muito ligada à geografia dentro do jornal e à proximidade com a redação. Se você está perto, é mais fácil sugerir pautas e se tornar parte do processo; vendo pelo outro lado, estar fora de vista é estar literalmente fora da cabeça do repórter. Antes do WikiLeaks, nós ficávamos num andar diferente, com quem faz gráficos. Desde o surgimento do WikiLeaks, nós passamos a ficar no mesmo andar, perto da redação. Isso significa que é mais fácil para nós sugerir ideias para as editorias, e faz com que repórteres da redação lembrem-se de nós para ajudá-los com suas reportagens.

Não faz muito tempo, jornalistas eram os guardiões dos dados oficiais. Nós escrevíamos reportagens sobre números e soltávamos para um público agradecido, que não estava interessado nas estatísticas puras. A ideia de liberarmos informações brutas nos jornais era um anátema.

Agora a dinâmica mudou completamente. Nosso papel é nos tornarmos intérpretes; ajudando as pessoas a compreenderem os dados, ou até mesmo apenas publicá-los, já que eles são interessantes por si mesmos.

Mas os números sem análise são só números, e é aí que entramos. Quando o Primeiro Ministro britânico declarou que os protestos em Agosto de 2011 não tinham a ver com a pobreza, nós fomos capazes de mapear os endereços dos manifestantes e verificá-los com indicadores de pobreza a fim de mostrar a verdade por trás desta declaração.

Há um processo por trás de toda reportagem ligada ao jornalismo de dados. Ele muda constantemente conforme usamos novas ferramentas e técnicas. Algumas pessoas dizem que a resposta é se tornar um super hacker, escrever códigos, e imergir no SQL. Você pode escolher esta abordagem. Mas muito do trabalho que fazemos utiliza apenas o Excel.

Primeiramente, localizamos os dados ou os recebemos de uma variedade de fontes, das últimas notícias, de dados do governo ou das pesquisas de jornalistas, e por aí vai. Começamos então a ver o que fazer com esses dados; é preciso misturá-los com outra base de dados? Como podemos demonstrar as mudanças ocorridas ao longo do tempo? As planilhas muitas vezes devem ser organizadas — todas as colunas esquisitas e as células estranhamente mescladas realmente não ajudam. E isso assumindo que não estejam em PDF, o pior formato para dados conhecido da humanidade.

Muitas vezes, dados oficiais vem com códigos oficiais; cada escola, hospital, distrito eleitoral, e autoridade local tem um único código identificador.

Os países têm também (o código do Reino Unido é GB, por exemplo). Eles são úteis caso se deseje começar a misturar as bases de dados, e é impressionante a quantidade de maneiras diferentes de escrever uma mesma informação que podem atrapalhar a análise. Há Burma e Myanmar, por exemplo, ou o Condado Fayette nos Estados Unidos (há 11 destes nos Estados de Georgia e West Virginia). Códigos nos permitem fazer comparações nesses casos em que um dado se confunde com outro.

Ao final do processo está o resultado: será uma reportagem, um gráfico, ou uma visualização e, quais as ferramentas que iremos utilizar? As ferramentas mais utilizadas por nós são as grátis com as quais podemos rapidamente produzir algo. Nossa equipe de desenvolvimento produz os gráficos mais sofisticados.

Isto significa que nós comumente utilizamos o Google charts (programa de gráficos do Google) para fazer pequenos gráficos de linha ou de pizza, ou o Google Fusion Tables para criar mapas mais rápidamente e facilmente.

Isto pode parecer novo, mas realmente não é.

Na primeira versão do Manchester Guardian (no sábado, 5 de maio de 1821), as notícias estavam na página de trás, assim como em todos os jornais daquela época. O primeiro item da capa era um aviso sobre um labrador desaparecido.

Por entre as reportagens e excertos de poemas, um terço dessa página de trás foi tomado por... fatos. Uma tabela completa mostrava os custos das escolas naquela área, "algo nunca antes informado ao público", escreve "N.H.".

N.H. queria seus dados publicados porque, caso contrário, os fatos seriam reportados por clérigos destreinados. Sua motivação era que "O conteúdo de tal informação é valoroso; porque, sem saber em que medida a educação... prevalece, as melhores opiniões que podem ser formadas sob a condição e o progresso futuro da sociedade serão necessariamente incorretas." Em outras palavras, se as pessoas não sabem o que está acontecendo, como a sociedade pode melhorar?

Não consigo pensar numa análise melhor para o que nós estamos tentando fazer. O que antes era reportagem para a página de trás do jornal pode, hoje, ser a notícia da primeira página.

— Simon Rogers, The Guardian

Jornalismo de dados no Zeit Online

O projeto PISA based Wealth Comparison é uma visualização interativa que permite comparar padrões de vida em diferentes países. Ele utiliza dados do PISA 2009, um abrangente relatório da OCDE sobre o nível de educação no mundo, publicado em dezembro de 2010. O relatório é baseado em um questionário aplicado a crianças de quinze anos sobre as condições de vida delas em casa.

A ideia foi analisar e visualizar esses dados com o objetivo de fornecer uma forma original de comparar os padrões de vida em diferentes países.

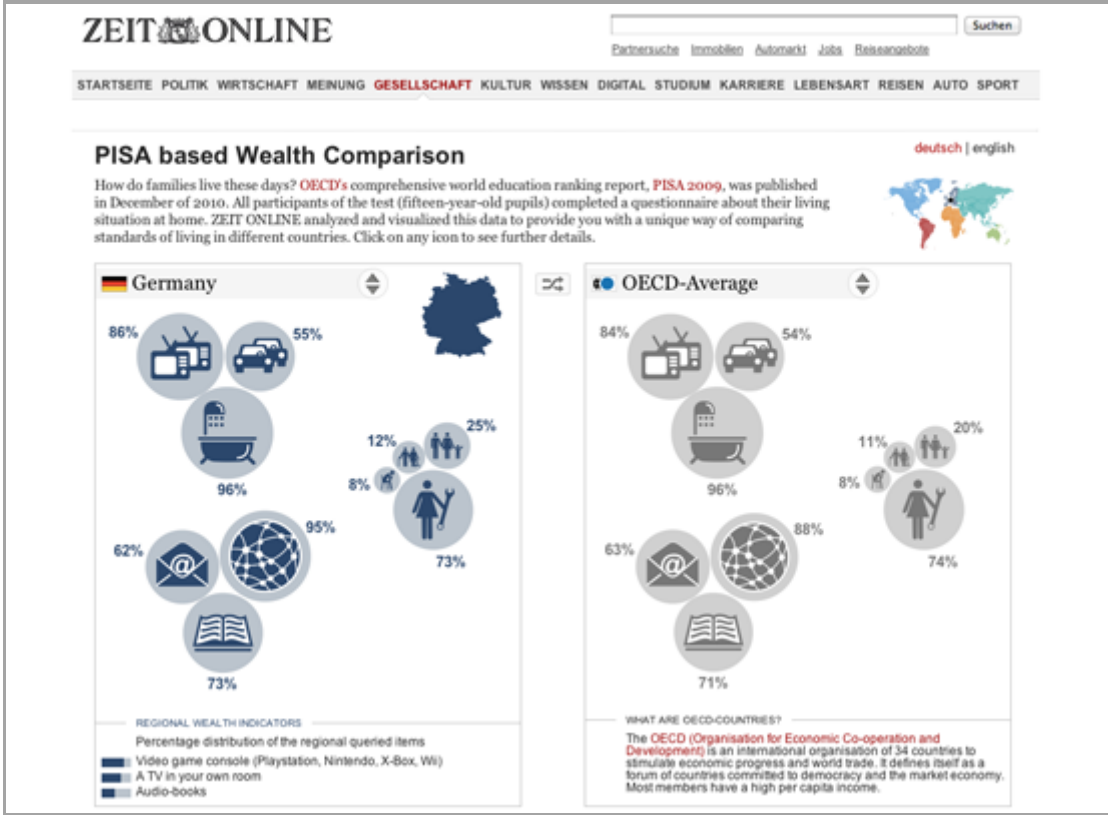


Imagem 6. PISA based Wealth Comparison (Zeit Online)

Primeiro, a nossa equipe editorial decidiu quais fatos pareciam úteis para tornar os padrões de vida comparáveis e quais deveriam ser visualizados, incluindo:

- Riqueza (número de TVs, carros e banheiros disponíveis em casa)
- Situação familiar (se os avós estão vivendo com a família, percentual de famílias com apenas um filho, desemprego dos pais, e condição de trabalho das mães)

- Acesso a fontes de conhecimento (Internet em casa, frequência no uso de email e quantidade de livros possuídos)
- Três indicadores adicionais sobre o nível de desenvolvimento de cada país

Com a ajuda da equipe de design, esses fatos foram traduzidos em ícones auto-explicativos. Uma programação de design foi construída para fazer comparações entre diferentes países, olhando para eles como se fossem cartas de baralho.

Depois, nós entramos em contato com o pessoal do German Open Data Network para procurar desenvolvedores que poderiam ajudar com o projeto. Essa comunidade de pessoas altamente motivadas nos sugeriu Gregor Aisch, um talentoso designer de informação, para codificar os aplicativos que fariam os nossos sonhos se tornar realidade (isso sem utilizar o Flash, o que era muito importante para nós!). Gregor criou uma visualização interativa de alta qualidade com um lindo estilo de bolhas, baseado noRaphaël-Javascript Library.

O resultado da nossa colaboração foi um sucesso interativo que gerou muito tráfego na internet. É simples comparar quaisquer dois países, o que faz o aplicativo útil como uma ferramenta de referência. Nós podemos reutilizá-lo no nosso trabalho editorial diário. Por exemplo, se estamos cobrindo algo relacionado à situação de vida na Indonésia, podemos rapidamente e facilmente embutir um gráfico comparando a situação de vida da Indonésia com a da Alemanha. O know-how ganho pela a nossa equipe foi um grande investimento para projetos futuros.

No Zeit Online, nós descobrimos que nossos projetos de jornalismo de dados têm aumentado o tráfego e ajudado a envolver o público de novas formas. Por exemplo, houve muita cobertura sobre a situação da usina nuclear em Fukushima depois do tsunami no Japão. Depois que o material radioativo escapou da usina nuclear, os moradores que estavam em um raio de 30 quilômetros foram retirados de suas casas. As pessoas podiam ler um monte de coisas sobre as evacuações. O Zeit Online encontrou uma forma inovadora para explicar o impacto ao público alemão. Nós perguntamos: quantas pessoas moram perto de uma usina nuclear na Alemanha? Quantas vivem em um raio de 30 quilômetros? Um mapa mostra quantas pessoas poderiam ter de deixar suas casas se algo semelhante acontecesse na Alemanha. O resultado: muitos acessos; na verdade, o projeto tornou-se viral nas mídias sociais. Projetos de jornalismo de dados podem ser relativamente fáceis de se adaptar a outros

idiomas. Nós criamos uma versão em inglês sobre a proximidade de usinas nucleares nos EUA, que foi uma grande fonte de tráfego. Organizações de notícias querem ser reconhecidas como fontes confiáveis e de autoridade entre os leitores. Nós achamos que projetos baseados no jornalismo de dados, combinados com o fato de que permitimos aos nossos leitores olhar e reutilizar os dados brutos, nos traz um elevado grau de credibilidade.

Há dois anos o departamento de pesquisa e desenvolvimento e o redator-chefe do Zeit Online, Wolfgang Blau, defendem o jornalismo de dados como uma importante maneira de contar histórias. Transparência, credibilidade e envolvimento do usuário são partes importantes da nossa filosofia. É por isso que o jornalismo de dados é uma parte natural do nosso trabalho atual e futuro. Visualizações de dados podem agregar valor para a recepção de uma matéria e são uma forma atraente para toda a equipe editorial apresentar conteúdos.

Por exemplo, em 9 de novembro de 2011, o Deutsche Bank prometeu parar de financiar a fabricação de bombas de fragmentação. Mas de acordo com um estudo da organização sem fins lucrativos Facing Finance, mesmo depois da promessa, o banco continuou a aprovar empréstimos para os produtores de bombas de fragmentação. A nossa visualização de dados mostra aos leitores os vários fluxos desse dinheiro. As diferentes subsidiárias do Deutsche Bank estão dispostas no topo; as companhias acusadas de envolvimento na construção de bombas de fragmentação, embaixo. No meio, os empréstimos estão representados ao longo de uma linha do tempo. Sobre os círculos são mostrados os detalhes de cada transação. Claro, a história poderia ter sido contada somente em texto. Mas a visualização permite aos nossos leitores entender e explorar as relações financeiras de uma forma mais intuitiva.

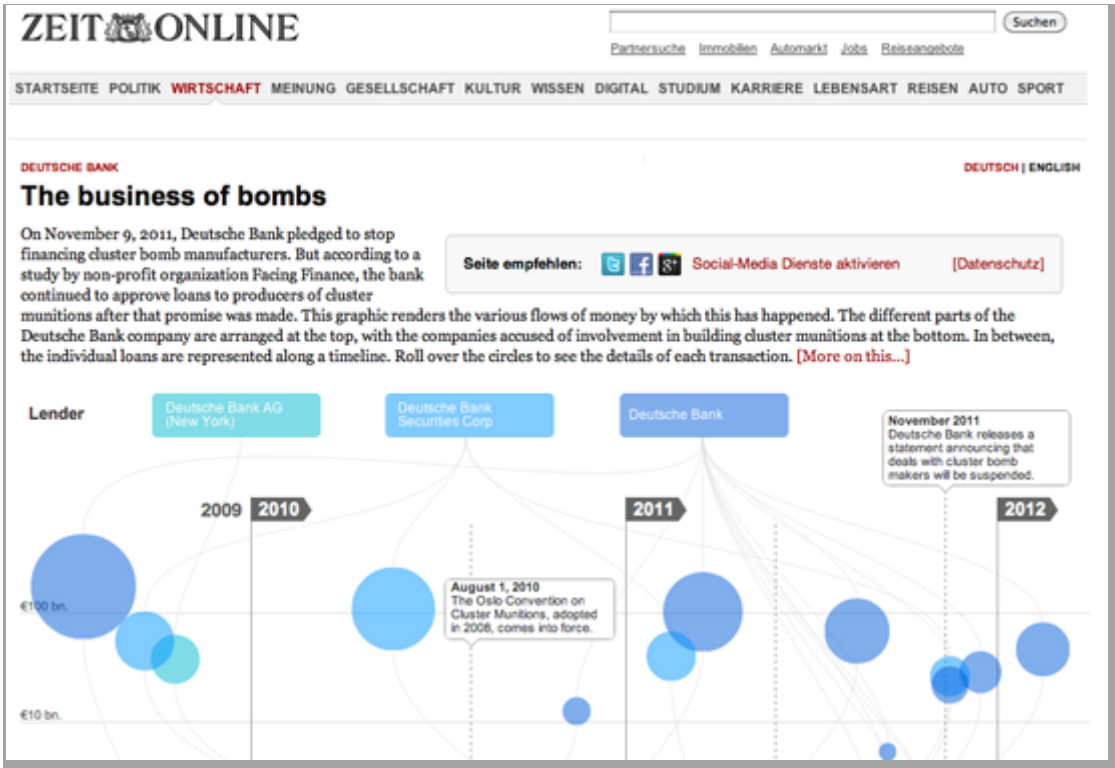


Imagem 7. O negócio das bombas (Zeit Online)

Outro exemplo: a Agência Federal de Estatística da Alemanha tem publicado uma grande base de dados sobre estatísticas vitais para o país, incluindo vários modelos de cenários demográficos até 2060. A típica maneira de representar isso é uma pirâmide populacional, tal como publicada pela agência.

Com os nossos colegas do departamento de ciência, tentamos dar aos nossos leitores uma forma melhor para explorar as projeções de dados demográficos sobre o futuro da nossa sociedade. Na nossa visualização, apresentamos um grupo estatisticamente representativo de 40 pessoas de diferentes idades desde 1950 até 2060. Elas estão organizadas em oito grupos diferentes. Parece uma foto da sociedade alemã em diferentes momentos. Os mesmos dados visualizados em uma tradicional pirâmide populacional dão apenas uma sensação muito abstrata da situação, mas ter um grupo com crianças, jovens, adultos e idosos faz com que nossos leitores possam relacionar os dados com mais facilidade. Você precisa somente apertar play para iniciar uma viagem através de 11 décadas. Você pode também digitar sua data de nascimento e o sexo para se tornar parte do grupo: para ver a sua viagem demográfica através das décadas e a sua própria expectativa de vida.

— Sascha Venhor, Zeit Online

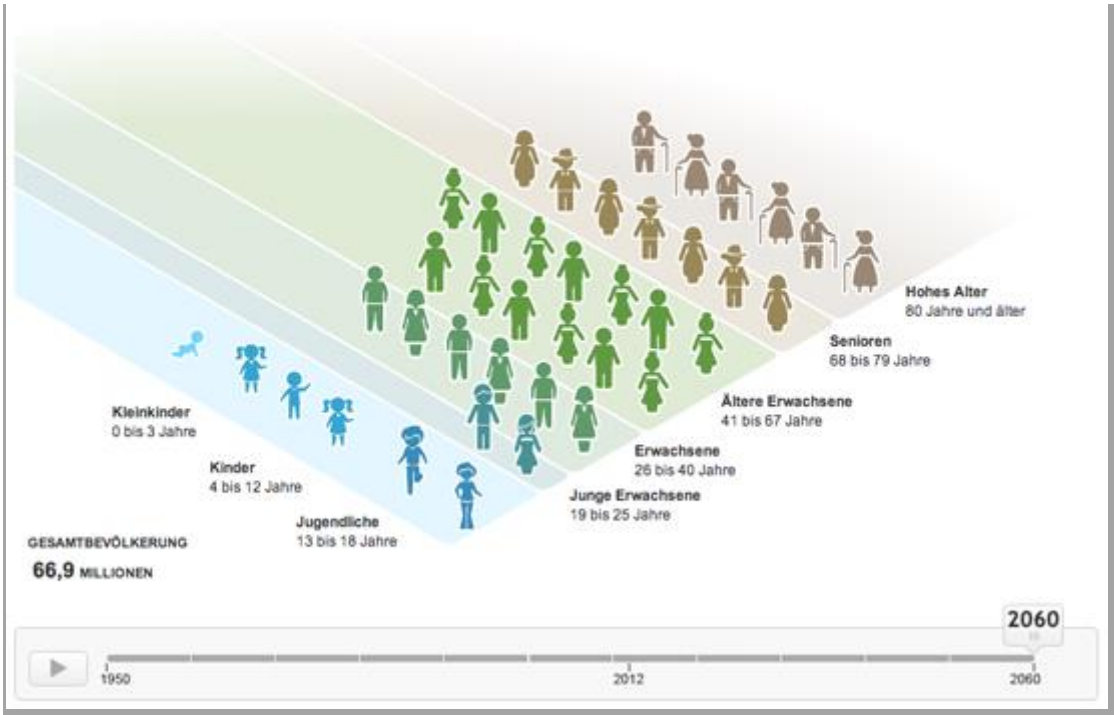


Imagem 8. Visualização de dados demográficos (Zeit Online)

0 0

Como contratar um hacker

Uma das coisas que jornalistas me perguntam regularmente é "como conseguir um programador para me ajudar com meu projeto?" Não se engane em pensar que este é um processo de mão única; hackers com consciência cívica e aficcionados por dados geralmente têm a mesma ansiedade para manter contato com jornalistas.

Jornalistas são usuários acima da média de serviços e ferramentas de dados. Do ponto de vista de desenvolvedores, jornalistas pensam fora da caixa para usar ferramentas de dados em contextos que desenvolvedores nem sempre consideraram (o feedback é inestimável!). Eles também ajudam a construir cenários, repercutir projetos e a torná-los relevantes. É uma relação simbiótica.

Felizmente, isso significa que se você estiver querendo contratar um hacker ou procurando por possíveis colaborações com um orçamento limitado, é mais que provável que haja alguém por aí interessado em te ajudar.

Então como você vai achá-los? Aron Pilhofer, do New York Times, responde:

Você pode achar que sua empresa já tem pessoas com todas as habilidades necessárias, mas essas pessoas não estão necessariamente dentro de sua redação. Perambule por aí, visite os departamentos de tecnologia e de TI e você possivelmente vai encontrar algo brilhante. É importante também saber apreciar a cultura de programação: encontre alguém que tenha um computador como esse aqui... Figure 9... e daí provavelmente você terá o que está procurando.

Imagem 9. Figurinha carimbada: hackers são geralmente fáceis de serem notados (foto de Lucy Chambers)

Mais algumas ideias de como fazer isso:

Publique em sites de emprego

Identifique e publique em sites voltados a desenvolvedores que trabalham com diferentes linguagens de programação. Por exemplo, Python Job Board.

Faça contato com listas de e-mail relevantes

Por exemplo, a lista de email do NICAR-L e do Data Driven Journalism

Contate organizações relevantes

Por exemplo, se você quiser arrancar ou depurar dados da web, você pode contatar uma organização como a Scraperwiki, que tem uma grande lista de contatos de programadores motivados e confiáveis.

Entre em redes e grupos relevantes

Procure por iniciativas como o Hacks/Hackers que reúnem jornalistas e aficionados por tecnologia. Grupos de Hacks/Hackers estão se espalhando pelo mundo. Você também pode tentar publicar algo na lista de empregos deles.

Grupos locais

Você pode tentar uma busca rápida por uma área de conhecimento na sua região (por exemplo, "javascript" + "london"). Sites como o Meetup.com também são um excelente lugar para se começar.

Competições e Hackathonas

Tendo ou não um prêmio em dinheiro, competições de visualização e de aplicativos e maratonas de desenvolvimento são geralmente solo fértil para a colaboração e para estabelecer conexões.

Pergunte a um geek!

Geeks andam com outros geeks. O boca a boca é sempre um bom modo de achar gente boa com quem se trabalhar.

> *— Lucy Chambers, Open Knowledge Foundation*

Habilidades Hacker

Depois de achar um hacker, como você vai saber se ele é bom? Nós pedimos a Alastair Dant do Guardian suas dicas sobre como identificar um bom hacker:

Eles codificam de todas as formas

Quando se trata de prazos, é melhor ser um mediano versátil que um mestre de uma coisa só. Novos aplicativos requerem esmiuçar dados, gráficos dinâmicos e obstinação.

Eles vêem as coisas dentro do contexto mais amplo

Abordagens holísticas favorecem a narrativa a detalhes técnicos. Eu preferiria ouvir uma nota tocada com sentimento a um virtuosismo incessante em escalas obscuras. Descubra o quão feliz a pessoa fica ao ter de trabalhar junto com um designer.

Eles contam uma boa história

Apresentações narrativas requerem organizar coisas no espaço e tempo. Descubra de qual o projeto eles têm mais orgulho e peça a eles que mostrem o caminho que fizeram para construir o projeto. Isso revelará tanto sobre a habilidade deles em se comunicarem quanto sobre o conhecimento técnico que possuem.

Eles dialogam ao longo dos processos

Fazer coisas rápido requer grupos mistos trabalhando em função de objetivos comuns. Cada participante deve respeitar seus colegas e estar

disposto a negociar. Imprevistos geralmente necessitam de rápido replanejamento e compromisso coletivo.

Eles se ensinam

A tecnologia move-se rápido. É uma luta manter-se atualizado. Tendo encontrado bons desenvolvedores de todos os tipos de formação, posso dizer que o traço mais comum entre eles é a disposição para aprender coisas novas e necessárias ao projeto.

> *— Lucy Chambers, Open Knowledge Foundation*

Como achar o desenvolvedor dos sonhos

A diferença de produtividade entre um bom desenvolvedor e um ótimo não é linear - é exponencial. Contratar bem é extremamente importante. Infelizmente, contratar bem também é muito difícil. É tarefa dura vetar candidatos se você não for um gerente técnico experiente. Junte a isso os salários que as empresas de jornalismo podem pagar e você tem um desafio e tanto.

No Tribune, nós recrutamos a partir de dois ângulos: um apelo emocional e outro técnico. O apelo emocional é este: jornalismo é essencial para uma democracia efetiva. Trabalhe aqui e você pode mudar o mundo. Tecnicamente, nós promovemos o quanto você aprenderá. Nossos projetos são pequenos, rápidos e frequentes. Cada projeto usa um novo conjunto de ferramentas, uma nova linguagem, um novo assunto (segurança de incêndio, o regime de pensões), que você tem que aprender. A redação é a prova de fogo. Eu nunca gerenciei um grupo que tenha aprendido tanto e tão rápido quanto a nossa equipe.

Sobre onde procurar, nós tiramos a sorte grande achando ótimos hackers na comunidade de open government (pró-transparência governamental). A lista de email do Sunlight Labs é onde nerds que fazem o bem mas com empregos sacais de dia passam a noite. Outra fonte com potencial é o Code for America. Todo ano, um grupo de colegas emerge do CfA, procurando pelo seu próximo grande projeto. De bônus, o CfA tem um processo de seleção rigoroso: eles já peneiraram para você. Atualmente, jornalistas interessados por programação também estão saindo

das escolas de jornalismo. Eles são novos, mas têm potencial gigantesco.

Por último, contratar desenvolvedores não é o suficiente. Você precisa de gerenciamento técnico. Um desenvolvedor solitário (especialmente recém-saído da escola de jornalismo e sem experiência de trabalho) irá tomar muitas decisões ruins. Até mesmo o melhor programador, quando deixado a seus próprios aparelhos, irá escolher o que é tecnicamente interessante em vez do que é mais importante para o seu público.

Chame isso de contratar um "editor de aplicativos de notícia", um "gerente de projetos" ou o que seja. Assim como escritores, programadores precisam de editores, acompanhamento e alguém que dialogue com eles em função de fazer um programa no prazo.

— Brian Boyer, Chicago Tribune

Aproveitando a expertise dos outros com Maratonas Hacker

Em Março de 2010, a organização de cultura digital SETUP, na cidade holandesa de Utrecht, formulou um evento chamado Hacking Journalism. O evento foi organizado para encorajar maior colaboração entre desenvolvedores e jornalistas.

"Nós organizamos hackatonas (competições hacker) para fazer aplicativos legais, mas nós não conseguimos reconhecer histórias interessantes nos dados. O que nós construímos não tem relevância social," disseram os programadores. "Nós reconhecemos a importância de jornalismo de dados, mas não temos todas as habilidades técnicas para construir as coisas que queremos," disseram os jornalistas.

Imagem 10. Jornalistas e desenvolvedores na RegioHack (foto por Heinze Havinga)

Trabalhando em um jornal regional, não via dinheiro ou incentivo para contratar um desenvolvedor para a redação. Jornalismo de dados ainda era uma incógnita para os jornais holandeses na época.

O modelo de hackathona era perfeito; um ambiente relax para colaboração, com bastante pizza e bebidas energéticas. A RegioHack foi uma hackathona organizada pelo meu empregador, o jornal regional De Stentor, nossa publicação irmã, TC Tubantia, e o Saxion Hogescholen Enschede, que cedeu o espaço para o evento.

O combinado foi: qualquer um poderia se alistar para uma hackathona de 30 horas. Nós providenciaríamos comida e bebidas. Tivemos como meta 30 participantes, os quais dividimos em seis grupos. Os grupos se focaram em tópicos diferentes, como crime, saúde, transporte, segurança, envelhecimento e poder. Para nós, os três maiores objetivos eram:

Encontrar matérias

Para nós, jornalismo de dados é algo novo e desconhecido. A única maneira que temos de provar o quanto é útil é através de reportagens bem elaboradas. Nós queríamos produzir pelo menos três matérias usando dados.

Criar conexões entre as pessoas

Nós, os jornalistas, não sabemos como jornalismo de dados é feito e não fingimos saber. Colocando jornalistas, estudantes e programadores numa mesma sala por 30 horas, queremos que eles compartilhem conhecimento e insights.

Organizar um evento social

Jornais não organizam muitos eventos socias, ainda mais hackathonas. Nós queríamos testar como um evento conseguiria gerar resultados. Na verdade, o evento bem que poderia ter sido tenso: 30 horas com estranhos, um monte de jargão, fritando o cérebro com questões básicas, e trabalhando fora da sua zona de conforto. Fazendo da hackatona um evento social (lembra da pizza e das bebidas?), nós buscamos criar um ambiente no qual jornalistas e programadores poderiam sentir-se confortáveis e colaborar efetivamente.

Antes do evento, a publicação TC Tubantia fez uma entrevista com a viúva de um policial que havia escrito um livro sobre os anos de serviço de seu marido. Ela também tinha um documento com todos os homicídios registrados no leste da Holanda, atualizado pelo seu marido desde 1945. Normalmente, nós publicaríamos o documento no nosso site. Desta vez, optamos por fazer um aplicativo usando o software Tableau. Nós também blogamossobre como isto foi produzido no nosso site do RegioHack.

Durante a hackathona, um grupo veio com o assunto sobre o que aconteceria com as escolas com o envelhecimento da população em nossa região. Fazendo a visualização de projeções futuras, entendemos quais

cidades iriam ter problemas em alguns anos com declínio de matrículas. A partir deste insight, nós escrevemos uma matéria sobre como isso iria afetar as escolas em nossa região.

Nós também iniciamos um projeto muito ambicioso chamado De Tweehonderd van Twente (os Duzentos de Vinte) para determinar quem tinha mais poder em nossa região e construir uma base de dados das pessoas mais influentes. Através de um cálculo no estilo Google — sobre quem tem mais conexões com organizações poderosas — uma lista das pessoas mais influentes da região foi criada. Isso poderia levar a uma série de matérias, além de ser uma ferramenta poderosa para os jornalistas. Quem tem conexões com quem? Você pode fazer questões para essa base de dados e usar isto em seu cotidiano. A base de dados também tem valor cultural. Artistas já perguntaram se poderiam usar o banco de dados quando finalizado, para fazer instalações de arte interativa.

Imagem 11. Novas comunidades em volta do jornalismo de dados (foto por Heinze Havinga)

Depois da RegioHack, percebemos que jornalistas consideram o jornalismo de dados um incremento viável para o jornalismo tradicional. Meus colegas continuaram usando e desenvolvendo as técnicas aprendidas naquele dia para criar projetos técnicos mais ambiciosos, como um banco de dados de custos administrativos de uma moradia. Com esses dados, fiz um mapa interativo usando Fusion Tables. Nós pedimos para nossos leitores

brincarem um pouco com os dados e colaborarem com resultados no sitehttp://bit.ly/scratchbook-crowdsourcing, por exemplo. Depois de várias questões sobre como nós fizemos o mapa usando Fusion Tables, gravei um video tutorial.

O que nós aprendemos? Aprendemos muito, mas também encontramos muitos obstáculos, como esses quatro:

Por onde começar: pela questão ou pelos dados?

Quase todos os projetos travaram quando buscaram por informações. Na maior parte das vezes, eles começaram com uma questão jornalística. Mas e então? Quais dados estão disponíveis? Onde podemos achá-los? E, quando achar esses dados, você poderá responder sua pergunta com eles? Jornalistas geralmente sabem onde achar informação quando fazem pesquisa para uma matéria. Com jornalismo de dados, a maioria dos jornalistas não sabem qual informação está disponível.

Pouco conhecimento técnico

Jornalismo de dados é uma disciplina bem técnica. Algumas vezes você tem de saber arrancar os dados das fontes, outras vezes você tem de fazer uma programação para visualizar os seus resultados. Para fazer um excelente jornalismo de dados, você precisa de duas coisas: os insights jornalísticos de um jornalista experiente e o conhecimento técnico de alguém bem versado digitalmente. Durante a RegioHack, isto não foi comum.

Isto é notícia?

Participantes na sua maioria usaram um conjunto de dados para descobrir notícias, em vez de procurar interconexões entre fontes diferentes. Isso acontece por que você necessita algum conhecimento estatístico para checar as notícias vindas do jornalismo de dados.

Qual é a rotina?

Tudo que escrevi acima se resume a uma coisa: não há rotina. Os participantes têm algumas habilidades na manga, mas não sabem como e quando usá-las. Um jornalista comparou isso a fazer um bolo. "Nós temos todos os ingredientes: farinha, ovos, leite, etc. Daí jogamos tudo num saco, misturamos e esperamos que o bolo saia." De fato, temos todos os ingredientes, mas não sabemos como é a receita.

Quais os planos agora? Nossas primeiras experiências com jornalismo de dados podem ajudar outros jornalistas ou programadores entrar no mesmo campo de trabalho, e estamos trabalhando para produzir um relatório.

Nós também estamos pensando em como continuar com a RegioHack num formato de hackathona. Nós achamos isto divertido, educacional, produtivo e uma ótima introdução ao jornalismo de dados.

Mas para o jornalismo de dados funcionar, temos de integrar isto com a redação. Além de discursos, coletivas de imprensa e encontros com autoridades, jornalistas têm de começar a pensar nos dados. Através da execução da RegioHack, provamos para nossa audiência que o jornalismo de dados não é somente hype. Nós podemos escrever matérias mais embasadas e diferenciadas dando a oportunidade aos leitores de consumirem material impresso e online.

— Jerry Vermanen, NU.nl

Seguindo o Dinheiro: Jornalismo de dados e Colaboração além das Fronteiras

Jornalistas investigativos e cidadãos interessados em desvelar corrupção e crime organizado que afetam as vidas de bilhões têm conquistado, a cada dia que passa, um acesso sem precedentes a informação. Imensos volumes de dados de governos e outras organizações estão disponíveis online, e parece que esse tipo tão necessário de informação está mais ao alcance de todos. Ao mesmo tempo, oficiais corruptos nos governos e grupos de crime organizado se empenham em ocultar os seus malfeitos. Há um esforço para manter as pessoas sem informação enquanto conduzem negócios espúrios que causam problemas em todos os níveis da sociedade, levando a conflitos, fome e outras crises.

É dever do jornalista investigativo expor os malfeitos e, fazendo isto, desmantelar mecanismos corruptos e criminosos.

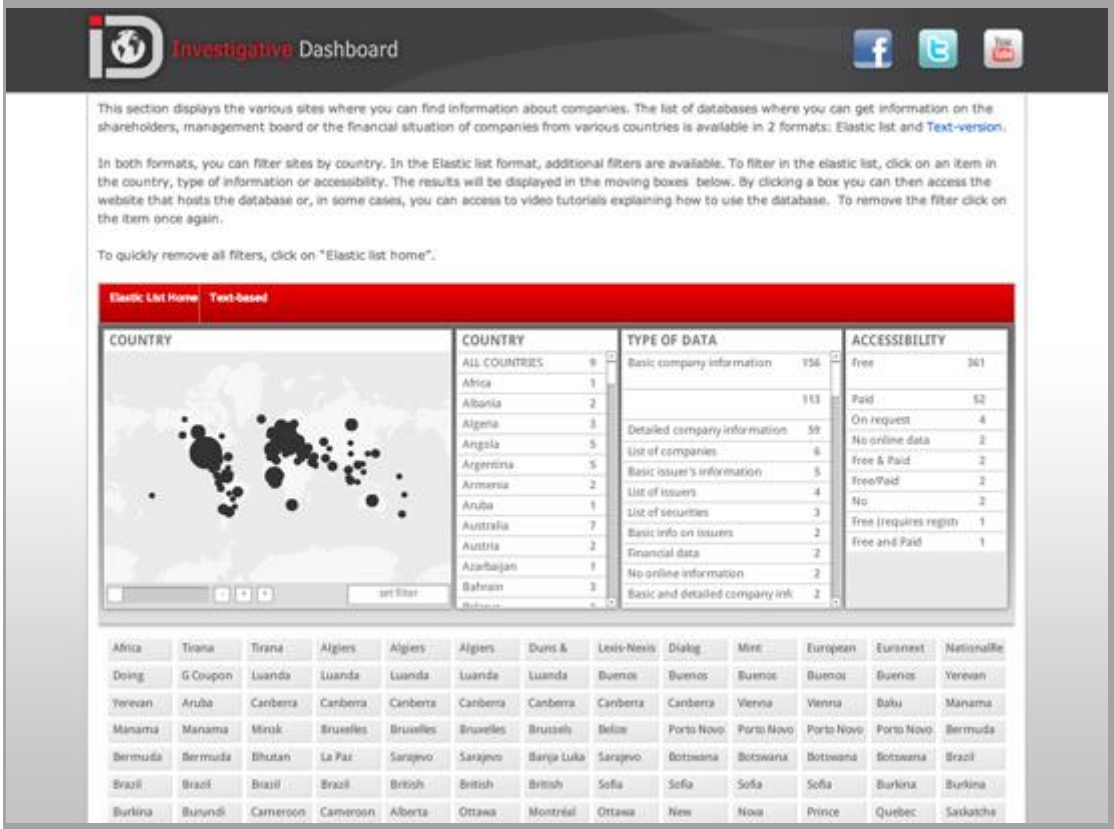


Imagem 12. O Painel Investigativo (OCCRP)

Há três diretrizes que, se seguidas, podem levar a um jornalismo bom e profundo ao investigar grandes atos de corrupção e crime, mesmo nos ambientes de informação mais escassa.

Pense fora do seu país

Em vários casos, é muito mais fácil obter informação fora do país onde o jornalismo investigativo opera. Informação obtida via bancos de dados estrangeiros ou por meio de leis de acesso à informação de outros países pode ser exatamente o que você precisa para fechar o quebra-cabeças de uma apuração investigativa. Criminosos e oficiais corruptos não mantêm o dinheiro no mesmo lugar onde ele foi roubado. Eles preferem depositar em bancos estrangeiros ou investir em outros países. Crime é global. Bases de dados que ajudam o jornalista investigativo a rastrear o dinheiro podem ser encontradas em vários lugares na internet. Por exemplo, o Investigative Dashboard (da imagem acima) permite a jornalistas seguir o dinheiro através das fronteiras.

Faça uso das redes existentes de jornalismo investigativo

Jornalistas investigativos de todo mundo se juntam em organizações como The Organized Crime and Corruption Reporting Project, The African Forum for Investigative Reporting, The Arab Reporters for Investigative Journalism, and The Global investigative Journalism Network. Jornalistas podem também fazer uso de plataformas de jornalismo profissional como a IJNet, onde informação relacionada ao jornalismo global é trocada diariamente. Muitos dos repórteres dessas redes trabalham em problemas similares e encontram situações parecidas, portanto faz muito sentido trocar informações e métodos. Listas de discussão por email e grupos de redes sociais são agregadas a esses fóruns, então é bastante fácil entrar em contato com seus companheiros jornalistas e pedir informações ou aconselhamento. Ideias para reportagens investigativas podem também partir daí.

Use a tecnologia e colabore com hackers

Softwares ajudam os jornalistas investigativos a acessar e processar informação. Eles são úteis para depurar, fuçar, coletar e entender um grande volume de dados, e também para achar os documentos certos para a matéria sair. Há muitos programas já prontos que podem ser usados como ferramenta para analisar, colher, ou interpretar informação — e, mais importante, jornalistas investigativos precisam saber que há muitos programadores prontos para ajudar se requisitados. Estes sabem como obter e manejar a informação, e podem ajudar muito na investigação. Os programadores, alguns membros de movimentos de

opendata globais, podem se tornar inestimáveis aliados na luta contra o crime e a corrupção, ajudando jornalistas a colher e analisar informações.

Um bom exemplo de interface entre programadores e cidadãos é o ScraperWiki, um lugar onde jornalistas podem pedir ajuda com extração de dados de sites. O Investigative Dashboard mantém uma lista de ferramentas prontas para recolher, modelar, e analisar dados.

A utilidade das diretrizes que mencionei tem sido visível em vários casos. Um bom exemplo é o trabalho de Khadija Ismayilova, uma experiente jornalista investigativa do Azerbaijão que trabalha num ambiente bem austero, em se tratando de acesso à informação. Ismayilova tem de sobrepujar obstáculos diariamente para oferecer ao público azeri informação boa e confiável. Em Junho de 2011,a repórter da Radio Free Europe/Radio Liberty's (RFE/RL) em Baku (capital do Azerbaijão) mostrou que as filhas do presidente do país, Ilham Aliyev, secretamente comandavam uma empresa de telefonia em rápido crescimento, a Azerfon por meio de firmas offshore com sede no Panamá. A Azerfon tem aproximadamente 1,7 milhão de assinantes, cobre 80% do território do país, e era (naquela época) o único provedor de serviços 3G no Azerbaijão. Ismayilova gastou três anos tentando descobrir quem eram os donos da companhia, mas o governo se negava a abrir informações dos acionistas e mentiu diversas vezes sobre a propriedade da empresa. As autoridades chegaram a anunciar que a companhia era de propriedade da empresa alemã Siemens AG, o que foi depois negado pelos alemães. Depois de muito investigar, a repórter conseguiu descobrir que a Azerfon pertencia às empresas privadas com sede no Panamá, mas isso parecia ser o fim da linha para a reportagem. Até que ela contou com ajuda de fora. No começo de 2011, Ismayilova descobriu, através do Investigative Dashboard, que companhias com sede no Panamá podem ser rastreadas através de um aplicativodesenvolvido pelo programador e ativista Dan O'Huiginn. Com a ferramenta, ela revelou o envolvimento das duas filhas do presidente.

Na verdade, O'Huiginn criou uma ferramenta que ajudou jornalistas de todo o o mundo relatar na corrupção — o Panamá, conhecido paraíso fiscal, tem sido usado por vários oficiais corruptos para esconder dinheiro roubado (dos comparsas do antigo presidente egípicio, Hosni Mubarak, a oficiais corruptos nos Balcãs ou na América Latina). O que o programador-ativista fez é chamado de web scraping: um método que permite a extrair e

reconstituir a informação para que possa ser usada na investigação. O'Huiginn teve de forçar a extração de informações do Registro das companhias do Panamá porque o site, mesmo aberto ao público, só permite buscas se o repórter souber o nome da companhia. Isto limita as possibilidades da investigação, já que os repórteres geralmente procuram pelo nome das pessoas para rastrear as suas propriedades. Com a extração de dados, ele criou um novo site onde buscas de nome também são possíveis. Com isso, o site permite a repórteres investigativos de muitos países buscar pelos nomes autoridades e checar se eles secretamente são proprietários de corporações no Panamá.

Há outras vantagens em usar as diretrizes que mencionei, além de obter melhor acesso à informação. Uma delas é minimizar o risco e garantir melhor proteção aos repórteres investigativos que trabalham em ambientes hostis. Quando numa rede, o jornalista trabalha com colegas em outros países, então é mais difícil para criminosos identificarem o responsável pela exposição dos seus crimes. Como resultado, fica muito mais difícil para governos e oficiais corruptos tentarem uma retaliação ao jornalista.

Outra dica para guardar é que uma informação que não parece muito valiosa num local pode ser crucial em outro. A troca de dados por redes de jornalistas investigativos pode levar a novas matérias importantes. Por exemplo, a informação que um romeno foi pego na Colômbia com 1 kg de cocaína não ganhará a primeira página de um jornal em Bogotá, mas pode ser muito importante para o público romeno se um repórter descobre que essa pessoa está trabalhando para o governo de Bucareste.

Reportagem investigativa eficiente é o resultado de cooperação entre jornalistas investigativos, programadores, e outros que querem usar os dados para contribuir com uma sociedade mais limpa, justa e global.

— Paul Radu, Organized Crime and Corruption Reporting Project

Nossas Histórias Vêm Como Código

O OpenDataCity foi fundado no final de 2010. Não havia quase nada que pudesse ser chamado de jornalismo de dados acontecendo na Alemanha na época.

Por que fizemos isso? Muitas vezes ouvimos pessoas que trabalham para jornais e TVs dizerem: "Não, nós não estamos prontos para começar uma área dedicada ao jornalismo de dados na nossa redação. Mas ficaríamos felizes em terceirizar isso para alguém."

Até onde sabemos, somos a única companhia exclusivamente especializada em jornalismo de dados na Alemanha. Atualmente, estamos em três: dois com uma formação jornalística e um com um profundo conhecimento de códigos e visualização. Trabalhamos também com hackers, designers e jornalistas free-lancers.

Nos últimos doze meses fizemos quatro projetos de jornalismo de dados com jornais, e oferecemos treinamento e consultoria para trabalhadores de mídia, cientistas, e escolas de jornalismo. O primeiro aplicativo que fizemos foi o TAZ, uma ferramenta interativa sobre barulhos de aeroporto sobre o recém-construído aeroporto em Berlim. Nosso próximo projeto notável foi uma aplicação sobre retenção de dados em cima da quantidade imensa de dados que uma companhia telefônica guardava de cada pessoa (um político alemão requisitou todos os dados dele na justiça e mostramos o que podia ser feito com esses dados). Neste projeto, com o jornal Zeit Online, ganhamos um Grimme Online Award e um Lead Award na Alemanha, e um Prêmio de Jornalismo Online da Associação de Jornalismo Online nos Estados Unidos. No momento em que escrevemos este texto, temos vários projetos na linha de produção, que vão de simples infográficos interativos até o desenho e desenvolvimento de um tipo de middleware (software usado para transportar informações entre programas de diferentes) de jornalismo de dados.

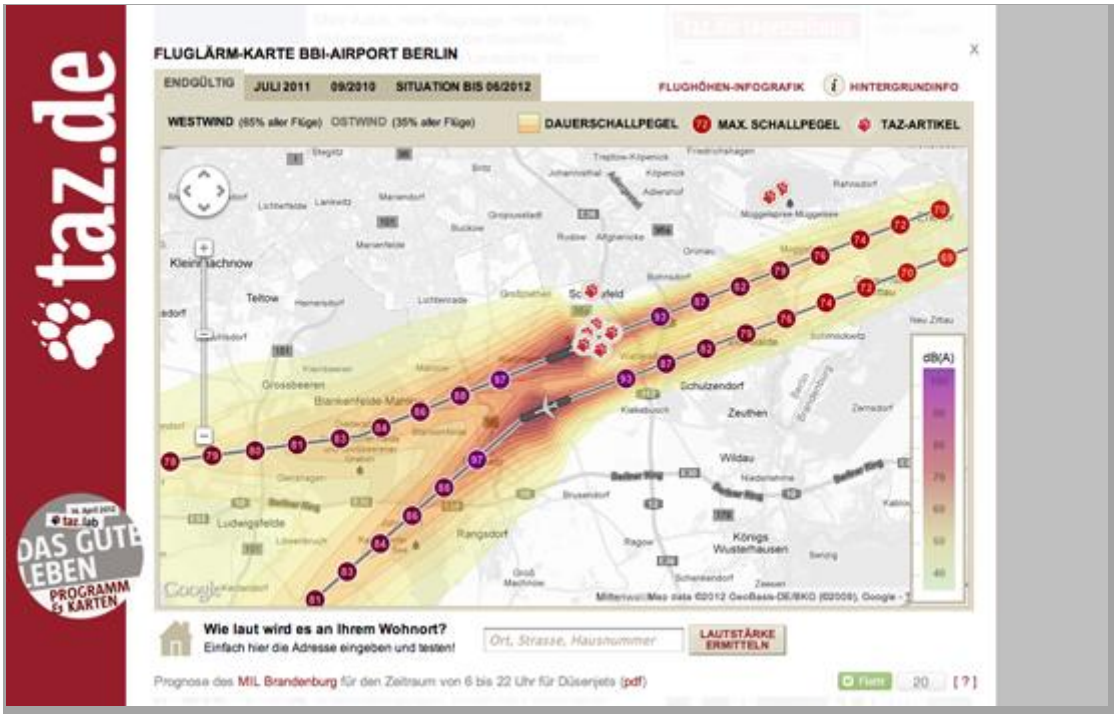


Imagem 13. Mapa do barulho do aeroporto (Taz.de)

Claro, ganhar prêmios ajuda a construir uma reputação. Mas quando conversamos com os publishers, que têm de aprovar nossos projetos, nosso argumento para investir em jornalismo de dados não é ganhar prêmios. É sobre receber atenção através de um longo período de tempo e de uma maneira sustentável. Ou seja, construir coisas devido ao seu impacto no longo prazo; o objetivo não é o furo, que é frequentemente esquecido depois de alguns dias.

Aqui estão três argumentos que usamos para encorajar os editores a empreenderem projetos de longo prazo:

Projetos de dados não envelhecem

Dependendo de seu design, novos materiais podem ser adicionados a aplicativos de jornalismo de dados. E, além de atender aos usuários, os projetos também podem ser usados internamente para reportagem e análise. Se você se preocupa que sua concorrência também se beneficie do seu investimento, é possível manter alguns atributos ou algum dado para uso interno apenas.

Você pode se beneficiar do trabalho já feito

Quando trabalhando num projeto de dados, você frequentemente irá criar pequenos códigos que podem ser reutilizados ou atualizados. O próximo projeto pode demorar metade do tempo porque você sabe muito

melhor o que fazer (e não fazer), e porque tem partes e pedaços que podem ser reaproveitados.

Jornalismo de dados se paga

Projetos de dados são mais baratos que tradicionais campanhas de marketing. O mercado de notícias online frequentemente investe em estratégias como o SEO ou o SEM (táticas que visam fazer um site aparecer melhor nas páginas de busca). Um projeto de dados normalmente irá gerar vários cliques e alvoroço, e pode se tornar viral. Editores normalmente irão pagar menos por isso do que tentando gerar a mesma atenção com cliques e vínculos através do SEM.

Nosso trabalho não é muito diferente de outras agências de novas mídias: oferecemos aplicações ou serviços para o mercado de notícias. Mas talvez nosso diferencial seja em nos pensar, primeiramente, como jornalistas. Ao nosso ver, os produtos que entregamos são reportagens, embora sejam fornecidas não em palavras, imagens, áudio ou vídeo, mas em códigos. Quando falamos de jornalismo de dados, temos de falar de tecnologia, software, aparelhos, e como contar uma história com eles.

Para exemplificar, recentemente terminamos uma aplicação que puxa, em tempo real, dados do site da ferrovia alemã. Isso permitiu desenvolver um monitor de trem interativo para o diário alemão Süddeutsche Zeitung, mostrando os atrasos de trens de longa distância em tempo real. Os dados da aplicação são atualizados a cada minuto, e nós estamos oferecendo um API para ele. Começamos no projeto há alguns meses, e desde então coletamos um imenso banco de dados, que cresce a cada hora. Neste momento, acumulamos centenas de milhares de linhas de dados. O projeto permite explorar esses dados em tempo real e pesquisar nos arquivos dos meses anteriores. No fim, a história que contamos será bastante definida pela ação individual dos usuários.

No jornalismo tradicional, devido à característica linear da mídia escrita ou de rádio e TV, temos de pensar sobre um começo, um fim, um desenvolvimento da história, o tamanho e o ângulo da que a obra adotará. Com o jornalismo de dados as coisas são diferentes. Sim, existe um começo. A pessoa vem ao site e tem uma primeira impressão da interface. Mas depois ela está por si. Talvez fique por um minuto, ou por meia hora.

Nosso trabalho como jornalistas de dados é oferecer a estrutura ou o ambiente para isto. Assim como a codificação e o tratamento de bits de dados, nós temos de pensar em maneiras inteligentes para criar experiências. A Experiência do Usuário (UX) vem principalmente da Interface (Gráfica) do Usuário (GUI). No final, essa é a parte que vai decolar ou afundar um projeto. Você pode ter o melhor código operando no fundo através do manejo de um excitante conjunto de dados. Mas se a interface ao usuário é ruim, ninguém vai se importar com ele.

Ainda há muito o que aprender e com o que experimentar. Mas por sorte existe a indústria de games, que tem inovado há muitas décadas com respeito a narrativas, ecossistemas e interfaces digitais. Quando desenvolvemos aplicações de jornalismo de dados devemos observar de perto como o design de games funciona e como as histórias são contadas nesses jogos. Por que jogos simples como Tetris são tão divertidos? E o que faz os mundos abertos de games como Grand Theft Auto ou Skyrim serem geniais?

Nós achamos que o jornalismo de dados veio para ficar. Em alguns anos, o fluxo de trabalho do jornalismo de dados vai ser naturalmente inserido em redações porque sites de notícias terão que mudar. A quantidade de dados que está disponível publicamente vai continuar crescendo. Mas, felizmente, novas tecnologias vão continuar a nos permitir encontrar novas maneiras de contar histórias. Algumas destas histórias serão guiadas por dados, e muitas aplicações e serviços terão uma característica jornalística. A questão interessante é qual estratégia as redações vão desenvolver para estimular este processo. Elas vão organizar grupos de jornalistas de dados integrados às suas redações? Existirão departamentos de pesquisa e desenvolvimento parecidos com startups dentro da empresa? Ou partes do trabalho serão terceirizadas para companhia especializadas? Nós ainda estamos no começo e apenas o tempo dirá.

— Lorenz Matzat, OpenDataCity

Kaas & Mulvad: Conteúdo pré-produzido para comunicação segmentada

A chamada stakeholder media (formada por serviços de reportagem segmentados, por empresas não jornalísticas e outros grupos) é um setor emergente, mas amplamente desprezado por teóricos da mídia. Ele possui um tremendo impacto potencial por meio de redes online ou para gerar conteúdo. O setor pode ser definido como meios de comunicação controlados por grupos organizacionais ou institucionais interessados em promover determinados interesses ou certas comunidades. ONGs frequentemente criam esse tipo de mídia, assim como grupos de consumidores, associações profissionais, sindicatos, entre outros. O ponto que limita a sua habilidade de influenciar a opinião pública ou outros grupos de interesse é o fato de que normalmente eles carecem da capacidade de descobrir informações importantes, até mais que grupos tradicionais de mídia que sofreram com cortes de funcionários. Kaas & Mulvad, uma empresa Dinamarquesa com fins lucrativos, é um dos primeiros empreendimentos investigativos de mídia a prover a capacidade de experts para esse segmento.

A empresa começou em 2007 como uma spinoff do Instituto Dinamarquês para Reportagem Com Auxílio de Computador (Dicar, na sigla em inglês), uma instituição sem fins lucrativos. Começou com a venda de reportagens investigativas para a mídia e treinava jornalistas em análise de dados. Seus fundadores, Tommy Kaas e Nils Mulvad, eram repórteres da grande mídia. A nova empresa ofereceu o que eles chamam de ”dados mais insights jornalísticos” (conteúdo semi-finalizado que requer edição), principalmente para meios de comunicação segmentados, que finalizam o conteúdo com releases ou matérias e distribuem na mídia tradicional ou pelos seus canais diretos (como sites das empresas). Clientes diretos incluem instituições governamentais, empresas de relações públicas, sindicatos e organizações não governamentais como a EU Transparency e WWF. O trabalho para as ONGs incluiu o monitoramento agrícola e de subsídios à pesca, além de atualizações periódicas sobre as atividades de lobistas da União Europeia gerados por meio da extração de dados (prática do "scraping") de sites na internet. Clientes indiretos incluem fundações que financiam projetos de ONGs. A empresa também trabalha com a grande mídia; um tabloide comprou um serviço de monitoramento de celebridades, por exemplo.

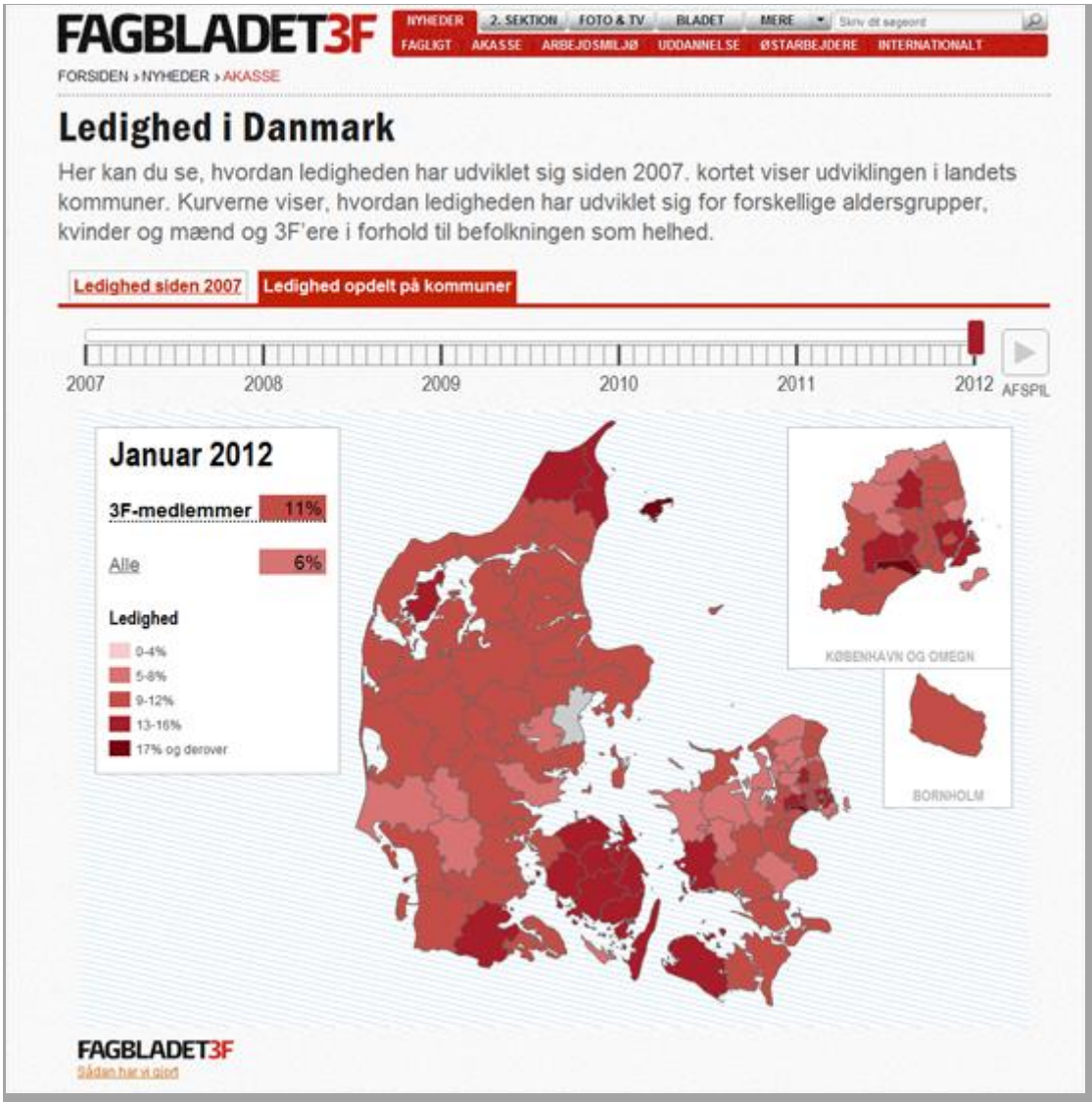


Imagem 14. Empresas de comunicação segmentada - Stakeholder Media (Fagblaget3F)

Os projetos de jornalismo de dados no portifólio deles incluem:

Mapa do Desemprego para o 3F

Uma visualização de dados com indicadores chave sobre o desemprego na Dinamarca realizado para o 3F, o maior sindicato da Dinamarca.

Condições de vida para o 3F

Outro projeto para o 3F que mostra a desigualdade de condições de renda e qualidade de vida em diferentes partes da Dinamarca. O mapa usa 24 indicadores diferentes.

Mapa dos municípios endividados para o jornal "Ugebrevet A4"

Um projeto que calcula um "índice de endividamento" dos municípios e mostra numa visualização de dados as diferenças na economia.

Instalações perigosas na Dinamarca

Projeto que mapeia e analisa instalações perigosas próximas a creches e a outras instituições infantis, realizado por "Born & Unge", revista publicada pela BUPL - Associação Dinamarquesa de Professores de Educação Infantil.

Dados sobre Responsabilidade Corporativa para a empresa Vestas

Visualização de dados sobre cinco áreas de responsabilidade corporativa para a empresa dinamarquesa de turbinas eólicas Vestas. O texto é gerado automaticamente. As informações são atualizadas automaticamente a cada três meses em 400 webpages, desde de dados de escala mundial até sobre cada uma das unidades de produção.

Mapa de nomes para a Experian

Escreva seu sobrenome e veja a distribuição de pessoas com o mesmo nome em diferentes áreas da Dinamarca.

Smiley Map para Ekstra Bladet

Diariamente a Kaas & Mulvad extraiu dados de todas as inspeções sanitárias que indicavam comida de má qualidade e mapearam as últimas delas para o tabloide dinamarquês Ekstra Bladet.

Kaas & Mulvad não são os primeiros jornalistas a trabalhar com mídia segmentada. O Greenpeace, por exemplo, frequentemente coloca jornalistas para trabalhar como colaboradores nos seus relatórios. Mas não sabemos de nenhuma outra empresa cujas ofertas de mídia segmentada é focada em jornalismo de dados; é muito mais comum jornalistas trabalharem em ONGs como repórteres, editores e redatores. O foco atual em Reportagem com Auxílio de Computadores (RAC) está na pesquisa e na descoberta (pense no WikiLeaks). Aqui, novamente, Kaas & Mulvad inovam, focando na análise de dados. Sua abordagem requer não apenas habilidades de programação, mas também de compreensão de que tipo de informação pode trazer uma história de impacto. Pode-se dizer com segurança que qualquer um que quiser imitar seu serviço provavelmente teria de adquirir esses dois conjuntos de habilidades por meio de parcerias, porque as pessoas raramente possuem ambos.

Processos: TI inovadora mais análises

A empresa conduz cerca de 100 projetos por ano, com duração que varia de algumas horas a alguns meses. Além disso, investe continuamente em projetos que ampliam sua capacidade e suas ofertas. O serviço de monitoramento de celebridade era um experimento desse tipo. Outro envolveu coleta de dados na internet (scraping) para notícias de execuções hipotecárias e criação de mapas delas. Os sócios dizem que o primeiro critério para iniciar projetos é o quanto eles gostam do trabalho e podem aprender com ele. A busca da empresa pelo mercado vem depois que um novo serviço está definido. Eles deixam claro que, dentro da grande mídia, encontraram dificuldade para desenvolver novos métodos e novos negócios.

Mulvad comenta que:

Não temos editores ou chefes para decidir quais projetos podemos fazer, qual software ou hardware devemos comprar. Podemos comprar nossas ferramentas de acordo com o que o projeto precisa, bem como as melhores soluções para a coleta de informações da internet e mineração de dados. Nosso objetivo é ser vanguarda nestas áreas. Tentamos obter clientes que estão dispostos a pagar, ou se o projeto é divertido, fazemos isso por um custo menor.

Valor criado: Marcas Pessoais, Coorporativas e Receitas

O volume de negócios em 2009 foi de cerca de 2,5 milhões de coroas dinamarquesas, ou 336 mil euros. A empresa também sustenta a reputação dos sócios como jornalistas de ponta, o que mantém uma demanda para serviços de palestras e aulas. Suas aparições públicas, por sua vez, apoiam a marca da empresa.

Insights principais deste exemplo

- A crise de redução de capacidade da grande mídia é também uma crise de sub-utilização de capacidades. Kaas e Mulvad tiveram de deixar a grande mídia para fazer o trabalho que eles valorizam, e isso dá dinheiro. Nada impediu uma empresa de notícias de absorver esse valor.

- Ao menos em alguns mercados, existe uma possibilidade de lucro para "conteúdo semi-acabado", que pode servir aos interesses de grupos de mídia segmentada.
- No entanto, esta oportunidade levanta a questão de quanto controle os jornalistas podem exercer sobre a apresentação e o uso do seu trabalho por terceiros. Lembramos que essa questão já existe dentro da grande mídia (onde editores podem impor mudanças no trabalho de um jornalista), e isso tem existido dentro das demais indústrias midiáticas (como no cinema, aonde conflitos entre diretores e estúdios sobre os "cortes finais" não são raros). Não é particularmente um perigo moral da mídia segmentada, mas não irá desaparecer, também. Mais atenção é necessária para a ética dessa realidade e mercado crescente
- Do ponto de vista das receitas, um único produto ou serviço não é suficiente. Empreitadas de jornalismo investigativo bem-sucedidas precisam adotar uma abordagem de portifólio, na qual consultoria, ensino, palestra e outros serviços podem trazer receitas adicionais para apoiar a marca.

— Trecho extraído e editado de Mark Lee Hunter and Luk N. Van Wassenhove,"Disruptive News Technologies: Stakeholder Media and the Future of Watchdog Journalism Business Models". INSEAD Working Paper, 2010

Modelos de Negócio para o Jornalismo de Dados

Dentre todos os interesses e esperanças no que diz respeito ao jornalismo de dados, existe uma questão sobre a qual as redações sempre se mantêm curiosas: quais são os modelos de negócio?

Devemos ter cuidado ao fazer previsões, mas um olhar para a história recente e a situação atual da indústria da mídia pode esclarecer algumas questões. Atualmente, há muitas organizações jornalísticas que se beneficiaram ao adotar novas abordagens.

Termos como "jornalismo de dados" e o mais novo chavão, "ciência dos dados", podem soar como se descrevessem algo novo, mas isso não é bem verdade. Ao contrário, esses novos rótulos são apenas formas de caracterizar uma mudança que vem ganhando força ao longo de décadas.

Muitos jornalistas parecem não ter conhecimento do tamanho da receita que já é gerada através da coleta, análise e visualização de dados. Trata-se de um negócio de refinamento de informação. Com ferramentas de dados e tecnologias, é possível cada vez mais lançar luz sobre questões altamente complexas, sejam elas finanças internacionais, dívida, demografia, educação e assim por diante. O termo "business intelligence" descreve uma variedade de conceitos de TI que têm por objetivo proporcionar uma visão clara sobre o que está acontecendo nas empresas comerciais. As grandes e rentáveis empresas do nosso tempo, incluindo McDonalds, Zara e H&M, apostam em um rastreamento constante de dados para se tornarem lucrativas. E isso funciona muito bem para elas.

O que está mudando agora é que as ferramentas desenvolvidas para essa área agora estão se tornando disponíveis para outros domínios, incluindo a mídia. E há jornalistas que as entendem. Citemos, como exemplo, Tableau, uma empresa que fornece um conjunto de ferramentas de visualização. Ou o movimento "Big Data", no qual empresas de tecnologia usam pacotes de software (muitas vezes de código aberto) para trabalhar intensamente através de pilhas de dados, extraindo insights em milésimos de segundo.

Estas tecnologias podem ser aplicadas ao jornalismo. Equipes do The Guardian e The New York Times estão constantemente forçando os limites neste campo emergente. E o que estamos vendo atualmente é apenas a ponta do iceberg.

Mas como isso gera dinheiro para o jornalismo? O grande mercado que está se abrindo em todo o planeta tem a ver com a transformação de dados disponíveis publicamente em algo que podemos processar: tornar os dados visíveis e humanos. Queremos ser capazes de nos relacionar com os grandes números que ouvimos todos os dias no noticiário — o que os milhões e bilhões significam para cada um de nós.

Há algumas empresas baseadas em mídia de dados muito rentáveis, que simplesmente aplicaram este princípio antes que outras. Elas gozam taxas de crescimento saudáveis e lucros às vezes impressionantes. Um exemplo é a Bloomberg. A empresa opera cerca de 300 mil terminais e fornece dados financeiros aos seus usuários. Se você está no negócio financeiro, esta é uma ferramenta poderosa. Cada terminal vem com um teclado com código de cores e até 30.000 ações para pesquisar, comparar, analisar e ajudar você a decidir o que fazer em seguida. Esse negócio gera cerca de US$ 6,3 bilhões (EUA) por ano — pelo menos é o que foi estimado em uma matéria de 2008 no The New York Times. Como resultado, Bloomberg tem contratado jornalistas de direita, esquerda e centro. Eles compraram a venerável mas deficitária “Business Week,” e assim por diante.

Outro exemplo é o conglomerado de mídia canadense conhecido atualmente como Thomson Reuters. Eles começaram com um jornal, compraram alguns títulos bem conhecidos no Reino Unido, e então decidiram há duas décadas sair do negócio de jornais. Em vez disso, eles cresceram com base em serviços de informação, com o objetivo de fornecer uma perspectiva mais profunda a clientes de uma série de áreas. Se você se preocupa em ganhar dinheiro com informação especializada, meu conselho seria ler sobre a história da empresa na Wikipédia.

E observem a Economist. A revista tem construído uma marca excelente, influente em seu aspecto de mídia. Ao mesmo tempo, a "Economist Intelligence Unit" agora é mais uma empresa de consultoria, elaboração de relatórios sobre tendências relevantes e previsões para quase todos os países do mundo. Eles estão empregando centenas de jornalistas e alegam servir cerca de 1,5 milhão de clientes em todo o mundo.

E existem muitos nichos de serviços de dados que podem servir como inspiração: eMarketer nos EUA, que fornece comparações, gráficos e conselhos para qualquer pessoa interessada em marketing na internet; Stiftung Warentest,

na Alemanha, uma instituição que verifica a qualidade de produtos e serviços; Statista, também da Alemanha, uma startup que ajuda a visualizar informações publicamente disponíveis.

Em todo o mundo, existe uma onda de empresas iniciantes no setor, cobrindo uma vasta gama de áreas, por exemplo, a Timetric, que tem por objetivo "reinventar a pesquisa em negócios", OpenCorporates, Kasabi, Infochimps, e Data Market. Muitas delas são, indiscutivelmente, experimentos, mas juntas, podem ser consideradas um importante sinal de mudança.

Depois, existem os meios de comunicação públicos, o que em termos de jornalismo de dados, são um gigante adormecido. Na Alemanha, 7,2 bilhões de euros estão migrando para este setor, anualmente. O jornalismo é um produto especial: se bem feito, não se trata apenas de gerar lucros, mas de prestar um papel importante para a sociedade. Uma vez que esteja claro que o jornalismo de dados pode fornecer percepções melhores e mais confiáveis, com maior facilidade, uma parte deste dinheiro poderia ser usado para novos postos de trabalho nas redações.

Com o jornalismo de dados, não se trata apenas de ser o precursor, mas de ser uma fonte confiável de informação. Neste mundo repleto de canais, a atenção pode ser gerada em abundância, mas *confiança* é um recurso cada vez mais escasso. Os jornalistas de dados podem ajudar a reunir, sintetizar e apresentar fontes de informação diversas e muitas vezes difíceis, de modo a fornecer percepções reais sobre questões complexas para a audiência. Ao invés de apenas reciclar press releases e reescrever matérias vistas ou ouvidas anteriormente em outros lugares, os jornalistas de dados podem fornecer aos leitores uma perspectiva clara, compreensível e, de preferência personalizável, com gráficos interativos e de acesso direto a fontes primárias. Nada muito trivial, mas certamente valioso.

Então, qual é a melhor abordagem para que os aspirantes a jornalistas de dados possam explorar este campo e convencer a chefia a apoiar projetos inovadores?

O primeiro passo deve ser procurar oportunidades imediatas perto de casa: frutos mais fáceis de colher. Por exemplo, você pode já ter coleções de textos estruturados e dados que poderia usar. Um bom exemplo disso é o "Banco de Dados de Homicídios" do Los Angeles Times. Aqui, os dados e as visualizações são a parte central, não algo pensado depois. Os editores coletam informações sobre todos os crimes que encontram e só então escrevem artigos com base

neles. Com o tempo, tais coleções se tornam melhores, mais profundas e mais valiosas.

Isto pode não funcionar da primeira vez. Mas funcionará ao longo do tempo. Um indicador muito esperançoso é que o Texas Tribune e a ProPublica, duas empresas que podemos considerar de mídia pós-impressa, informaram que o financiamento para as suas organizações de jornalismo sem fins lucrativos ultrapassou suas metas muito mais cedo do que o planejado.

Tornar-se proficiente em os dados sobre tudo — seja como um generalista ou como um especialista focado em um aspecto da cadeia de dados — fornece uma perspectiva valiosa para as pessoas que acreditam no jornalismo. Como um editor bem conhecido na Alemanha disse recentemente em uma entrevista,"existe este novo grupo que se diz jornalistas de dados. E não estão mais dispostos a trabalhar por mixaria."

— Mirko Lorenz, Deutsche Welle

Estudos de Caso

Nesta seção, nós mostramos com mais profundidade os bastidores de muitos projetos de jornalismo de dados — de aplicativos desenvolvidos em um dia a investigações de nove meses. Nós aprendemos sobre como fontes de dados podem ser usadas para aumentar e melhorar a cobertura de tudo, de eleições a gastos públicos, de protestos à corrupção, do desempenho de escolas ao preço da água. Veremos grandes organizações de mídia, como BBC, Chicago Tribune, Guardian, Financial Times, Helsingin Sanomat, La Nación, Wall Street Journal e o Zeit Online, mas também iniciativas menores, como California Watch,Hack/HackersBuenos Aires, ProPublica e um grupo de jornalismo cidadão brasileiro chamado de Amigos de Januária.

O que há neste capítulo?

- Eleições parlamentares finlandesas e financiamento de campanha
- Hack Eleitoral em tempo real (Hacks/Hackers Buenos Aires)
- Dados no Noticiário: WikiLeaks
- Hackatona Mapa76
- A cobertura dos protestos violentos no Reino Unido pelo The Guardian
- Boletins escolares de Illinois (EUA)
- Faturas de hospitais
- Care Home Crisis: A crise da empresas de saúde em domicílio
- O telefone conta tudo
- Quais modelos se saem pior na inspeção veicular britânica?
- Subsídios de ônibus na Argentina
- Jornalistas de dados cidadãos
- O Grande Quadro com o Resultado das Eleições
- Apurando o preço da água via crowdsourcing

Basômetro: Passando o poder da narrativa para o usuário

O Basômetro foi a primeira ferramenta criada pelo Estadão Dados, que, por sua vez, foi o primeiro núcleo de jornalismo de dados das redações brasileiras. Não é um infográfico, não é um banco de dados, não é uma tabela, mas é tudo isso ao mesmo tempo.

O propósito do Basômetro é medir, partidária e individualmente, o grau de apoio de deputados e senadores ao governo federal no Congresso Nacional a qualquer tempo. Por que o governo como base de medida? Porque é assim que funcionam a Câmara e o Senado no Brasil: aprovando ou rejeitando proposições do Executivo.

O outro motivo é que o apoio ao governo é condicional, melífluo, temporário, pendular. Nenhum voto é líquido e certo. A "base governista" é um vício de linguagem. Apoio vira oposição de uma votação para outra, e vice-versa. Tudo depende de negociação, de agrados, de liberações de verbas, de concessões de cargos.

Daí a importância de medir essas oscilações e saber quem e quantos estão onde no espectro político a cada instante. O Basômetro é o termômetro do presidencialismo de coalizão que governa o Brasil desde a redemocratização. É uma interface simples para analisar um sistema complexo e volumoso.

Estão computados no Basômetro centenas de milhares de votos nominais (nos quais os parlamentares se identificam) dados na apreciação de matérias em que o governo tenha encaminhado a votação contra ou a favor - sejam projetos de lei, emendas constitucionais, medidas provisórias, destaques de votação, ou simples requerimentos de inversão de pauta.

Não são consideradas no Basômetro votações simbólicas, secretas ou nominais nas quais o líder do governo não tenha orientado sua bancada para votar a favor ou contra - seja porque é impossível saber como votou cada parlamentar, seja porque o governo não tinha um interesse manifesto no resultado.

O Basômetro registra se o deputado ou senador votou a favor, contra, se absteve ou não votou. A sua taxa de governismo é obtida pela divisão do número de votos a favor do governo pelo total de votos dados pelo parlamentar (incluídas as abstenções).

Assim, um deputado que tenha votado 62 vezes junto com o líder do governo, que por 9 vezes tenha votado contra e que por 2 vezes tenha se abstido de votar

terá uma taxa de governismo de 85%. Não importa que o deputado seja, como é, do DEM (partido de oposição) nem que ele tenha faltado a 112 votações. Conta apenas e exclusivamente o que ele fez transparentemente em plenário.

A taxa de governismo das bancadas partidárias é obtida pela média das taxas de todos os parlamentares do partido que tenham participado de alguma votação computada pelo Basômetro. Todos entram nesse cálculo, inclusive os que por uma razão ou outra não exerçam mais o mandato. Busca-se assim medir o comportamento histórico da bancada, não apenas o instantâneo.

O Basômetro foi concebido para permitir ao usuário fazer recortes temporais, partidários ou geográficos simplesmente deslizando seus marcadores ou acionando seus filtros.

É possível comparar, por exemplo, o tamanho da base governista na Câmara dos Deputados durante o primeiro ano do governo Dilma Rousseff (79% de apoio médio) com o da primeira metade do terceiro ano (71%). Ou, mais especificamente, o grau de fidelidade do PMDB: nesse período, caiu de 93% para 73%.

Além dos "sliders" de tempo localizados no eixo horizontal da interface, outro "slider" colocado no eixo vertical permite contar, automaticamente, quantos deputados ou senadores votaram com que frequência junto com o governo.

Se o usuário pesquisar o tamanho do "núcleo duro" da base de Dilma na Câmara, por exemplo, descobrirá que ele foi reduzido a um terço entre 2011 e o primeiro semestre de 2013. No começo do governo, nada menos do que 306 deputados votavam pelo menos 90% das vezes junto com o governo. Entre janeiro e junho de 2013 esse número havia sido reduzido a 103 (e 79 deles são do PT).

Imagem 1. Basômetro mostra a redução do "núcleo duro" do governo

Trata-se de uma medida objetiva do grau de apoio que o governo de ocasião dispõe a cada momento no Congresso. Ao contrário da cobertura jornalística tradicional, passa longe do discurso político, dos bastidores, das negociações, do mise-en-scène. É mais preciso, é mais conciso, é menos dependente das fontes de informação humanas e, por consequência, menos manipulável.

É também uma revolução na narrativa jornalística. Em vez de o jornalista contar para o leitor/espectador/ouvinte o que aconteceu, o Basômetro transfere ao usuário o poder de narrar a história para si próprio. O jornalista perdeu a exclusividade de descrever o que se passou. Qualquer um pode fazer isso - sem intermediários, preferências ou preconceitos que não os seus.

Como em toda boa ferramenta, o uso do Basômetro é permanente - ao menos enquanto houver Congresso Nacional e/ou meios de o Estadão Dados alimentá-lo. A base cresce a cada votação no Senado e na Câmara. Pode incorporar votações de governos passados (inclui os dos governos de Luiz Inácio Lula da Silva), futuros e novas dimensões. Isso provoca problemas, porém.

O código do Basômetro transfere a maior parte das operações para o navegador do usuário. Isso torna as transições e cálculos mais rápidos, mas aumenta o tempo de espera para o carregamento das bases de votações quando o usuário acessa a ferramenta pela primeira vez. A cada novo governo, maior o tamanho dessa base a ser transferida, o que acaba sendo uma limitação.

Em outra inovação nas redações jornalísticas, o código do Basômetro está disponível no Github com licença livre. Qualquer um pode fazer o download e construir um basômetro para a Assembleia Legislativa de seu Estado ou para a Câmara Municipal de sua cidade. Sem pagar nada pelos direitos autorais. Basta citar a fonte.

O Basômetro só existe porque é um trabalho coletivo. Ele reúne habilidades de profissionais com distintas formações: jornalistas, engenheiros/desenvolvedores e designers. Também não teria sido possível se vários níveis de chefias no Estadão não tivessem comprado a ideia do projeto e destinado os recursos humanos e materiais necessários à sua realização.

Para além do seu uso cotidiano na redação pelos jornalistas que acompanham política, a aceitação do Basômetro surpreendeu seus criadores. Quem temíamos que abominasse a novidade - a academia - adorou. E quem imaginávamos que usaria a ferramenta com estrondo - os políticos - se calou.

Logo após seu lançamento, a ferramenta inspirou uma série de artigos escritos por professores universitários e pesquisadores, todos eles publicados no portal estadao.com.br. A editora da Fundação Getúlio Vargas se propôs a editar um livro com versões ampliadas e atualizadas desses artigos. Uma ferramenta eminentemente digital, o Basômetro acabou no papel.

— José Roberto de Toledo, coordenador do Estadão Dados

InfoAmazônia: o diálogo entre jornalismo e dados geográficos

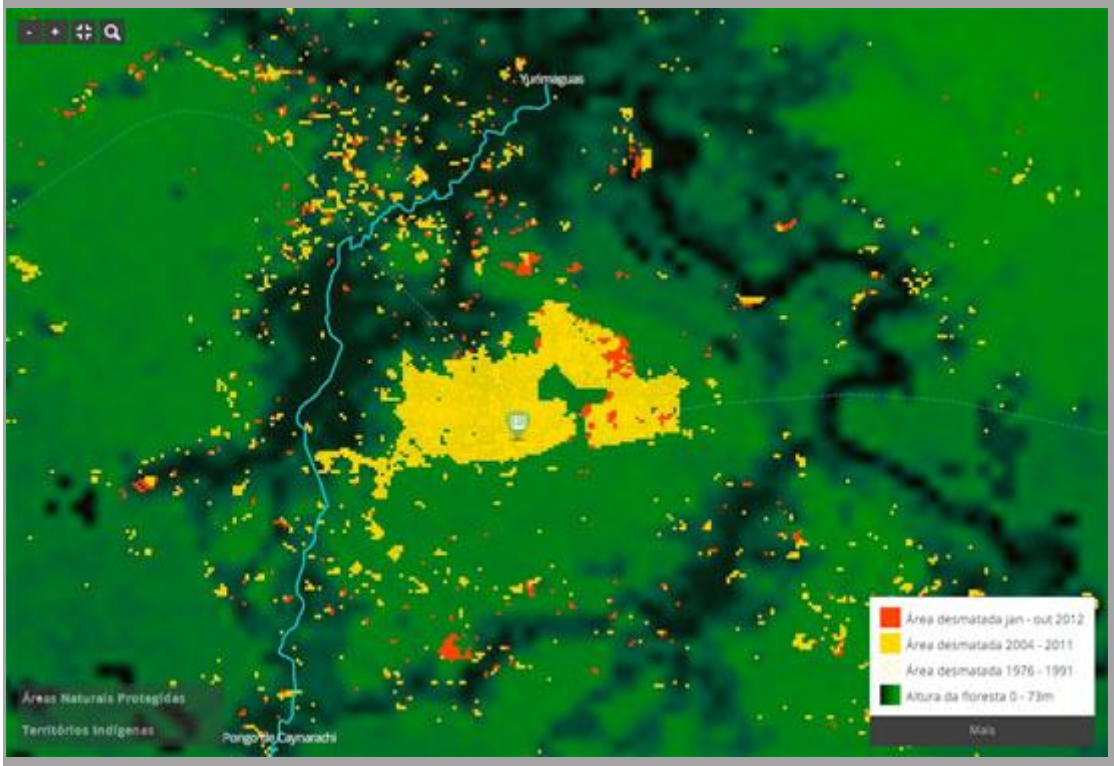


Imagem 2. Mapa do desmatamento mostrando o desmatamento na selva do Peru causado pelo avanço da indústria de óleo de palma (InfoAmazônia)

Em 2008, a necessidade de reportar sobre a alta incidência de incêndios florestais dentro de parques e reservas no Brasil me colocou em contato com as novas tecnologias de mapeamento digital. Naquela ocasião, a simples descoberta de que era possível incluir dados próprios em ferramentas do Google descortinou para mim uma revolução na forma de fazer a cobertura de meio ambiente.

Meu primeiro instinto foi exatamente localizar as reportagens sobre as queimadas em uma mapa interativo. Nos meses que seguiram, fui descobrindo, fascinado, outros instrumentos que permitiam visualizações ainda mais poderosas e que facilitavam a navegação do público pelos dados. Como bem observou a minha esposa, tornei-me amante do Google Earth.

O especial Monitor - Fogo nas Áreas Protegidas, lançado pelo site O Eco foi a experiência precursora do que, 4 anos depois, viria a ser o InfoAmazônia: uma plataforma digital que reúne jornalismo e dados ambientais em uma espécie de diálogo onde o fio condutor é a referência geográfica. Dados emprestavam contexto para as reportagens, mas o inverso também era verdadeiro: o

jornalismo qualificava a informação vinda do satélite. Nossa frase de efeito na época foi “o que satélite capta do espaço, o repórter conta em terra”

Uma exposição em 2010 na British Library me mostrou com enorme clareza que mapas são bons instrumentos de informação há séculos. Mas a utilização deles, sempre bastante restrita. Basta visitar a suntuosa sala de mapas do Museu do Vaticano para entender como a confecção das cartas servia aos que detinham poder. O surgimento da geoweb, como tem sido classificado o crescente uso de mapas digitais, democratizou o conhecimento geográfico e abriu uma nova porta para o jornalismo: transmitir conteúdo sobre os mapas, criando distintas camadas de informação. O mapa se tornou um meio de publicação, onde a teia de longitude e latitude pode ser vista da mesma forma como as antigas marcas da lauda no papel.

A inspiração para o InfoAmazônia foi reforçada pela enorme quantidade de dados gratuitos; séries históricas sobre fogo e desmatamento, por exemplo, são encontradas em formatos abertos nos sites da NASA ou do Instituto Brasileiro de Pesquisas Espaciais (INPE). Nossa ideia, logo de início, era usar o dado de satélite como contexto e guia para reportagens que deveriam ser feitas em campo pelos jornalistas. Assim surgiu um nome para a prática, o geojornalismo - uma espécie de galho dentro da frondosa árvore do jornalismo de dados.

Tenho enfatizado que o termo geojornalismo apareceu mais por conta de um desejo de propagandear o que estamos fazendo do que como um conceito bem formado. No entanto, após anos amadurecendo a plataforma, nos demos conta de que existem muitos fundamentos que surgiram exatamente do desejo de transformar o jornalismo em uma camada relevante para entender um determinado território, neste caso a maior floresta tropical do planeta.

A arquitetura do InfoAmazônia

O projeto InfoAmazônia foi lançado em junho de 2012 através de uma parceria entre O Eco e Internews, uma organização americana dedicada a fomentar a mídia em países em desenvolvimento, com apoio do Centro Internacional de Jornalistas (ICFJ), que financia o meu trabalho através das Bolsas Knight.

Imagem 3. A equipe de desenvolvedores do InfoAmazonia reunida em 17 de junho poucas horas antes do lançamento no Rio de Janeiro (foto: Gustavo Faleiros)

A primeira decisão, e certamente a mais difícil, foi a escolha da ferramenta de mapas. Desde o planejamento da plataforma, em 2008, a escolha era utilizar as ferramentas do Google. Mas notamos que, por conta da grande quantidade de informação coletada, necessitávamos de algo diferente, e acabamos nos unindo em uma parceria com a empresa americana MapBox.

A decisão por usar a tecnologia de MapBox foi guiada pelo fato de que os mapas funcionam como imagens interativas, suportando uma enorme quantidade de dados. Ao contrário de outras ferramentas, as camadas são renderizadas antes de irem para nuvem e um recurso conhecido como UTF Grid permite a interação entre os usuários e a base de dados com uma rapidez incrível. Isso nos permite ter hoje mapas como o do desmatamento, com até 15 camadas diferentes com séries históricas representando dados dos últimos 20 anos.

Para montar o InfoAmazonia contamos com 8 pessoas. Do MapBox - cuja equipe liderada pelo programador Alex Barth enriqueceu o projeto com novas ideias - havia o designer do site, um designer de mapas e um programador para o sistema de publicação (CMS). Do nosso lado, no Brasil, tínhamos uma gestora de desenvolvimento (Juliana Mori, que coordenava a execução das etapas do projeto) e dois jornalistas organizando a base de dados das reportagens. Eu e James Fahn (da Internews) cuidamos da parte institucional e concepção editorial.

Uma das questões fundamentais foi criar uma base de dados de reportagens sobre os temas que seriam representados nos mapas. Usando uma planilha de Google Docs, onde havia uma coluna de coordenadas geográficas, começamos a acumular notícias em português, inglês e espanhol sobre desmatamento, queimadas, conservação, mineração e outras questões relevantes. No lançamento, a tabela possuía 180 matérias. Um ano depois, cerca de 800 já tinham sido agregadas .

Modelo para distribuir e replicar

É exatamente a acumulação de dados que nos faz mover em novas direções. Acreditamos que o aplicativo InfoAmazonia tem algumas características que o tornam único. Este é o único local na web onde se pode encontrar concentradas informações sobre Amazônia como um todo, não apenas do Brasil, mas dos 9 países que detêm a floresta tropical. Esta vantagem também se torna um desafio na gestão dos dados.

Nossa primeira ação para lidar com o desafio foi criar um tema de Wordpress exclusivo para a gestão dos mapas e notícias por jornalistas. Para isso, trabalhamos com dois estúdios de São Paulo, Cardume e Memelab. Em maio de 2013, esse tema do Wordpress - batizado de Mappress - se tornou livre para utilização e seu código pode ser encontrado no GitHub. Potencialmente, outros projetos com informações do Cerrado, da Caatinga ou da Mata Atlântica poderão surgir, testando a validade do olhar territorial na cobertura jornalística.

Recentemente, criamos uma seção dedicada à customização dos mapas pelo público e por instituições parceiras. É possível levar toda essa informação que batalhamos para agregar simplesmente embedando - ou seja incorporando - o código em seu próprio site. Os mapas podem ser desagregados por camadas ou filtrados por tipo de notícias. Nossa esperança é uma só: aumentar o alcance e o impacto dos dados sobre a Amazônia.

— Gustavo Faleiros, InfoAmazônia

The Opportunity Gap: projeto de oportunidades em escolas

The Opportunity Gap usou dados de direitos civis do Departamento de Educação americano nunca antes liberados e mostrou que alguns estados dos EUA, como a Flórida, aumentaram o nível de educação e ofereceram aos estudantes ricos e pobres acesso praticamente igual a cursos de alto nível, enquanto outros, como Kansas, Maryland, e Oklahoma oferecem menos oportunidades em bairros com famílias mais pobres.

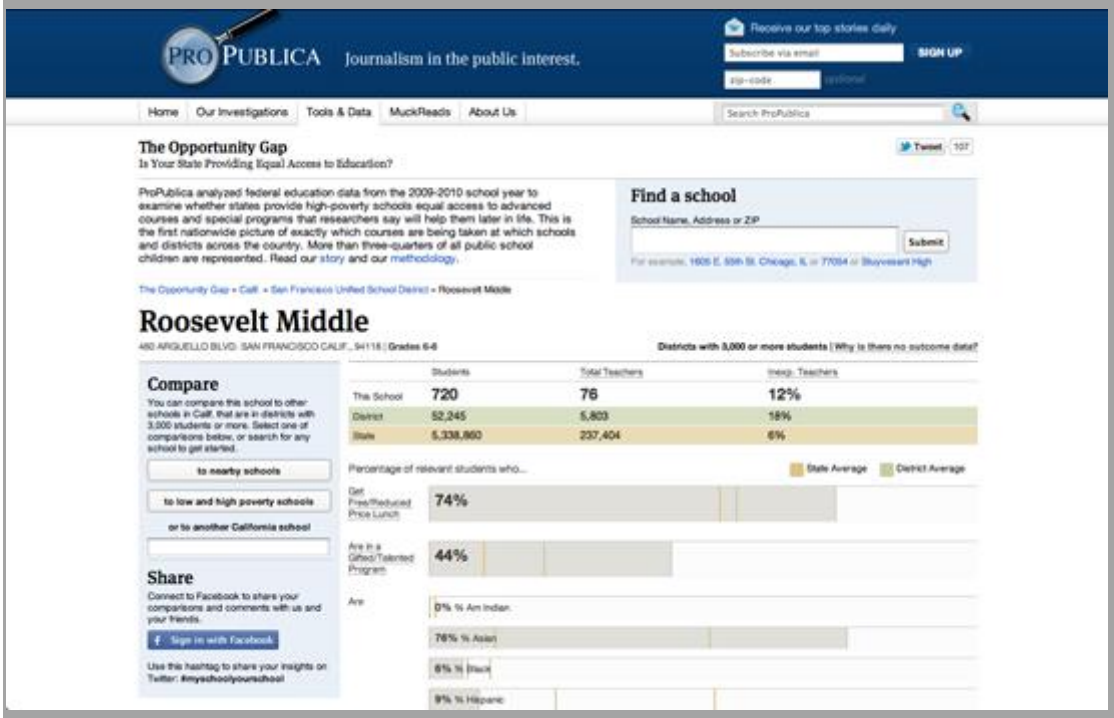


Imagem 4. O projeto The Opportunity Gap (ProPublica)

Os dados incluíram todas as escolas públicas em bairros com três mil alunos ou mais. Mais de três quartos de todos os estudantes de escolas públicas foram representados. Um repórter de nossa redação obteve os dados e nosso diretor de Reportagem com Auxílio do Computador (RAC) os limpou extensivamente.

Foi um projeto com aproximadamente três meses de duração. Ao todo, seis pessoas trabalharam juntas na matéria e no aplicativo de notícias: dois editores, um repórter, uma pessoa de RAC e dois desenvolvedores. A maioria de nós não estava trabalhando exclusivamente no projeto durante este período.

O projeto realmente exigiu a combinação de nossas habilidades: profundo conhecimento na área, entendimento das melhores práticas com dados, design e habilidades em programação, e por aí vai. Mais importante foi a habilidade de encontrar a história dentro dos dados. O projeto também exigiu edição, não só

para a matéria que resultaria dos dados, mas também para próprio aplicativo de notícias.

Para o tratamento e análise dos dados foram utilizados principalmente Excel e scripts de tratamento, bem como o Microsoft Access. O aplicativo de notícias foi escrito em Ruby on Rails e usa muito JavaScript.

Além de uma reportagem mais geral sobre o problema, nossa cobertura incluiu um aplicativo de notícias interativo permitindo encontrar exemplos na imensa base de dados. Usando nosso aplicativo, um leitor poderia identificar sua escola local — por exemploCentral High School in Newark, N.J. — e imediatamente ver a performance dela em áreas variadas. Apertando o botão Comparar com escolas de alto e baixo índice de pobreza, veria uma comparação outros colégios, sua pobreza relativa e seu nível de ensino de matemática, participação no "Advanced Placement" (programa criado nos Estados Unidos para oferecer matérias de nível universitário a alunos do Ensino Médio) e outros cursos importantes. A situação de pobreza dos estudantes é mostrada pelo percentual de alunos que podem ingressar num programa de almoço grátis do governo.

Em nosso exemplo, ao clicar no botão, Central High é comparada a Millburn Sr. High (menos pobre) e International High (mais pobre). O Opportunity Gap mostra que apenas 1% dos estudantes de Milburn podem ter almoço gratuito e 72% deles cursaram ao menos uma disciplina do Advanced Placement (AP). No outro extremo,a escola International High, 85% dos seus alunos são elegíveis ao almoço grátis, mas somente 1% deles cursou disciplinas do AP.

Por meio deste exemplo, o leitor pode usar algo que ele conheça - uma escola de ensino médio - para entender algo que não conheça: a distribuição do acesso à educação e o quanto a pobreza é um indicador desse acesso.

Nós também integramos o aplicativo ao Facebook, de maneira que ele informasse automaticamente os leitores sobre as escolas de seu interesse quando estes acessassem a rede social.

O tráfego para todos os nossos aplicativos de notícias é excelente, e estamos particularmente orgulhosos da maneira como este app conta uma história complexa — indo mais direto ao ponto, ele ajuda os leitores a contar suas próprias histórias para si mesmos.

Tal como em muitos projetos que começam a partir de dados governamentais, foi necessário limpar muito os dados. Por exemplo, enquanto existem apenas

cerca de 30 cursos no programa Advanced Placement, algumas escolas relataram centenas deles. Isso levou à verificação manual e ligações para as escolas para confirmação e correções.

Também trabalhamos arduamente para ter certeza de que o app contasse uma história que fosse "distante" e uma "próxima". Ou seja, o aplicativo precisava apresentar ao leitor uma visão geral e ampla nacional – especificamente, uma maneira de comparar o que faziam os estados no que diz respeito ao acesso à educação. Mas, uma vez que a uma visão geral por vezes deixa os leitores confusos sobre o que os dados significam para eles, nós também queríamos que eles fossem capazes de encontrar a sua própria escola local e compará-la com escolas mais ricas e pobres na sua área.

Se fosse aconselhar aspirantes a jornalistas de dados interessados em assumir esse tipo de projeto, diria que você tem que conhecer o assunto e ser curioso! Todas as regras aplicadas a outros tipos de jornalismo valem aqui. Você tem que conhecer os fatos direito, certificar-se de contar bem a história e, principalmente, verificar se o aplicativo de notícias não discorda da história que você está escrevendo – porque, se isso acontecer, um dos dois deve estar errado.

Além disso, se você quiser aprender a programar, a coisa mais importante é começar. Você pode gostar de aprender por meio de aulas, livros ou vídeos, mas certifique-se que você tem realmente uma boa ideia para um projeto e um prazo suficiente para completá-lo. Se há uma história em sua cabeça que só pode sair como um aplicativo de notícias, então a falta de conhecimento de programação não irá te parar!

— Scott Klein, ProPublica

Uma investigação de nove meses dos Fundos Estruturais Europeus

Em 2010, o Financial Times e o Bureau of Investigative Journalism (BIJ) somaram forças para investigar os Fundos Estruturais Europeus. O objetivo era identificar quem são os beneficiários desses fundos e se o dinheiro era bem aplicado. Com 347 bilhões de euros em sete anos, os Fundos Estruturais são o segundo maior programa de subsídios da União Europeia (UE). O programa existe há décadas, mas exceto por alguns panoramas gerais, havia pouca transparência sobre seus beneficiários. Como parte de uma série de mudanças na atual rodada de financiamento, as autoridades foram obrigadas a tornar públicas suas listas de beneficiários, incluindo a descrição dos projetos e o montante de recursos recebidos da UE e do fundos nacionais.

Imagem 5. Investigação dos Fundos Estruturais Europeus (Financial Times e Bureau of Investigative Journalism)

A equipe do projeto foi composta por 12 jornalistas e um programador em tempo integral que colaboraram por nove meses. Apenas a coleta de dados levou vários meses.

O projeto resultou em cinco dias de cobertura do Financial Times e no BIJ, um documentário de rádio da BBC, e diversos documentários para TV.

Antes de encarar um projeto com esse nível de esforço, você deve ter certeza que os achados serão originais, e que ao fim você terá boas histórias que ninguém mais tem.

O processo foi dividido em diferentes passos.

1. Identificar quem possui os dados e como estão armazenados

A Direção-Geral de Política Regional da Comissão Europeia (DG REGIO) mantém umportal para agregar as páginas de autoridades regionais que publicam dados. Acreditávamos que a Comissão tivesse uma base de dados abrangente com informações sobre seus projetos e que esta poderia ser acessada diretamente, ou que ao menos pudéssemos solicitar os dados por meio de pedidos pela lei de informação. Mas essa base não existia no nível de detalhamento que precisávamos. Rapidamente percebemos que muitos dos links que a Comissão fornecia estavam quebrados e que a maior parte das autoridades publicava dados em formato PDF, em vez de formatos como CSV ou XML, mais adequados para análises.

Um time de até 12 pessoas trabalhou para identificar os dados mais recentes e compilar os links em uma planilha que usamos colaborativamente. Uma vez que os campos não estavam uniformes (por exemplo, os cabeçalhos estavam em diferentes idiomas, algumas bases usavam moedas diferentes e algumas incluíam ainda separações por financiamento da UE ou fundos nacionais), precisávamos ter o máximo de precisão possível para traduzir e descrever os campos disponíveis em cada base de dados.

2. Download e tratamento dos dados

O próximo passo consistiu em fazer download de todas as planilhas, PDFs e, em alguns casos, arrancar os dados com scripts dos sites internet.

Cada base de dados precisava, então, ser padronizada. Nossa maior tarefa era extrair os dados dos PDFs, alguns com centenas de páginas. Muito desse trabalho foi feito por meio do UnPDF e do ABBYY FineReader, que permitem a extração de dados para formatos como CSV ou Excel.

Essa etapa também envolvia a checagem e rechecagem para verificar se as informações extraídas do PDF estavam corretas. Isso era feito por meio de filtragem, classificação e soma de totais (para assegurar que correspondiam ao que estava registrado nos PDFs).

3. Criar o banco de dados

O programador da equipe montou um banco de dados SQL. Cada um dos arquivos preparados foi então utilizado como um bloco de construção para a base global em SQL. A cada dia, um upload dos arquivos individuais era feito para essa base de dados SQL, que podia ser consultada em tempo real por meio de palavras-chave em uma interface amigável.

4. Rechecagem e análise

A equipe analisou os dados de duas formas principais:

Pela interface (front end) da base de dados

Isso envolvia entrar com palavras-chave de interesse (ex.: "tabaco", "hotel", "companhia A") no mecanismo de busca. Com ajuda do Tradutor do Google, que foi incluído como funcionalidade de busca em nossa base de dados, essas palavras-chave foram traduzidas para 21 idiomas e retornavam resultados mais adequados. Estes podiam ser baixados e os repórteres podiam aprofundar a pesquisa nos projetos individuais de seu interesse.

Por meio de macroanálises usando toda a base de dados

Ocasionalmente, era possível baixar toda a base de dados, que poderia então ser analisada (por exemplo, usando palavras-chave ou agregando dados por país, região, tipo de gasto, número de projetos por beneficiário etc.)

Nossas pautas surgiam a partir desses dois métodos, mas também por meio de investigação em campo e pesquisas secundárias.

A rechecagem da integridade das informações (agregando e confrontando com aquilo que as autoridades disseram estar sendo alocado) levou um tempo considerável. Um dos principais problemas era que as autoridades em sua maioria divulgavam somente o montante de "financiamento da UE e nacional". De acordo com as regras da UE, cada programa pode financiar determinados percentuais do total de dinheiro para os subsídios. O financiamento da UE é estabelecido, no nível do programa, pela chamada taxa de co-financiamento. Cada programa (por exemplo, competitividade regional) é composto de numerosos projetos. Um projeto pode, tecnicamente, receber 100% de financiamento da UE e outro, nada; contanto que estejam agrupados, o montante de financiamento do programa não pode ser maior que a taxa de co-financiamento aprovada.

Isso significava que precisávamos checar cada montante de financiamento que citávamos em nossas reportagens com a empresa beneficiária em questão.

— Cynthia O'Murchu, Financial Times

A crise da Zona do Euro

Nós estamos cobrindo cada passo da crise da Zona do Euro. O drama à medida que os governos quebram e as poupanças de uma vida são perdidas, a reação dos líderes mundiais, as medidas de austeridade e os protestos contra elas. Todos os dias, no Wall Street Journal, existem gráficos de desemprego, queda do PIB, queda dos mercados mundiais. É gradual. É paralizante.

Os editores da Primeira Página marcam uma reunião para discutir ideias para a cobertura do fim de ano e, assim que saímos da reunião, fico a pensar: como deve ser viver esta situação?

Será como 2008, quando fui despedido e não parava de aparecer más noticias? Lembro que falávamos dos nossos empregos, trabalho e dinheiro todas as noites durante o jantar, quase esquecendo como isso poderia irritar a minha filha. E os fins de semana foram os piores. Tentei negar o medo que parecia estar permanentemente fungando na minha nuca e a ansiedade comprimindo as minhas costelas. Era assim que se sentia agora uma família na Grécia ou na Espanha?

Voltei e segui Mike Allen, o editor da primeira página, até ao seu escritório e lancei a ideia de contar a crise através das famílias na Zona do Euro, olhando primeiro para os dados, encontrando perfis demográficos para entender o que constituía uma família e depois realçando isso juntamente com imagens e entrevistas em áudio. Utilizaríamos belos retratos, as vozes e os dados.

De volta à minha mesa, escrevi um resumo e desenhei um logotipo.

Imagem 6. A Crise da Zona do Euro: resumo (Wall Street Journal)

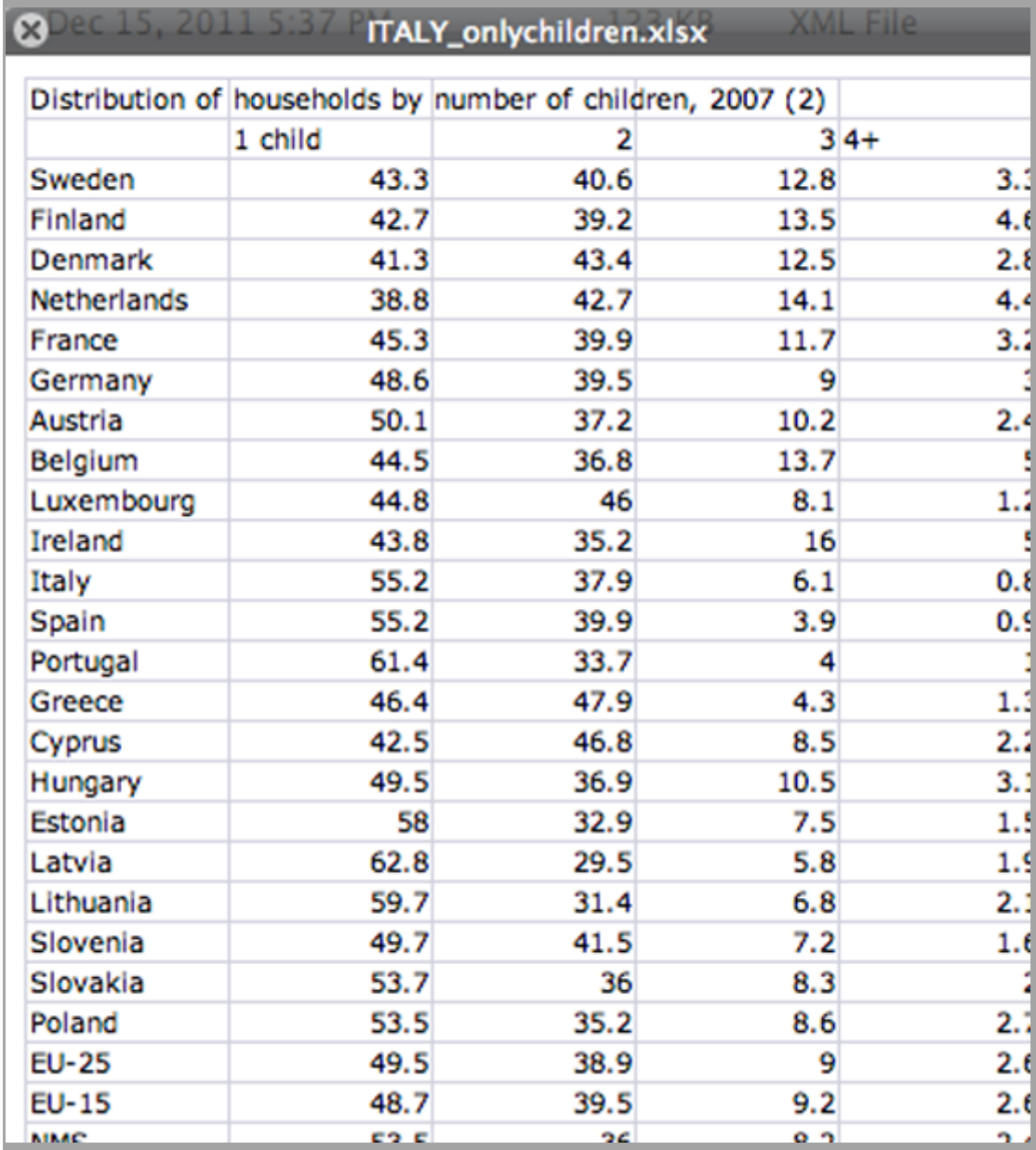
Durante as três semanas seguintes fui à caça dos números: métricas sobre casamento, mortalidade, tamanho da família e gastos com a saúde. Li sobre condições de vida e números de divórcio, vi questionários sobre bem-estar e taxas de poupança. Pesquisei nos departamentos nacionais de estatísticas, telefonei ao escritório do Population Bureau da ONU, ao FMI, Eurostat e OCDE até que encontrei um economista que tinha passado a sua carreira acompanhando famílias europeias. Ele me levou até uma especialista em composições familiares, que me indicou vários documentos sobre o assunto.

Com o meu editor, Sam Enriquez, reduzi o número de países. Juntamos uma equipe para discutir a abordagem visual e quais repórteres poderiam nos entregar palavras, áudios e histórias. Matt Craig, o editor de fotografia da primeira página, iniciou o trabalho de encontrar os fotógrafos. Matt Murray, Vice-Chefe de Redação para cobertura global, enviou um memorando aos diretores das sucursais solicitando a ajuda dos repórteres (isto foi crucial: aprovação da direção).

Mas primeiro, aos dados. Durante as manhã, exportava os dados para planilhas e construía gráficos para identificar tendências: redução das poupanças, desaparecimento das pensões, mães voltando ao trabalho, gastos na saúde, juntamente com a dívida do governo e desemprego. Durante as tardes eu analisava os grupos de dados, comparando países para encontrar histórias.

Fiz isto durante uma semana até me perder e começar a duvidar de mim mesmo. Talvez fosse a abordagem errada. Talvez não fosse sobre países, mas sobre pais e mães, e crianças e avós. Os dados cresciam.

E encolhiam. Às vezes passava horas coletando informação apenas para perceber que ela me dizia, bem, nada. Que eu tinha obtido conjuntos de dados completamente errados. Algumas vezes os dados eram muito velhos.


ITALY_onlychildren.xlsx

Distribution of	households by	number of children, 2007 (2)		
	1 child	2	3	4+
Sweden	43.3	40.6	12.8	3.
Finland	42.7	39.2	13.5	4.
Denmark	41.3	43.4	12.5	2.
Netherlands	38.8	42.7	14.1	4.
France	45.3	39.9	11.7	3.
Germany	48.6	39.5	9	
Austria	50.1	37.2	10.2	2.
Belgium	44.5	36.8	13.7	
Luxembourg	44.8	46	8.1	1.
Ireland	43.8	35.2	16	
Italy	55.2	37.9	6.1	0.
Spain	55.2	39.9	3.9	0.
Portugal	61.4	33.7	4	
Greece	46.4	47.9	4.3	1.
Cyprus	42.5	46.8	8.5	2.
Hungary	49.5	36.9	10.5	3.
Estonia	58	32.9	7.5	1.
Latvia	62.8	29.5	5.8	1.
Lithuania	59.7	31.4	6.8	2.
Slovenia	49.7	41.5	7.2	1.
Slovakia	53.7	36	8.3	
Poland	53.5	35.2	8.6	2.
EU-25	49.5	38.9	9	2.
EU-15	48.7	39.5	9.2	2.

Imagem 7. Julgar a utilidade de um conjunto de dados pode ser uma tarefa bastante demorada (Sarah Slobin)

E então os dados ganharam corpo novamente assim que percebi que ainda tinha perguntas, e que ainda não entendia as famílias. Precisava ver, dar forma a eles. Então fiz um conjunto de gráficos no Illustrator e comecei a ajustá-los e editá-

los. Assim que que os gráficos surgiam, também surgia um retrato coeso das famílias.

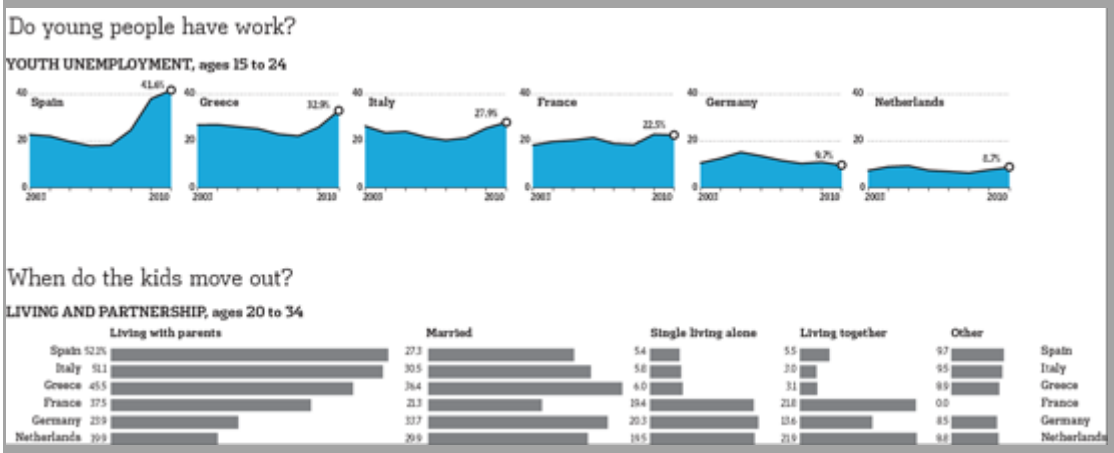


Imagem 8. Visualização de Gráficos: entendendo tendências e padrões escondidos nas bases de dados (Sarah Slobin)

Imagem 9. Números são pessoas: o valor dos dados está nas histórias individuais que eles representam (Wall Street Journal)

Então, começamos. Liguei para cada repórter. Enviei-lhes os gráficos, a ideia geral e um convite aberto para encontrarem histórias que sentissem serem significativas, que aproximassem a crise aos nossos leitores. Precisávamos de uma pequena família em Amsterdã, e outras maiores na Espanha e na Itália. Queríamos ouvir múltiplas gerações para ver como a história pessoal moldava as respostas de cada uma.

A partir daí, acordava cedo para verificar o meu e-mail, levando em conta a diferença de fuso horário. Os repórteres responderam com belos assuntos, sumários, e surpresas que eu não tinha previsto.

Para a parte fotográfica, sabíamos que queríamos retratos de gerações. A ideia do Matt era fazer com que os seus fotógrafos acompanhassem um membro da família ao longo de um dia de suas vidas. Ele escolheu jornalistas visuais que tinham coberto assuntos internacionais, noticias e até guerras. Matt queria que cada sessão de fotos terminasse na mesa de jantar. Sam sugeriu que incluíssemos os menus.

A partir de então, foi uma questão de esperar para ver que história as fotos contavam. Para ver o que as famílias diziam. Desenhamos o visual do aplicativo interativo. Roubei uma paleta dum livro do Tintin, trabalhamos na interação. E quando estava tudo reunido e tínhamos os storyboards, voltamos a acrescentar alguns (não muitos, mas alguns) dos gráficos originais. Apenas o suficiente para pontuar cada história, apenas o suficiente para dar corpo aos temas. Os dados tornaram-se uma pausa na história, uma maneira de alterar o ritmo.

No fim, os dados eram as pessoas: elas eram as fotografias e as histórias. Elas eram o que emoldurava cada narrativa e conduzia a tensão entre os países.

Quando publicamos, logo antes do Ano Novo, conhecia todos os membros das famílias pelo nome. Ainda penso em como estão agora. E se isto não parece um projeto de dados, por mim tudo bem. Porque todos esses momentos que estão documentados no Vida na Zona do Euro, essas histórias de sentar para uma refeição e falar sobre o trabalho e a vida com a sua família eram algo que podíamos dividir com os nossos leitores. Entender os dados foi o que tornou isso possível.

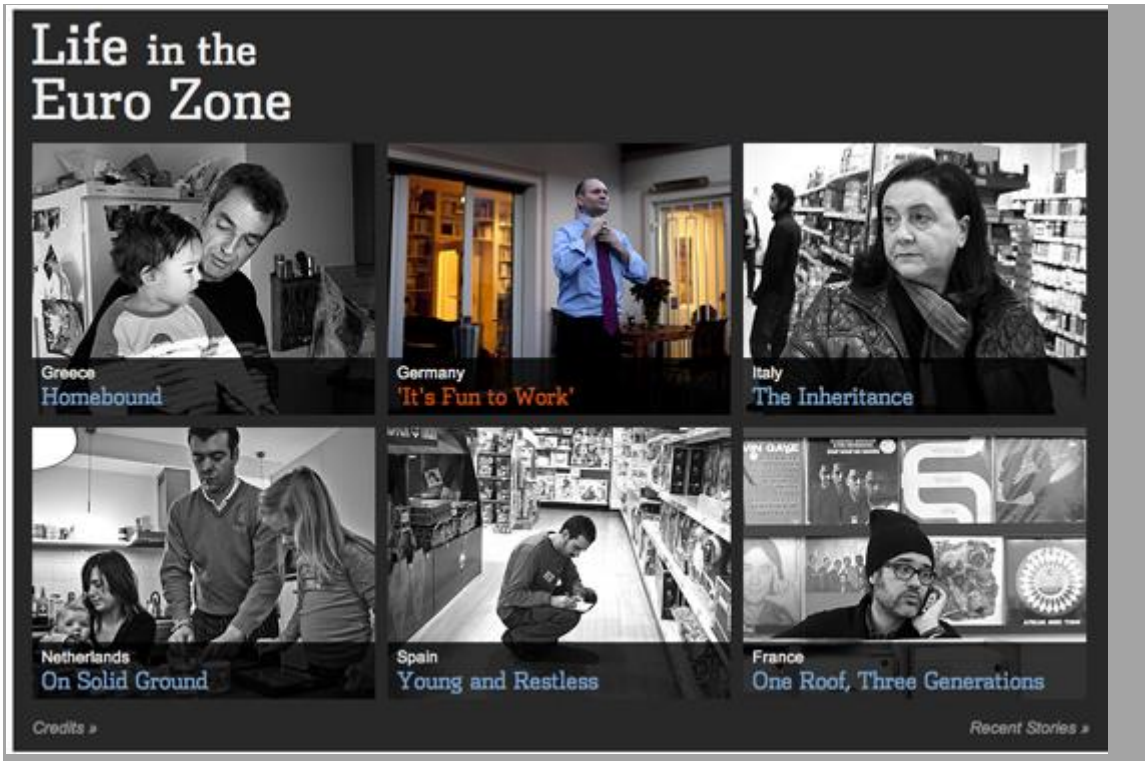


Imagem 10. Vida na Zona do Euro (Wall Street Journal)

— Sarah Slobin, Wall Street Journal

lamentares finlandesas e financiamento de campanha

0 0

Cobrindo o gasto público com OpenSpending.org

Em 2007, Jonathan chegou à Open Knowledge Foundation com uma proposta de uma página para um projeto chamado *Where Does My Money Go?* (Para onde vai o meu dinheiro?), que tinha o objetivo de tornar mais fácil aos cidadãos do Reino Unido entender como as verbas públicas eram gastas. O projeto foi pensado como o protótipo de uma iniciativa maior para reproduzir visualmente informações púbicas, baseada no trabalho pioneiro do Isotype Institute de Otto e Marie Neurath, na década de 1940.

Imagem 11. Where Does My Money Go? (Open Knowledge Foundation)

O projeto permitiu aos usuários explorar dados públicos de várias fontes usando ferramentas intuitivas de código aberto. Ganhamos um prêmio para ajudar a desenvolver um protótipo, e posteriormente recebemos financiamento do 4IP (fundo de inovação do canal Channel 4) para transformá-lo num aplicativo web completo. O guru do design da informação David McCandless (do Information is Beautiful) criou visualizações diferentes dos dados que ajudaram as pessoas a se relacionar com os grandes números — incluindo a "Country and Regional Analysis", que mostra como o dinheiro é gasto nas diferentes partes do país e "Daily Bread", que mostra aos cidadãos um detalhamento de quantas libras são pagas por dia em impostos.

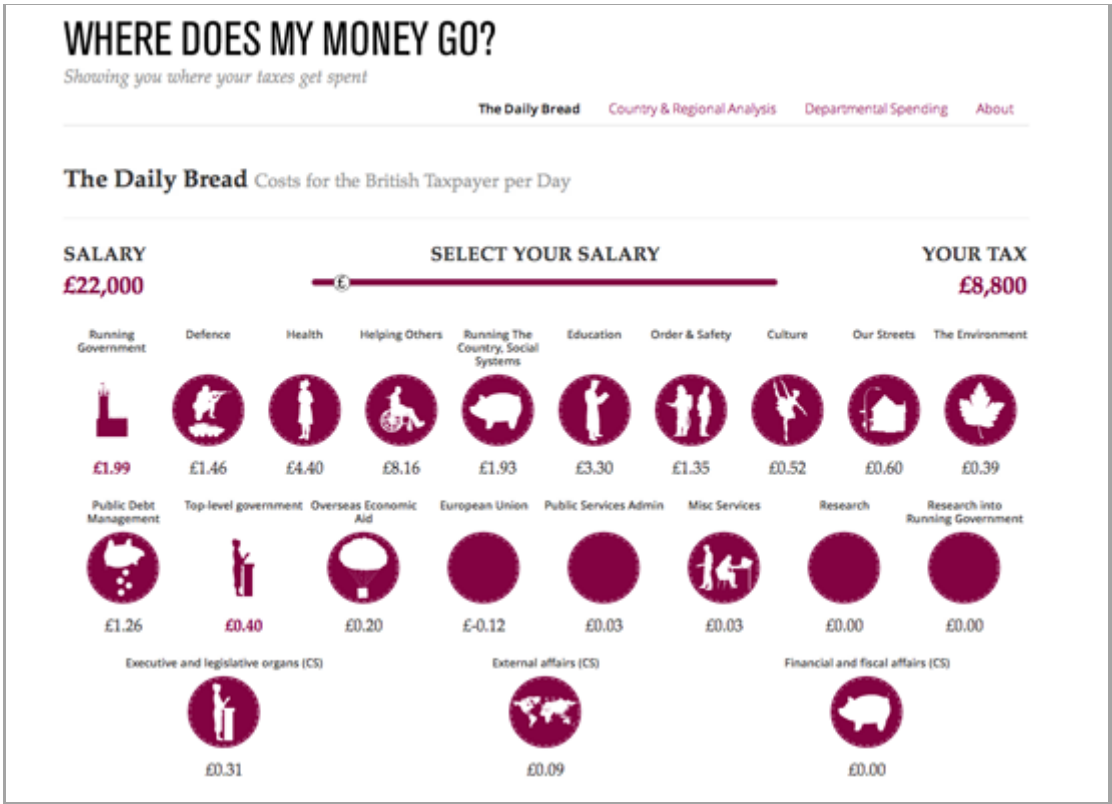


Imagem 12. A calculadora de impostos do Daily Bread do projeto "Where Does My Money Go?" (Open Knowledge Foundation)

Naquela época, o santo graal para o projeto eram os dados do COINS, acrônimo paraCombined Online Information System (Sistema Online de Informações Combinadas), o banco de dados mais abrangente e detalhado das finanças do governo do Reino Unido. Trabalhando com Lisa Evans (antes de ela integrar o time do Guardian Datablog), Julian Todd, Francis Irving (agora no famoso Scraperwiki) e Martin Rosenbaum (BBC), entre outros, nós preenchemos inúmeros requerimentos para obter os dados — sem sucesso em muitos deles.

Quando os dados foram finalmente liberados, em meados de 2010, o fato foi considerado uma grande vitória pelos defensores da transparência. Ganhamos acesso avançado aos dados para carregá-los no nosso projeto, e recebemos uma atenção significativa da imprensa quando isso se tornou público. No dia da liberação dos dados, havia dúzias de jornalistas no nosso canal no IRC questionando sobre como abri-los e explorá-los (os arquivos tinham dezenas de gigabytes). Enquanto alguns especialistas afirmaram que a liberação em massa dos dados era tão complicada que estava escondendo por meio de transparência, muitos jornalistas se debruçaram sobre os eles para dar a seus leitores um retrato sem precedentes de como as verbas públicas são gastas. O Guardian criou um blog em tempo real sobre a liberação e muitos veículos da

mídia cobriram o assunto e ofereceram análises e descobertas a partir dos dados.

Não demorou até que começássemos a receber solicitações e pedidos de informação para a execução de projetos semelhantes em outros países. Pouco tempo depois de lançarOffenerHaushalt — uma versão do projeto para o orçamento do Estado alemão criado por Friedrich Lindenberg — nós lançamos o OpenSpending, uma versão internacional para ajudar os usuários a mapear os gastos públicos ao redor do mundo, um pouco como o OpenStreetMap os ajudou a mapear aspectos geográficos. Implementamos novos designs com a ajuda do talentoso Gregor Aisch, parcialmente baseados nos designs originais de David McCandless.

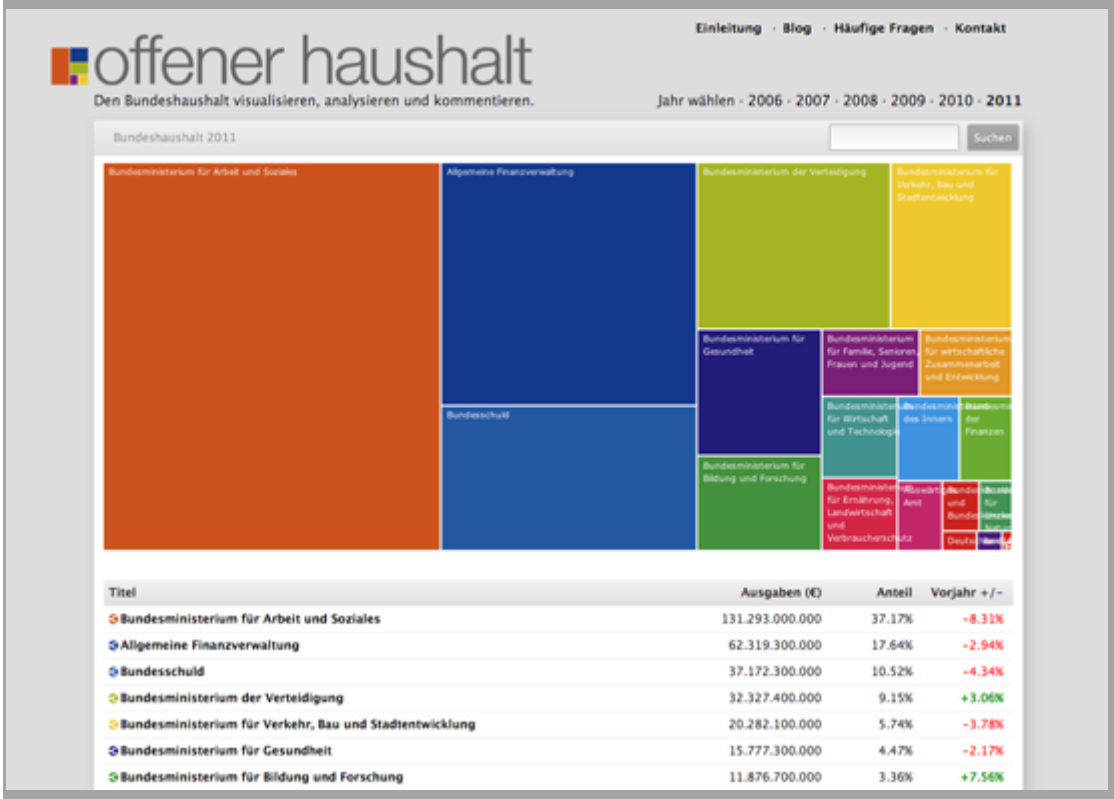


Imagem 13. OffenerHaushalt, a versão alemã do Where Does My Money Go? (Open Knowledge Foundation)

Com o projeto OpenSpending, trabalhamos extensivamente com jornalistas para obter, representar, interpretar e exibir dados sobre gastos ao público. OpenSpending é um enorme banco de dados pesquisável de gastos públicos — tanto de informações orçamentárias de alto nível quanto de operações de gastos efetivos. Qualquer um pode carregar informações de seu município e produzir visualizações a partir delas.

Inicialmente pensávamos que haveria maior demanda por nossas visualizações mais sofisticadas, mas depois de conversar com organizações jornalísticas percebemos que havia necessidades mais básicas, como a capacidade de inserir tabelas dinâmicas de dados nas postagens de seus blogs. Querendo encorajar as organizações jornalísticas a dar acesso público aos dados ao longo de suas histórias, construímos um programa para isso também.

Nosso primeiro grande lançamento foi na época do primeiro Festival Internacional de Jornalismo em Perugia. Um grupo de programadores, jornalistas e funcionários do governo colaboraram para carregar dados da Itália na plataforma OpenSpending, o que gerou uma rica visão de como os gastos estavam divididos entre a administração central e as administrações regionais e locais. O lançamento ganhou cobertura do Il Fatto Quotidiano, Il Post, La Stampa, Repubblica, e Wired Italia, assim como do Guardian.

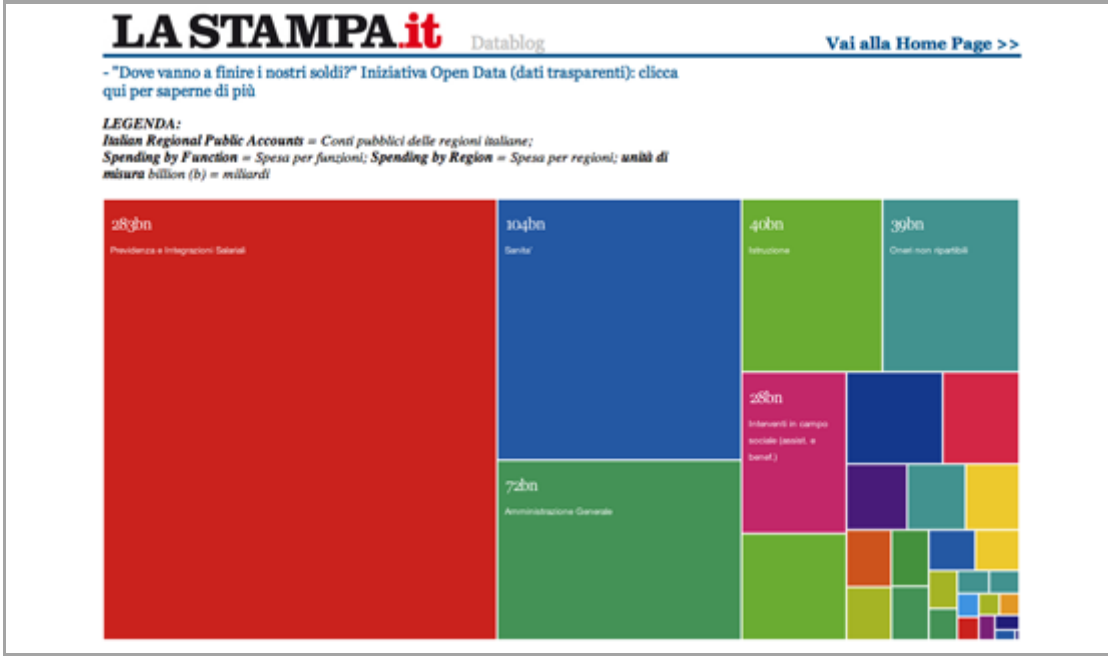


Imagem 14. A versão italiana do Where Does My Money Go? (La Stampa)

Em 2011 nós trabalhamos com o Publish What You Fund (Publique o que você financia) e oOverseas Development Institute para mapear o financiamento da ajuda humanitária a Uganda entre 2003 e 2006. Pela primeira vez você podia ver o fluxo do financiamento dentro do orçamento nacional — permitindo ver até que ponto as prioridades dos doadores se alinhavam com as prioridades do governo. Houve alguns resultados interessantes — por exemplo, tanto programas de combate ao HIV como de planejamento familiar se revelaram

como quase totalmente financiados por doadores externos. Isto foi coberto pelo Guardian.

Nós também vínhamos trabalhando com ONGs e grupos de ativistas para cruzar dados de gastos com outras fontes de informações. Por exemplo, a Privacy International nos procurou com uma grande lista de empresas de tecnologia de segurança e uma lista de agências que compareceram a uma famosa feira internacional de segurança, conhecida informalmente como o "baile dos arapongas". Ao relacionar os nomes das companhias com dados de gastos públicos, foi possível identificar quais delas possuíam contratos com o governo — que poderiam então ser investigados por meio de pedidos oficiais de informação com base no FOI (Freedom of Information Act). O Guardian cobriu essa história.

Trabalhamos atualmente para aumentar o conhecimento fiscal entre os jornalistas e o público, como parte de um projeto chamado Spending Stories, que permite aos usuários relacionar dados sobre gastos públicos com reportagens ligadas a esses gastos, para mostrar os números por trás das notícias.

Por meio de nosso trabalho nesta área, nós aprendemos que:

- Jornalistas frequentemente não estão acostumados a trabalhar com dados brutos, e muitos não consideram isto um fundamento necessário para sua reportagem.
- Analisar e compreender dados é um processo que requer dedicação intensiva de tempo, ainda que se possua as habilidades necessárias. Encaixar isto no ciclo curto do noticiário é difícil, de maneira que o jornalismo de dados é frequentemente usado em projetos investigativos de longo prazo.
- Dados divulgados por governos estão muitas vezes incompletos ou desatualizados. Muito frequentemente, bancos de dados públicos não podem ser usados para fins investigativos sem o acréscimo de informações mais específicas requisitadas por meio de lei de acesso à informação.
- Grupos de ativistas, especialistas e pesquisadores geralmente dispõem de mais tempo e recursos que jornalistas para conduzir pesquisas mais extensivas baseadas em dados. Pode ser muito proveitoso se juntar a eles para trabalhar em equipe.

— Lucy Chambers e Jonathan Gray, Open Knowledge Foundation

Eleições parlamentares finlandesas e financiamento de campanha

Recentemente houve julgamentos relacionados ao financiamento das campanhas nas eleições gerais finlandesas de 2007.

Depois das eleições de 2007, a imprensa descobriu que as leis sobre divulgação de financiamento de campanha não tiveram efeito sobre os políticos. Basicamente, o financiamento de campanha tem sido usado para comprar favores de políticos, que não declararam as origens de seus financiamentos como mandam as leis finlandesas.

Após esses incidentes, as leis tornaram-se mais rigorosas. Depois das eleições gerais de março de 2011, o jornal Helsingin Sanomat decidiu explorar cuidadosamente todos os dados disponíveis sobre o financiamento de campanha. A nova lei determina que o financiamento eleitoral deve ser declarado, e apenas doações abaixo de 1.500 euros podem ser anônimas.

1. Procura de dados e desenvolvedores

O jornal Helsingin Sanomat tem hackatonas desde março de 2011. Nós convidamos programadores, jornalistas e designers gráficos finlandeses para o porão do nosso prédio. Os participantes são divididos em grupos de três e encorajados a desenvolver aplicações e visualizações. Tivemos até agora, em cada um dos nossos três eventos, cerca de 60 participantes. Nós decidimos que os dados de financiamento de campanha deviam ser o foco da hackatona HS Open 2, de maio de 2011.

A Agência Nacional de Auditoria da Finlândia é a autoridade que mantém os registros de financiamento de campanha. Essa foi a parte mais fácil. O chefe de tecnologia de informação Jaakko Hamunen construiu um website que permite o acesso, em tempo real, ao banco de dados. A Agência de Auditoria fez o website em apenas dois meses depois do nosso pedido.

O website http://www.vaalirahoitus.fi disponibilizará ao público e à imprensa a partir de agora informações sobre o financiamento de campanha para cada eleição.

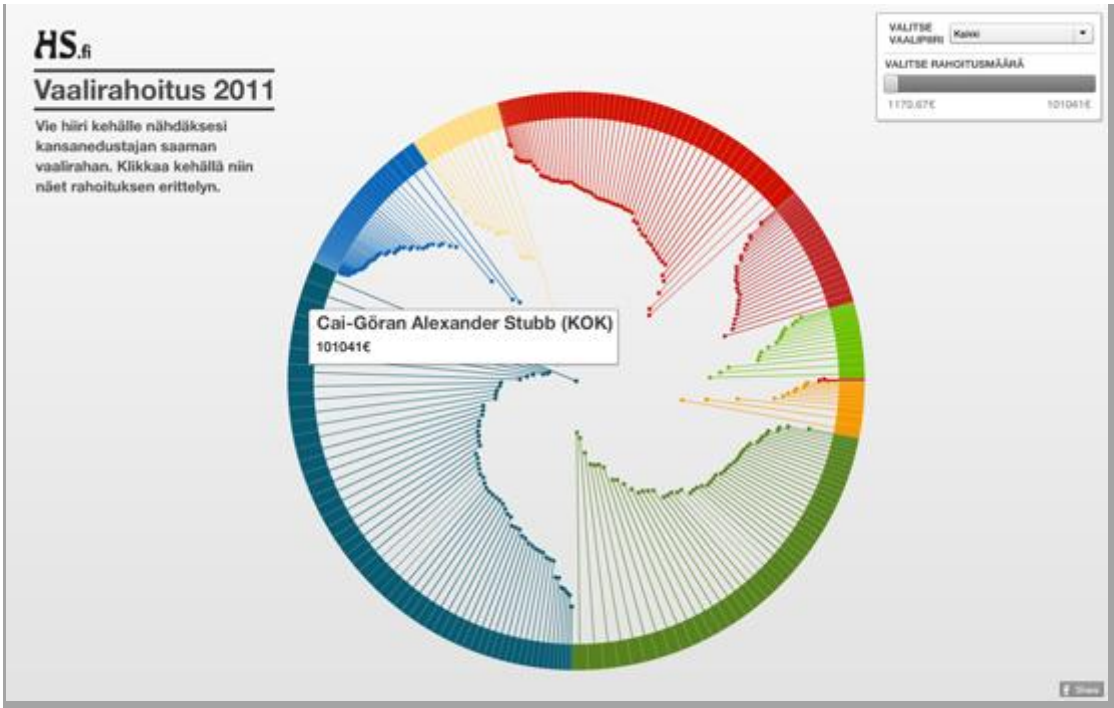


Imagem 15. Financiamento de campanhas (Helsingin Sanomat)

2. Brainstorm de ideias

Os participantes do HS Open 2 chegaram a vinte propostas diferentes sobre o que fazer com os dados. Você pode encontrar todos os protótipos em nosso website (texto em finlandês).

Uma pesquisadora de bioinformática chamada Janne Peltola notou que os dados de financiamento de campanha pareciam os dados genéticos que ela pesquisa, no que diz respeito a conter muitas interdependências. Em bioinformática existe uma ferramenta de código aberto chamada Cytoscape que é usada para mapear estas interdependências. Então nós movemos os dados através do Cytoscape, e construímos um protótipo muito interessante.

3. Implementar a ideia no papel e na web

A lei sobre o financiamento de campanhas estabelece que os membros eleitos do parlamento devem declarar o financiamento até dois meses após as eleições. Na prática, isso significa que conseguimos os dados na metade de junho. Durante o HS Open, tínhamos dados apenas da parcela de prestação de contas que os políticos haviam apresentado antes do prazo final.

Houve também um problema com o formato dos dados. A Agência Nacional de Auditoria providenciou os dados como dois arquivos CSV. Um continha o orçamento total das campanhas e o outro listava o total de doadores. Nós

tivemos que combinar esses dois, criando um arquivo que continha três colunas: doador, recebedor e total. Se os políticos tinham provido todo o dinheiro da própria campanha, no nosso formato de dados aparecia Politico A doou X euros para Politico A. Contra-intuitivo, talvez, mas isso funcionou no Cytoscape.

Quando os dados foram limpos e reformatados, logo os passamos pelo Cytoscape. Depois, o nosso departamento gráfico fez uma página inteira de gráficos externos.

Finalmente, criamos uma belíssima visualização no nosso site. Não foi um gráfico de análise de rede. Queríamos dar às pessoas uma maneira fácil de explorar quanto existe de financiamento de campanha e quem financia. A primeira visualização mostra a distribuição de financiamento entre os membros do parlamento. Quando você clica em um membro, você detalha os resultados dos financiamentos dele. Você também pode votar se determinado doador é bom ou não. A visualização foi feita por Juha Houvinen e Jukka Kokko, de uma agência chamada Satumaa. A versão web de visualização de financiamento de campanha utiliza os mesmos dados que a análise de rede.

4. Publicar os dados

Claro, a Agência Nacional de Auditoria já publicou os dados, por isso não houve a necessidade de republicar. Mas, como nós havíamos tratado os dados e os colocado em uma estrutura melhor, decidimos republicá-los. Distribuímos os nossos dados com a licençaCreative Commons Atribuição 3.0.

Usamos para o projeto Excel e Google Refine para a limpeza e a análise de dados; Cytoscape para a análise de rede; e Illustrator e Flash para a visualização. O Flash deveria ter sido HTML5, mas nós já estávamos trabalhando fora do tempo estipulado.

O que aprendemos? Talvez a lição mais importante foi a de que as estruturas de dados podem ser muito difíceis. Se os dados originais não estão no formato adequado, recalculá-los e convertê-los vai demorar muito tempo.

— Esa Mäkinen, Helsingin Sanomat

Hack Eleitoral em tempo real (Hacks/Hackers Buenos Aires)

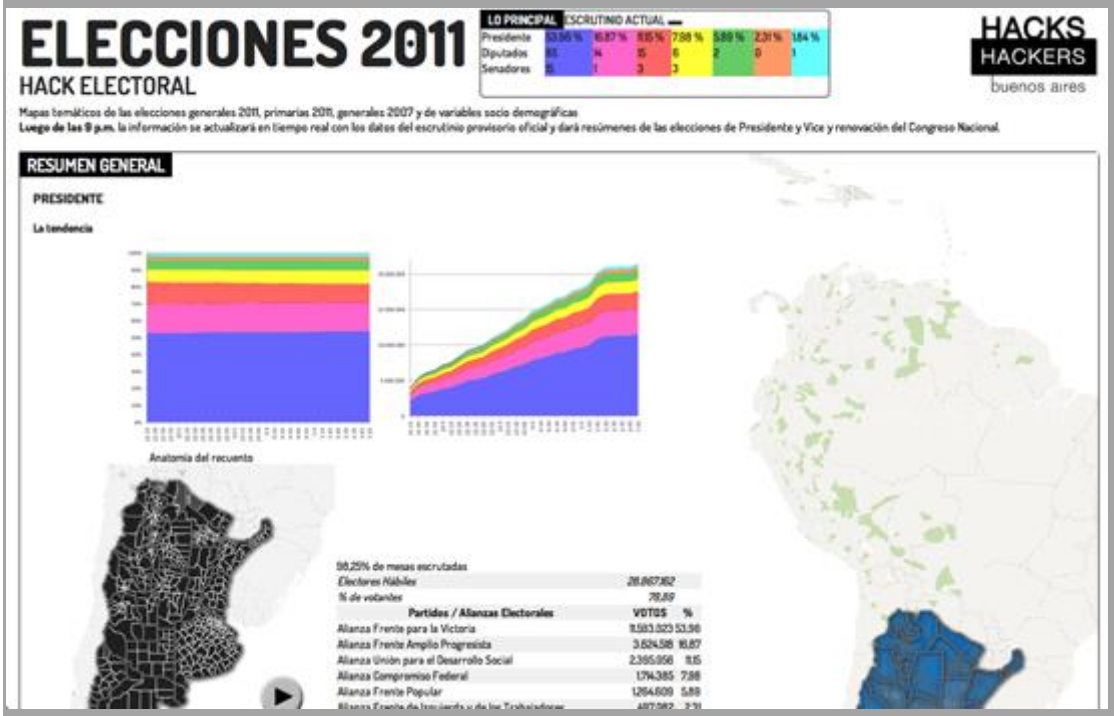


Imagem 16. Eleições 2011 (Hacks/Hackers Buenos Aires)

Hack Eleitoral é um projeto que exibe dados dos resultados parciais das eleições de outubro de 2011 na Argentina. O sistema também conta com informações de eleições anteriores e estatísticas sociodemográficas de todo o país. O projeto foi atualizado em tempo real com informações da contagem dos votos das eleições nacionais de 2011 na Argentina e fornecia parciais. Foi uma iniciativa do Hacks/Hackers Buenos Aires com o analista político Andy Tow. Um esforço colaborativo de jornalistas, programadores, designers, analistas, cientistas políticos e outros membros do Hacks/Hackers local.

Que dados nós usamos?

Todos os dados vieram de fontes oficiais: a Administração Nacional Eleitoral forneceu acesso aos dados da contagem provisória de votos pela Indra (empresa que compila o resultado da votação em todo o país); o Ministério do Interior forneceu os dados sobre os cargos eletivos e os candidatos dos diferentes partidos; um projeto de uma universidadeforneceu as informações biográficas e a plataforma política de cada chapa presidencial; informações sociodemográficas vieram do Censo Nacional da População e Habitação de 2001, do Censo 2010 (Indec) e do Ministério da Saúde.

Como o sistema foi desenvolvido?

O aplicativo foi gerado durante a hackatona (maratona hacker) Eleições 2011, promovida pelo Hacks/Hackers Buenos Aires na véspera das eleições. O evento teve a participação de 30 voluntários de diferentes áreas. O Hack Eleitoral foi desenvolvido como uma plataforma aberta que poderia ser melhorada com o tempo. Usamos as ferramentas Google Fusion Tables, Google Maps e bibliotecas de imagens vetoriais.

Nós trabalhamos na construção de polígonos para a exibição do mapeamento geográfico e da demografia eleitoral. Combinando polígonos de um software de GIS (Sistema de Informações Geográficas, na sigla em inglês) com a geometria de tabelas de dados públicos do Google Fusion Tables, geramos tabelas com chaves correspondentes ao banco de dados eleitoral do Ministério do Interior, da Indra, e aos dados sociodemográficos do Indec. A partir daí, criamos as visualizações no Google Maps.

Usando a API do Google Maps, publicamos diversos mapas temáticos representando a distribuição espacial da votação por meio de diferentes tons de cor, nos quais a intensidade da cor representava o percentual de votos de cada uma das várias chapas presidenciais nos diferentes departamentos administrativos e locais de votação, com destaque especial para os principais centros urbanos: a cidade de Buenos Aires, os 24 distritos da Grande Buenos Aires, a cidade de Córdoba, e Rosário.

Nós usamos a mesma técnica para gerar mapas temáticos de eleições anteriores (as primárias presidenciais de 2011 e a eleição de 2007), assim como da distribuição dos dados sociodemográficos, como níveis de pobreza, mortalidade infantil e qualidade de vida, permitindo uma comparação histórica. O projeto também mostrou a distribuição espacial dos diferentes percentuais de votação obtidos por cada chapa nas eleições gerais de outubro comparados às primárias de agosto.

Mais tarde, usando dados da contagem parcial dos votos, criamos um mapa animado representando a anatomia da contagem, no qual o progresso na contagem é mostrado desde o encerramento dos locais de votação até o dia seguinte.

Prós

- Nós partimos com o objetivo de encontrar e apresentar dados, e conseguimos fazer isso. Tínhamos à mão o banco de dados sociodemográficos do UNICEF sobre a infância, assim como o banco de dados dos candidatos, criado pelo grupo yoquierosaber.org da Universidade Torcuato Di Tella. Durante a hackathona, reunimos um grande volume de dados suplementares que terminamos não incluindo.
- Ficou claro que o trabalho jornalístico e de programação foi enriquecido pelo conhecimento acadêmico. Sem a contribuição de Andy Tow e de Hilario Moreno Campos, teria sido impossível alcançar os objetivos do projeto.

Contras

- Os dados sociodemográficos que conseguimos usar não estavam atualizados (a maioria era do censo de 2001) e não eram muito detalhados. Por exemplo, eles não incluíam detalhes sobre o PIB local, a principal atividade econômica, o nível de escolaridade, o número de escolas, a quantidade de médicos per capita, e muitas outras coisas que teriam sido ótimas de se ter.
- O sistema foi planejado inicialmente para ser uma ferramenta que pudesse ser usada para combinar e exibir quaisquer dados, assim os jornalistas poderiam facilmente exibir dados que os interessassem na internet. Mas tivemos que deixar isso para uma outra oportunidade.
- Como o projeto foi construído por voluntários num curto espaço de tempo, foi impossível fazermos tudo que queríamos. Entretanto, alcançamos um grande progresso na direção certa.
- Pelo mesmo motivo, todo o trabalho colaborativo de 30 pessoas terminou concentrado em apenas um programador quando os dados fornecidos pelo governo começaram a chegar, e nós também enfrentamos alguns problemas ao importar dados em tempo real. Esses problemas foram resolvidos em poucas horas.

Consequências

A plataforma Hack Eleitoral teve um grande impacto na mídia, com cobertura em televisão, rádio, impresso e on-line. Mapas do projeto foram utilizados pelos diferentes meios de comunicação durante a eleição e nos dias seguintes. Com o passar dos dias, os mapas e visualizações eram atualizados, o que aumentou ainda mais o tráfego de dados. No dia da eleição, o site criado na data recebeu

cerca de 20 mil visitantes únicos, e seus mapas foram reproduzidos na primeira página do jornal Página/12, por dois dias seguidos, assim como em reportagens do La Nación. Alguns mapas foram usados na edição impressa do jornal Clarín. Esta foi a primeira vez que a visualização interativa de mapas atualizados em tempo real foi usada na história do jornalismo argentino. Nos mapas principais era possível ver a vitória esmagadora de Cristina Fernandez de Kirchner, por 54% dos votos, ilustrada pela intensidade das cores. Isso também ajudou os usuários a compreender casos específicos em que candidatos locais tiveram vitórias esmagadoras nas províncias.

— Mariano Blejman, Mariana Berruezo, Sergio Sorín, Andy Tow e Martín Sarsale, do Hacks/Hackers Buenos Aires

Dados no Noticiário: WikiLeaks

Começou com um integrante do time de reportagem investigativa perguntando "Você é bom com planilhas, não?" E essa era uma bela de uma planilha: 92.201 linhas de dados, cada uma contendo uma detalhada análise de um evento militar no Afeganistão. Esse era oWikiLeaks war logs. Quer dizer, a Parte um. Havia mais dois episódios para acompanhar: o vazamento do Iraque e dos Telegramas. O termo oficial Base de Dados de Ações Significativas do exército dos Estados Unidos (em inglês, na sigla SIGACTS).

Os diários de guerra do Afeganistão — compartilhados com o The New York Times e o Der Spiegel — eram jornalismo de dados em ação. O que nós queríamos fazer era possibilitar que o nosso time de repórteres especialistas obtivessem grandes histórias por meio da informação — e queríamos analisá-la para obter a visão geral, para mostrar como a guerra está realmente acontecendo.

Era importante para o que faríamos que não publicássemos a base de dados completa. O WikiLeaks já iria fazer isso e nós queríamos ter certeza de que não revelaríamos nomes de informantes ou colocaríamos as tropas da OTAN em perigo desnecessariamente. Ao mesmo tempo, precisávamos tornar mais fácil o uso dos os dados para o nosso time de repórteres investigativos comandados por David Leigh e Nick Davies (que negociaram a liberaração dos dados com Julian Assange). Nós também queríamos tornar mais simples o acesso a informações principais tão clara e abertamente quanto nos era possível.

Os dados vieram a nós como um enorme arquivo de Excel: mais de 92,201 linhas de dados, algumas com nada dentro ou pobremente formatadas. Isso não ajudou em nada os repórteres que tentavam se arrastar entre os dados, em busca de histórias. A base de dados era grande demais para se extrair dali relatórios significativos.

Nosso time construiu um banco de dados simples, usando SQL. Agora, os repórteres poderiam procurar histórias para palavras-chave ou eventos. De repente, o conjunto de dados tornou-se acessível e a criação de histórias tornou-se mais fácil.

Os dados eram bem estruturados: cada evento tinha os seguintes campos: hora, data, uma descrição, número de baixas, e — o que era crucial — latitude e longitude detalhadas.

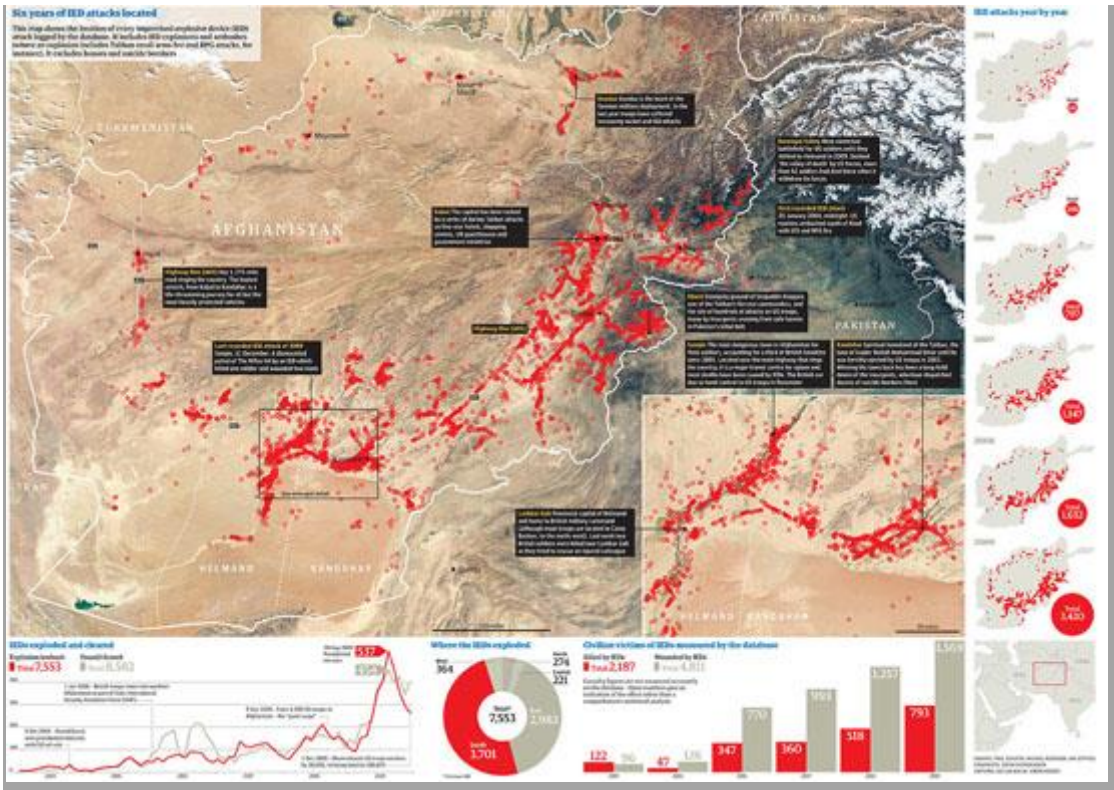


Imagem 17. Os diários de guerra do WikiLeaks (the Guardian)

Também começamos a filtrar os dados para que nos ajudassem a contar uma das principais histórias da guerra: o aumento de ataques com dispositivos explosivos improvisados (IED na sigla em inglês), bombas caseiras de beira de estrada, imprevisíveis e difíceis de combater. Esse conjunto específico de dados ainda era gigante, mas mais fácil de gerenciar. Houve cerca de 7.500 explosões com IEDs ou emboscadas (uma emboscada é onde o ataque é combinado com, por exemplo, pequenas armas de fogo ou granadas-foguete) entre 2004 e 2009. Outros 8.000 IEDs foram encontrados e desarmados. Esses dados nos permitiram ver que o sul do país, onde as tropas Inglesas e Canadenses estavam até então, era a pior área de impacto — o que confirmava as informações de nossos repórteres que cobriram a guerra.

O lançamento dos diários de guerra do Iraque em Outubro de 2010 liberou outros 391.000 registros da guerra para debate público. Em comparação com o vazamento do Afeganistão, atingiu um outro nível. Pode-se dizer que isso fez desta guerra a mais documentada na história. Cada mínimo detalhe estava lá agora, para que pudéssemos analisar e desvendar. Mas um fator se destacava: o volume absoluto de mortes, a maioria de civis.

Assim como com o Afeganistão, o Guardian decidiu não republicar a base de dados inteira, em grande parte porque não conseguíamos ter certeza de que o

campo do sumário poderia conter detalhes confidenciais de informantes e por aí vai.

Mas nós permitimos que nossos usuários fizessem o download da planilha contendo os registros de cada incidente onde alguém morreu, aproximadamente 60.000 no total. Removemos o campo do sumário, deixando apenas os dados básicos: o comando militar, número de mortes, e a classificação geográfica.

Nós também pegamos todos os incidentes em que alguém tenha morrido e os colocamos em um mapa usando Google Fusion tables. Não ficou perfeito, mas um começo na tentativa de mapear os padrões da destruição que devastou o Iraque.

O telegramas foram vazados em dezembro de 2010, em um nível completamente diferente. Era um conjunto enorme de dados de documentos oficiais: 251.287 remessas de mais de 250 embaixadas dos EUA em todo o mundo e consulados. É uma imagem única da atuação diplomática norte-americana — incluindo mais de 50 mil documentos relativos já à administração Obama. Mas o que tinha nos dados?

Os próprios despachos vieram por meio da vasta Rede Roteadora de Protocolos Secretos da Internet ou SIPRNet. A SIPRNet é o sistema militar mundial de internet norte-americano, mantido em separado da internet civil e gerido pelo Departamento de Defesa em Washington. Desde os ataques de setembro de 2001, há um movimento nos EUA para interligar arquivos de informações governamentais, na esperança de que a inteligência-chave não mais fique presa em "stovepipes" (meios de informações verticalizados e isolados). Um número crescente de embaixadas norte-americanas ligou-se à SIPRNet durante a última década, de forma que informações militares e diplomáticas pudessem ser compartilhadas. Em 2002, 125 embaixadas estavam na SIPRNet; Em 2005, eram 180, e, atualmente, a grande maioria das missões dos Estados Unidos em todo o mundo estão ligadas ao sistema — é por isso que a maior parte dos telegramas vazados são de 2008 e 2009. Como David Leigh escreveu:

Uma remessa de uma embaixada marcada como SIPDIS é automaticamente baixada para o website confidencial da embaixada. De lá, ela pode ser acessada não só por qualquer um do departamento de estado, mas, também, por qualquer um no exército dos EUA que possua uma licença de segurança até o nível "Secreto", uma senha, e um computador conectado à SIPRNet.

...o que surpreendentemente está acessível a mais de 3 milhões de pessoas. Há várias camadas de dados projetadas para nunca serem exibidas a cidadãos de fora dos EUA. Pelo contrário, elas deveriam ser lidas por oficiais em Washington do nível da Secretária de Estado Hillary Clinton. Os telegramas são normalmente esboçados pelo embaixador local ou subordinados. Os documentos "Altamente Secretos" e acima da inteligência estrangeira não podem ser acessados do SIPRNet.

Ao contrário dos vazamentos anteriores, isso era predominantemente texto, não quantificável. Isso era o que estava incluído:

Uma fonte

A embaixada ou órgão que o enviou.

Uma lista de destinatários

Normalmente, os telegramas eram enviados para algumas embaixadas e órgãos.

Um campo para assunto

Um resumo do despacho.

Códigos

Cada mensagem foi rotulada com algumas abreviações de palavras-chave.

Corpo de texto

A mensagem em si. Optamos pela não publicação completa destes por razões de segurança óbvias.

Uma nuance interessante é como os telegramas quase criaram vazamentos por demanda. Eles guiaram as notícias por semanas após serem publicados; agora, sempre que uma história sobre um regime corrupto ou escândalo internacional surge, o acesso aos telegrama nos dá a possibilidade de novas histórias.

A análise das correspondências é uma tarefa enorme que pode nunca ser terminada completamente.

— Essa é uma versão editada de um capítulo publicado em "Facts are Sacred: The Power of Data" (Fatos são Sagrados: O Poder dos Dados), de Simon Rogers, the Guardian (publicado no Kindle)

Hackatona Mapa76

Nós lançamos o Hacks/Hackers Buenos Aires em abril de 2011. Tivemos dois encontros iniciais para divulgar a ideia de uma maior colaboração entre jornalistas e desenvolvedores de software, que contaram com 120 a 150 pessoas em cada um dos eventos. Para o terceiro encontro, organizamos uma hackatona de 30 horas com oito pessoas durante uma conferência de jornalismo digital na cidade de Rosário, a 300 quilômetros de Buenos Aires.

Um tema recorrente nos encontros era o desejo de obter grandes volumes de dados da internet e representá-los visualmente. Para ajudar com isso, nasceu o projeto Mapa76, que ajuda usuários a extrair dados e mostrá-los usando mapas e linhas do tempo. Não foi uma tarefa fácil.

Imagem 18. Mapa76 (Hacks/Hackers Buenos Aires)

Por que Mapa76? Em 24 de março de 1976 houve um golpe na Argentina que durou até 1983. Durante esse período, estima-se que tenha havido 30 mil pessoas desaparecidas, milhares de mortes e 500 crianças nascidas durante o cativeiro foram apropriadas pela ditadura militar. Mais de 30 anos depois, o número de pessoas condenadas na Argentina por crimes contra humanidade cometidos durante a ditadura chega a 262 (até setembro de 2011). Há 14 julgamentos iniciados e 7 com datas de início definidas. Há 802 pessoas em vários processos judiciais abertos.

Esses processos geram grandes volumes de dados que são difíceis de serem processados por pesquisadores, jornalistas, organizações de direitos humanos, juízes, promotores e outras pessoas. Os dados são produzidos de forma dispersa e os pesquisadores muitas vezes não tiram proveito de softwares para ajudá-los

com a interpretação. No fim das contas, isto significa que, frequentemente, fatos são ignorados e hipóteses ficam limitadas. Mapa76 é uma ferramenta investigativa que dá livre acesso a essas informações para fins jornalísticos, legais, jurídicos e históricos.

Para nos preparar para a hackatona, criamos uma plataforna que desenvolvedores e jornalistas poderiam usar para colaborar no dia do evento. Martin Sarsale desenvolveu alguns algoritmos básicos que extraía dados estruturados a partir de documentos de texto simples. Algumas bibliotecas do projeto DocumentCloud.org também foram usadas, mas não muitas. A plataforma automaticamente analisava e extraía nomes, datas e locais dos textos — e permitia que os usuários explorassem fatos importantes sobre casos diferentes (por exemplo, data de nascimento, local de prisão, o suposto local do desaparecimento, e assim por diante).

Nosso objetivo era criar uma plataforma para extração automática de dados dos julgamentos da ditadura militar na Argentina. Nós queríamos uma maneira para automaticamente (ou, ao menos, semi-automaticamente) mostrar dados importantes relacionados a casos de 1976-1983 que fossem baseados em evidências escritas, argumentações e julgamentos. Os dados extraídos (nomes, lugares e datas) são coletados, armazenados e podem ser analisados e refinados pelo pesquisador, assim como ser explorado utilizando-se mapas, linhas do tempo e ferramentas de análise de redes.

O projeto vai permitir que jornalistas, pesquisadores, promotores e testemunhas sigam a história da vida de uma pessoa, incluindo o período de prisão e de desaparecimento ou soltura subsequente. Onde houver ausência de informação, os usuários poderão vasculhar um vasto número de documentos em busca de dados que poderão ser relevantes para o caso.

Para a hackatona, fizemos um anúncio por meio do Hacks / Hackers Buenos Aires, que, então, tinha cerca de 200 membros (no momento em que escrevo, são 540). Nós também entramos em contato com várias associações de direitos humanos. A reunião teve a presença de cerca de 40 pessoas, incluindo jornalistas, organizações de advogados, desenvolvedores e designers.

Durante a hackatona, identificamos as tarefas que os diferentes tipos de participantes poderiam exercer independentemente para ajudar as coisas a funcionarem bem. Por exemplo, pedimos aos designers que trabalhassem em uma interface que juntasse mapas e linhas do tempo, pedimos aos

desenvolvedores para analisar a possibilidades para extrair dados estruturados e algoritmos para remover a ambiguidade de nomes, e pedimos aos jornalistas para investigar o que aconteceu com pessoas específicas, para comparar diferentes versões de histórias, e passar um pente fino nos documentos para contar histórias sobre casos particulares.

Provavelmente, o principal problema que tivemos após a hackatona foi que o nosso projeto era muito ambicioso, nossos objetivos de curto prazo demandavam muito trabalho, e é difícil coordenar uma rede frouxa de voluntários. Quase todos os envolvidos com o projeto tiveram um dia intenso de trabalho e muitos também participaram de outros eventos e projetos. O coletivo Hacks/Hackers Buenos Aires fez 9 reuniões em 2011.

O projeto está em constante desenvolvimento. Há um time central de quatro pessoas trabalhando com mais de uma dúzia de colaboradores. Nós temos um grupo de emails público e um repositório de códigos através do qual qualquer um pode se envolver com o projeto.

— Mariano Blejman, Hacks/Hackers Buenos Aires

A cobertura dos protestos violentos no Reino Unido pelo The Guardian

No verão de 2011, o Reino Unido foi tomado por uma onda de manifestações violentas, depredações e saques. Políticos sugeriram que as ações não tinham ligação alguma com a pobreza e que aqueles que participaram dos saques eram simplesmente criminosos. O Primeiro Ministro, com outros líderes conservadores, culparam as mídias sociais por provocarem os quebra-quebras, sugerindo que os saques foram organizados via Facebook, Twitter e Blackberry Messenger (BBM). Houve pedidos para que as plataformas de mídias sociais fossem fechadas temporariamente. Como o governo britânico não investigou porque os quebra-quebras aconteceram, o The Guardian, em colaboração com a Escola Londrina de Economia, construiu o projeto inovador Reading the Riots ("Lendo os Protestos") para esclarecer essa questão.

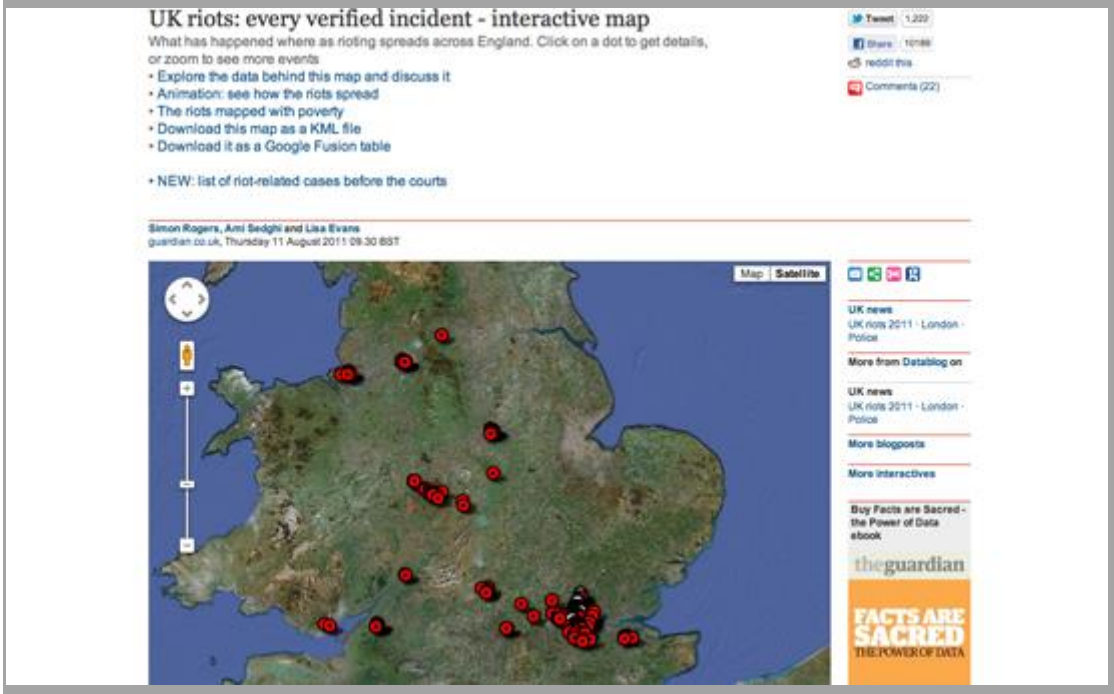


Imagem 19. Os tumultos ingleses: todo incidente checado (The Guardian)

O jornal usou extensivamente jornalismo de dados para entender melhor quem estava participando dos saques e o porquê. Além disso, trabalhou em conjunto com outro time de acadêmicos, liderados pelo professor Rob Procter da Universidade de Manchester, para entender o papel das mídias sociais, muito usadas pelo The Guardian nas reportagens sobre os protestos. A equipe do Reading the Riots foi liderada pelo editor de projetos especiais do The Guardian, Paul Lewis. Durante os protestos, Paul enviou relatos da linha de

frente em cidades ao longo da Inglaterra (principalmente através do seu perfil no Twitter, @paullewis). Esse segundo time trabalhou em cima de 2,6 milhões de tuítes cedidos pelo Twitter. O principal objetivo do trabalho ali foi enxergar como os rumores circularam no Twitter, a função que diferentes usuários/atores tiveram em propagar e espalhar fluxos de informação, ver se a plataforma foi usada para incitar e examinar outras formas de organização.

Em termos do uso do jornalismo de dados e da visualização de dados, é útil separar dois períodos-chave: o das maneiras com que os dados ajudaram a narrar as notícias enquanto os tumultos se desdobravam; e, em seguida, um período de pesquisa muito mais intensa com dois grupos de pesquisadores trabalhando com o The Guardian, para coletar dados, analisá-los e escrever profundas reportagens relatando as conclusões. Os resultados da primeira fase do projeto Reading The Riots foram publicados durante uma semana de exaustivas reportagens, no começo de dezembro de 2011. Abaixo há alguns exemplos de como o jornalismo de dados foi usado nos dois períodos.

Fase um: os tumultos enquanto aconteceram

Usando mapas simples, o time de dados do The Guardian mostrou os locais de tumultos confirmados e, ao integrar os dados de renda e pobreza à localização dos quebra-quebras, começou a desmontar a principal narrativa política de que não havia relação entre saques e pobreza. Ambos os exemplos usaram ferramentas de cartografia inéditas e, no segundo caso, combinou dados de localização com outro conjunto de dados para começar estabelecer outras conexões e relações.

Em relação ao uso das mídias sociais durante os tumultos (no caso, o Twitter), o jornal criou uma visualisação das hashtags relacionadas ao tumultos naquele período, o que ressaltou que o Twitter foi utilizado mais para reagir aos tumultos do que para organizar as pessoas que participariam dos saques, com a hashtag *#riotcleanup* (ou #limpezadotumulto) (campanha espontânea para limpeza das ruas após a confusão) apresentando o pico de crescimento mais significativo.

Fase Dois: Interpretando os protestos

Quando o jornal publicou suas conclusões, após meses de intensa pesquisa e trabalho íntimo com os dois times de acadêmicos, duas visualizações se destacaram e foram amplamente discutidas. A primeira, um pequeno vídeo, mostra o resultado da combinação entre os locais conhecidos em que pessoas

participaram dos quebra-quebras e seus endereços, mostrando assim o que chamamos de "trajeto do tumulto". Aqui o jornal trabalhou com um especialista em cartografia de transporte, ITO World, para modelar a rota mais provável percorrida pelos baderneiros em direção aos locais dos saques, destacando diferentes padrões para diferentes cidades, com alguns viajando grandes distâncias.

A segunda visualização aborda as maneiras com que os rumores se espalharam no Twitter. No debate com a equipe de acadêmicos, sete boatos foram selecionados para análise. Os acadêmicos em seguida coletaram todo os dados relativos a cada boato e bolaram um código de cores que classificou cada tuíte de acordo com quatro características: pessoas simplesmente repetindo o boato (fazendo uma afirmação), rejeitando-o (fazendo um desmentido), questionando-o (interrogação), ou simplesmente comentando-o (comentário). Todos os tuítes foram triplamente codificados e os resultados foram exibidos numa visualização feita pelo time de Interatividade do The Guardian. A equipe do jornaldescreveu como construiu a visualização.

O que é tão admirável nessa visualização é que ela mostra de maneira eloquente algo muito difícil de descrever: a natureza viral dos boatos e a maneira como seus ciclos de vida se desenvolvem ao longo do tempo. O papel da mídia tradicional é evidente em alguns desses boatos (por exemplo, desmascarando-os completamente ou de fato confirmando-os como notícia), como também é a natureza retificadora do próprio Twitter ao lidar com os rumores. Essa visualização não apenas deu grande ajuda à tarefa de contar bem essa história, mas também permitiu a compreensão real de como os rumores se comportam no Twitter, o que oferece informação útil para lidar com eventos como esses no futuro.

O que fica claro com o último exemplo é a sinergia poderosa entre o jornal e um grupo de acadêmicos capazes de analisar profundamente 2,6 milhões de tuítes ligados aos quebra-quebras. Apesar dos acadêmicos terem construído ferramentas originais para suas análises, eles agora estão trabalhando para torná-las disponíveis para qualquer um que queira usá-las, fornecendo uma plataforma para análise. Combinada com o passo-a-passo descrito pela equipe do The Guardian, isso fornece um estudo de caso útil de como a análise de mídias sociais e a visualização podem ser usadas para narrar histórias tão importantes. — *Farida Vis, Universidade de Leicester*

Boletins escolares de Illinois (EUA)

A cada ano, a Secretaria de Educação do Estado de Illinois (EUA) publica os chamados "boletins escolares", dados demográficos e de desempenho de todas as suas escolas públicas. É um conjunto de dados expressivo — a base, em 2011, possuía aproximadamente 9.500 *colunas* de largura. O problema quando se trabalha com essa quantidade de dados é escolher o que apresentar. (Assim como em qualquer projeto de software, o mais complicado não é construir o software, e sim o software *certo*).

Trabalhamos com os repórteres e o editor da equipe de educação para escolher os dados mais interessantes. (Há muitos dados ali que parecem interessantes, mas que um repórter te dirá que, na verdade, tem falhas ou pode levar a conclusões erradas).

Também fizemos uma enquete e entrevistamos colegas da redação que têm crianças em idade escolar. Isso por causa de uma lacuna na equipe de aplicativos de notícias — ninguém tinha filhos nessa faixa etária. Ao longo do caminho, aprendemos muito sobre nosso público e também sobre a usabilidade (ou a falta dela!) da versão anterior de nosso site de escolas.

Imagem 20. 2011 Boletins escolares de Illinois (Chicago Tribune)

Nosso objetivo era desenvolver um projeto para alguns usuários e tipos de uso específicos:

- Pais que querem saber como a escola de seu filho está avaliada

- Pais que estão procurando um lugar para morar, uma vez que a qualidade da escola tem peso significativo nessa decisão.

Na sua primeira versão, o site de escolas era um projeto de seis semanas e dois desenvolvedores. Na atualização que fizemos em 2011, passou a ser de quatro semanas e dois desenvolvedores (na realidade, havia três pessoas trabalhando ativamente no projeto, mas nenhuma em tempo integral — então consideremos duas pessoas).

Uma peça-chave desse projeto era o design da informação. Embora apresentemos uma versão reduzida dos dados, ainda assim há *muitos* dados, e fazer isso tudo ficar compreensível era um desafio. Felizmente, conseguimos trazer para o projeto um designer especialista em apresentar informações complexas. Ele nos guiou a uma apresentação amigável, mas que não subestima a habilidade ou a disposição do leitor de entender os números.

O site foi desenvolvido em Python e Django. Os dados estão hospedados em MongoDB — os dados sobre as escolas são heterogêneos e hierárquicos, não cairia bem numa base de dados relacional (senão, teríamos provavelmente usado PostgreSQL).

Experimentamos pela primeira vez o framework Twitter Bootstrap (um kit de desenvolvimento para criar interfaces na web) nesse projeto, e ficamos satisfeitos com os resultados. Os gráficos foram desenhados com o Flot.

O aplicativo também abriga uma série de reportagens que escrevemos sobre o desempenho das escolas. Funciona como uma espécie de portal no seguinte sentido; quando há uma nova reportagem sobre o desempenho escolar, colocamos no topo do aplicativo, ao lado de listas de escolas relevantes para a matéria (e quando uma nova reportagem ganha repercussão, os leitores do chicagotribune.com são redirecionados para o aplicativo, e não para a reportagem).

Relatórios recentes mostram que os leitores adoram o aplicativo. O retorno que recebemos foi altamente positivo (ou, ao menos, construtivo!), e o número de visitas está bem alto. Para completar, esses dados ainda devem gerar interesse por ao menos um ano — apesar de esperarmos que as visitas diminuam à medida que as reportagens sobre as escolas saiam da página inicial, nossa experiência passada mostra que os leitores continuam a acessar o site ao longo do ano.

Algumas ideias-chave que aprendemos com esse projeto:

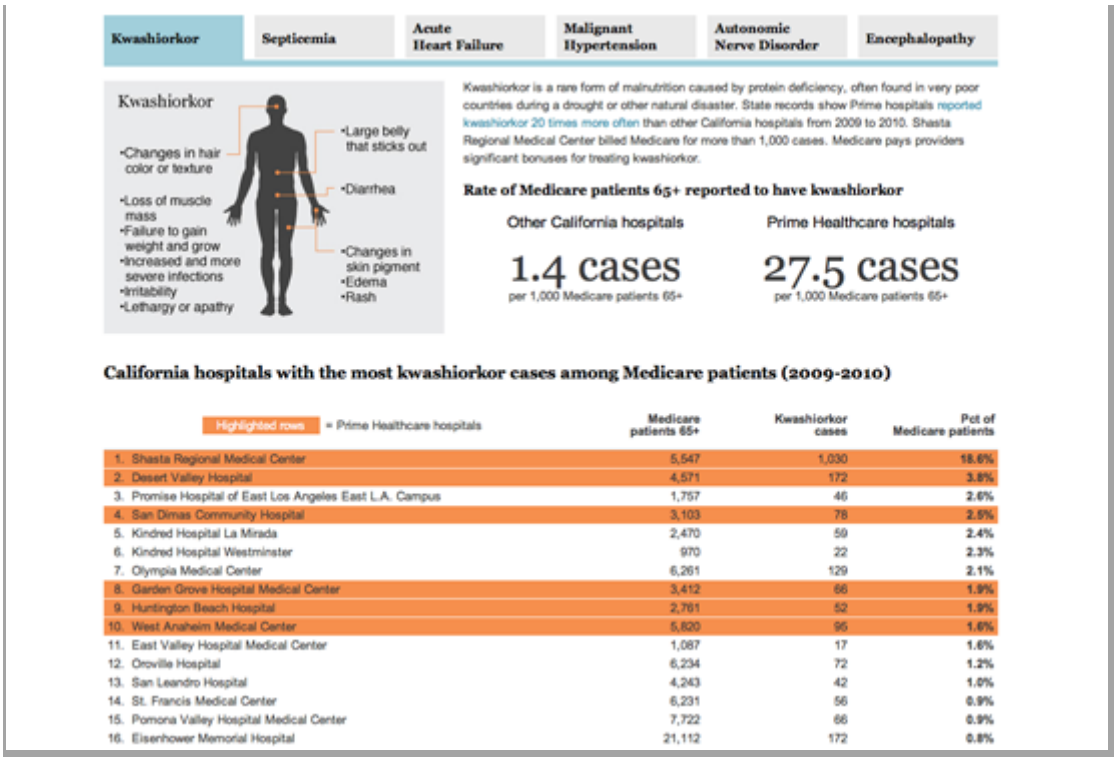
- Os infografistas são seus amigos. Eles são bons em fazer informações complexas ficarem mais palatáveis.
- Peça ajuda à redação. Esse foi o segundo projeto em que realizamos uma enquete e entrevistas com a redação, e foi uma excelente maneira de conhecer a opinião de pessoas atenciosas que, assim como seu público, têm diferentes bagagens e, em geral, sentem certo desconforto com computadores.
- Mostre seu trabalho! Muitos dos retornos que tivemos foram solicitações dos dados que usamos na aplicação. Disponibilizamos muitos deles publicamente via API, e em breve vamos lançar dados que não havíamos pensado em incluir inicialmente.

— Brian Boyer, Chicago Tribune

Faturas de hospitais

Repórteres investigativos da CaliforniaWatch receberam informações de que uma grande rede de hospitais na Califórnia poderia estar burlando de forma sistemática o programa federal Medicare, que paga os custos de tratamentos médicos de americanos com 65 anos ou mais. O esquema denunciado é chamado de *upcoding*, que significa relatar pacientes com condições de saúde mais complicadas — as quais dão o direito a receber um valor de reembolso maior — do que realmente existiam. Mas uma fonte-chave da denúncia era um sindicato que estava brigando com a gerência da rede de hospitais, e a equipe da CaliforniaWatch sabia que seria necessária uma verificação independente para que a história tivesse credibilidade.

Felizmente, o Departamento de Saúde da Califórnia tem documentos públicos que dão informações muito detalhadas sobre cada caso tratado em todos os hospitais do Estado. As 128 variáveis incluem até 25 códigos de diagnóstico da "Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde" (mais conhecida como CID-9), publicada pela Organização Mundial de Saúde (OMS). Embora os pacientes não sejam identificados pelo nome nos registros, outras variáveis dizem a idade do paciente, como os custos são pagos e qual hospital o tratou. Os jornalistas perceberam que, com esses registros, podiam ver se os hospitais pertencentes à rede estavam mesmo relatando certas condições raras a taxas significativamente mais altas do que as verificadas em outros hospitais.



Kwashiorkor | Septicemia | Acute Heart Failure | Malignant Hypertension | Autonomic Nerve Disorder | Encephalopathy

Kwashiorkor

- Changes in hair color or texture
- Loss of muscle mass
- Failure to gain weight and grow
- Increased and more severe infections
- Irritability
- Lethargy or apathy
- Large belly that sticks out
- Diarrhea
- Changes in skin pigment
- Edema
- Rash

Kwashiorkor is a rare form of malnutrition caused by protein deficiency, often found in very poor countries during a drought or other natural disaster. State records show Prime hospitals reported kwashiorkor 20 times more often than other California hospitals from 2009 to 2010. Shasta Regional Medical Center billed Medicare for more than 1,000 cases. Medicare pays providers significant bonuses for treating kwashiorkor.

Rate of Medicare patients 65+ reported to have kwashiorkor

Other California hospitals	Prime Healthcare hospitals
1.4 cases per 1,000 Medicare patients 65+	27.5 cases per 1,000 Medicare patients 65+

California hospitals with the most kwashiorkor cases among Medicare patients (2009-2010)

Highlighted rows = Prime Healthcare hospitals

	Medicare patients 65+	Kwashiorkor cases	Pct of Medicare patients
1. Shasta Regional Medical Center	5,547	1,030	18.6%
2. Desert Valley Hospital	4,571	172	3.8%
3. Promise Hospital of East Los Angeles East L.A. Campus	1,757	46	2.6%
4. San Dimas Community Hospital	3,103	78	2.5%
5. Kindred Hospital La Mirada	2,470	59	2.4%
6. Kindred Hospital Westminster	970	22	2.3%
7. Olympia Medical Center	6,261	129	2.1%
8. Garden Grove Hospital Medical Center	3,412	66	1.9%
9. Huntington Beach Hospital	2,761	52	1.9%
10. West Anaheim Medical Center	5,820	96	1.6%
11. East Valley Hospital Medical Center	1,087	17	1.6%
12. Oroville Hospital	6,234	72	1.2%
13. San Leandro Hospital	4,243	42	1.0%
14. St. Francis Medical Center	6,231	56	0.9%
15. Pomona Valley Hospital Medical Center	7,722	66	0.9%
16. Eisenhower Memorial Hospital	21,112	172	0.8%

Imagem 22. Kwashiorkor (California Watch)

As bases de dados eram muito grandes, quase 4 milhões de registros por ano. Os repórteres queriam estudar o equivalente a seis anos de registros, a fim de ver como os padrões mudaram ao longo do tempo. Eles pediram os dados à agência estatal, que chegaram em CD-ROMs facilmente copiados para um computador. O repórter encarregado da análise de dados usou um sistema chamado SAS para trabalhar com eles. O SAS é muito poderoso (permite a análise de muitos milhões de registros) e é usado por agências governamentais, incluindo o Departamento de Saúde da Califórnia, mas é caro — o mesmo tipo de análise poderia ter sido feito com qualquer uma de uma variedade de outras ferramentas de bancos de dados, como o Microsoft Access ou o open-source MySQL.

Com os dados em mãos e os programas apropriados para estudá-los, encontrar padrões suspeitos seria relativamente simples. Por exemplo, uma das alegações foi de que aquela rede estava relatando vários graus de desnutrição em taxas muito mais elevadas do que as taxas vistas em outros hospitais. Usando o SAS, o analista de dados extraiu tabelas de frequência que mostraram os números de casos de desnutrição relatados a cada ano por cada um dos mais de 300 hospitais de emergência da Califórnia. Em seguida, as tabelas de frequência foram importadas para o Excel para uma inspeção mais próxima dos padrões de

cada hospital. A capacidade do Excel para classificar, filtrar e calcular taxas dos números brutos fez com que os padrões fossem fáceis de identificar.

Foi particularmente notável o fato de existirem relatos de uma condição chamada Kwashiorkor, uma síndrome de deficiência de proteína vista quase que exclusivamente em crianças famintas nos países em desenvolvimento afetados pela falta de alimentos. Ainda assim, os hospitais da rede estavam diagnosticando casos de Kwashiorkor entre californianos idosos em taxas até 70 vezes maiores do que a média de todos os hospitais do Estado.

Em outras reportagens, a análise usou técnicas semelhantes para examinar as taxas informadas de condições raras como a septicemia, encefalopatia, hipertensão maligna e doenças do sistema nervoso autônomo. E outra análise examinou as alegações de que a rede estava transferindo da emergência os para leitos hospitalares percentuais acima do normal de pacientes do Medicare, cujo pagamento para a assistência hospitalar é mais certo do que para a emergência.

Resumindo, reportagens como essas se tornam possíveis quando você usa os dados para produzir evidências e testar de forma independente as denúncias feitas por fontes que poderiam estar enviesadas. Essas histórias também são um bom exemplo da necessidade de fortes leis de acesso à informação; a razão pela qual o governo obriga hospitais a informar esses dados é para que esse tipo de análise possa ser feita, seja por parte do governo, da academia, de pesquisadores, jornalistas ou mesmo cidadãos. O tema dessas reportagens é importante porque analisa se milhões de dólares de dinheiro público estão sendo gastos corretamente.

— Steve Doig, Walter Cronkite School of Journalism, Arizona State University

Care Home Crisis: A crise da empresas de saúde em domicílio

Uma investigação do Financial Times sobre o mercado de serviços de saúde em casa (home care) expôs como algumas empresas tornaram o cuidado de idosos uma máquina de lucro e destacou os custos humanos de um modelo de negócios que favoreceu o retorno do investimento em vez de bons cuidados.

A análise foi oportuna, pois os problemas financeiros da empresa Southern Cross, então a maior operadora de home care do país, estavam chegando a um estágio crítico. Há décadas o governo promoveu uma privatização no setor de cuidadores e continuou a atrair o setor privado para práticas astutas de negócios.

Nossa investigação começou com a análise de dados obtidos a partir do órgão regulador do Reino Unido responsável por fiscalizar serviços de saúde em domicílio. A informação era de utilidade pública, mas exigiu muita persistência para ser obtida em uma forma utilizável.

Os dados incluíram avaliações (agora extintas) sobre o desempenho dos serviços em domicílios e também se eles eram privados, estatais ou sem fins lucrativos. A Comissão de Qualidade da Assistência, até junho de 2010, avaliou cuidados domiciliares em nível de qualidade (que iam de 0 estrelas = ruim a 3 estrelas = excelente).

O primeiro passo necessário foi um grande tratamento de dados, pois aqueles dados continham categorias não-uniformes. Isso foi feito usando principalmente o Excel. Nós também determinamos — por meio de pesquisas secundárias ou por telefone — se determinados serviços domiciliares haviam sido adquiridos por meio de grupos de private-equity. Antes da crise financeira, o setor de home care era um ímã para private equity e investidores imobiliários, mas vários - como Southern Cross - começaram a enfrentar sérias dificuldades financeiras. Queríamos estabelecer se havia algum efeito no fato de uma empresa ser ligada a um fundo de private equity (que normalmente financia empresas em fase de expansão de forma agressiva).

Um conjunto relativamente simples de cálculos do Excel permitiu-nos estabelecer que os cuidadores sem fins lucrativos e geridos pelo governo tinham, em média, um desempenho significativamente melhor do que os do setor privado. Alguns grupos de private-equity de home care mostravam um desempenho acima da média, e outros bem abaixo da média.

Junto com a reportagem de campo, os estudos de casos de negligência jogaram um olhar mais profundo sobre falhas nas políticas de regulação, bem como mostraram outros dados sobre os níveis de remuneração, rotatividade, etc., e nossa análise foi capaz de evidenciar a verdadeira situação de cuidado ao idoso.

Algumas dicas:

- Certifique-se de manter suas anotações de como manipulou os dados originais.
- Mantenha uma cópia dos dados originais e nunca mude-os.
- Faça a checagem e rechecagem de seus dados. Faça a análise muitas vezes (e se precisar, desde o início).
- Se você mencionar empresas particulares ou pessoas, ofereça a eles a oportunidade de resposta.

— Cynthia O'Murchu, Financial Times

O telefone conta tudo

A compreensão da maioria das pessoas sobre o que pode realmente ser feito com os dados fornecidos pelos celulares é teórica; há poucos exemplos no mundo real. É por isso que Malte Spitz, do Partido Verde alemão, decidiu publicar seus próprios dados. Para acessar as informações, ele teve que abrir um processo contra a gigante das telecomunicações alemã Deutsche Telekom. Os dados, contidos em um gigantesco documento de Excel, foram a base para o mapa interativo publicado no Zeit Online. Cada uma das 35.831 linhas da planilha representa uma ocasião na qual o celular de Sptiz transferiu informações. O período de todos esses eventos foi de apenas seis meses.

Vistos individualmente, os dados são, na maioria das vezes, inofensivos. Mas se tomados em conjunto, podem fornecer o que investigadores chamam de perfil: uma clara imagem dos hábitos e preferências do indivíduo e, de fato, de sua vida. Este perfil revela quando Spitz andou pelas ruas, quando pegou um trem, quando estava em um avião. Os dados mostram que ele trabalha principalmente em Berlim e quais cidades ele visitou. Mostra ainda quando ele acordou e quando dormiu.

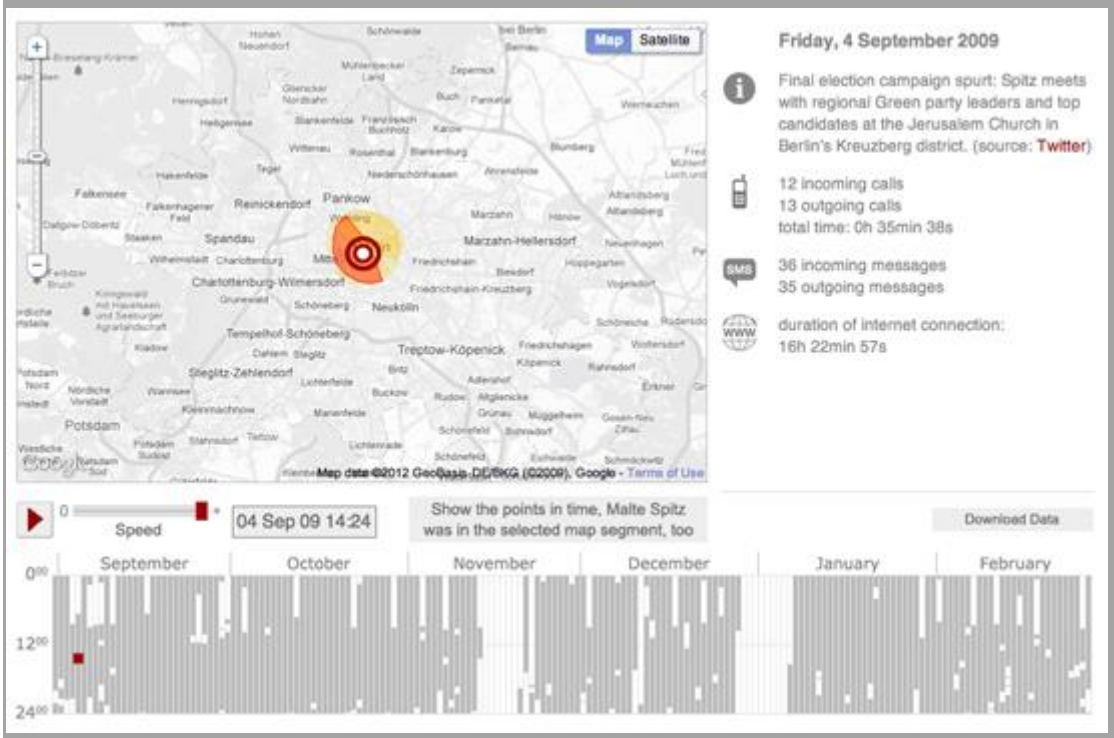


Imagem 23. O telefone conta-tudo (Zeit Online)

A base de dados da Deutsche Telekom manteve privada parte dos dados de Spitz: para quem ele ligou e quem ligou para ele. Este tipo de informação não só

infringiria a privacidade de várias outras pessoas, como também iria—mesmo se os números estivessem criptografados — revelar muito mais que o necessário sobre Spitz (mas agentes governamentais do mundo real teriam acesso a essa informação).

Pedimos a Lorenz Matzat e Michael Kreil, do OpenDataCity, que explorassem os dados e buscassem uma solução para a apresentação visual. "Primeiramente, usamos ferramentas como o Excel e o Fusion Tables para entender os dados; em seguida, desenvolvemos uma interface de mapa para permitir à audiência interagir com as informações de uma maneira não linear", disse Matzat. Para ilustrar quantos detalhes da vida de alguém podem ser obtidos por meio destes dados armazenados, a pesquisa foi ampliada com dados públicos sobre suas atividades (Twitter, registro em blogs, informação sobre partido político, entre outros). Este é o tipo de processo que qualquer bom investigador iria provavelmente seguir para traçar o perfil de uma pessoa que estivesse sob observação. Junto com a equipe de infográficos do Zeit Online, o time de pesquisa e desenvolvimento finalizou uma ótima interface de navegação: pressionando o botão "play", o usuário embarca em uma viagem pela vida de Malte Spitz.

Após o lançamento bem-sucedido do projeto na Alemanha, notamos que tínhamos um tráfego muito grande de acessos de fora do país, e então decidimos criar uma versão em inglês do aplicativo. Depois de ganhar o Germany Grimme Online Award, o projeto foi honrado com o Prêmio da ONA (Online News Association - Associação de Jornais Online) em setembro de 2011, sendo a primeira vez que isso ocorria com um site de notícias alemão.

Todos os dados estão disponíveis nesta planilha do Google Docs. Leia a reportagem no Zeit Online.

— Sascha Venohr, Zeit Online

Quais modelos se saem pior na inspeção veicular britânica?

Em janeiro de 2010, a BBC obteve as taxas de aprovação e reprovação da inspeção veicular do Ministério do Transporte para diferentes marcas e modelos de carros. Este teste atesta se um carro é seguro e se possui condições de trafegar pelas ruas; todo carro com mais de três anos deve passar pela verificação anual.

Obtivemos os dados por meio da lei de acesso à informação após uma longa batalha com a VOSA, a agência do Departamento de Transporte britânico que supervisiona a inspeção. A VOSA recusou nosso pedido para acesso a esses dados sob o argumento de que violaria a confidencialidade comercial. O órgão sustentou que isso poderia causar "danos comerciais" às fabricantes de veículos com alta taxa de reprovação. Apelamos ao Comissário de informação, que determinou a abertura dos dados em nome do interesse público. Só assim a VOSA divulgou os dados, 18 meses após a solicitação.

Analisamos os números com foco nos modelos mais populares e comparando carros da mesma idade. Isso apontou grandes discrepâncias. Por exemplo, entre carros de três anos, 28% dos Renault Mégane foram reprovados, em contraste com apenas 11% dos Toyota Corolla. Os dados foram divulgados na televisão, no rádio e na internet.

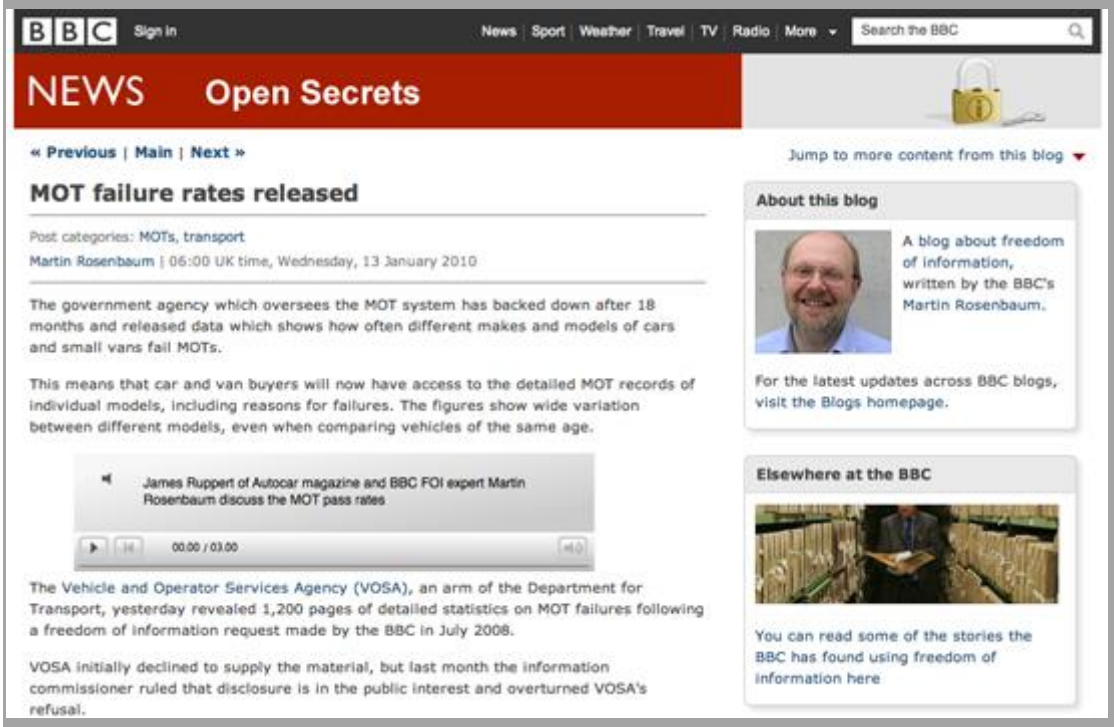


Imagem 24. Publicação das taxas de reprovação na inspeção veicular (BBC)

Os dados nos foram entregues em um documento PDF de 1,2 mil páginas, que tivemos que converter em uma planilha para análise. Além das nossas conclusões, publicamos o arquivo de Excel (com mais de 14 mil linhas de dados) no site BBC News junto com nossa reportagem. Isso permitiu que todos acessassem os dados em um formato mais simples.

O resultado foi que outras pessoas começaram a usar esses dados para suas próprias análises, as quais não tivemos tempo de fazer em função da pressa para publicar rapidamente a reportagem (algumas delas, na verdade, superaram nossas capacidades técnicas naquele momento). Isso incluiu a verificação dos índices de reprovação de carros com outras idades, comparando registros de fabricantes, e a criação de bases de dados para consulta por modelos individuais. Acrescentamos links para esses sites em nossa matéria, de modo que leitores pudessem conhecer os outros trabalhos.

Isso mostrou algumas vantagens de divulgar dados brutos para numa reportagem baseada em dados. Pode haver exceções (por exemplo, se você planeja usar os dados para reportagens posteriores e quer guardá-los enquanto isso), mas publicar as informações tem vários benefícios importantes:

- Seu trabalho é descobrir coisas e contá-las ao público. Se você se deu o trabalho de obter os dados, deve também divulgá-los.
- Outras pessoas podem descobrir pontos de interesse significativo que você não viu, ou simplesmente detalhes que sejam mais importantes para elas ainda que não tenham relevância para a sua reportagem.
- Outros podem se basear em seu trabalho para desenvolver uma análise mais detalhada, ou usar técnicas diferentes para apresentar ou visualizar os números, usando ideias ou capacidades próprias que podem investigar os dados de outras maneiras.
- É parte da incorporação de responsabilidade e de transparência ao processo jornalístico. Outros podem entender seus métodos e verificar seu trabalho, se desejarem.

— Martin Rosenbaum, BBC

Subsídios de ônibus na Argentina

Desde 2002, os subsídios para ônibus no sistema de transporte público da Argentina têm crescido exponencialmente, batendo recordes a cada ano. Mas em 2011, após vencer as eleições, o governo federal recém-eleito anunciou corte nos subsídios. Ao mesmo tempo, decidiu transferir a administração de linhas de ônibus e de metrô locais para a Prefeitura de Buenos Aires. Como a transferência dos subsídios para esse governo local não foi esclarecida e havia falta de verbas para garantir a segurança do sistema de transporte, a prefeitura da cidade de Buenos Aires rejeitou a decisão.

Enquanto isso acontecia, meus colegas do La Nación e eu nos encontrávamos pela primeira vez para discutir como começar nossa própria operação de jornalismo de dados. Nosso editor de Finanças sugeriu que os dados de subsídios publicados pela Secretaria de Transporte seriam um bom desafio inicial, pois era muito difícil tirar sentido daquilo em função do formato e da terminologia usados.

As condições precárias do sistema de transporte público atrapalham a vida de mais de 5,8 milhões de passageiros todos os dias. Atrasos, greves, panes de veículos ou até acidentes são frequentes. Decidimos investigar para onde vão os subsídios do sistema público de transporte na Argentina e tornar esses dados facilmente acessíveis para todo cidadão por meio de um "Explorador dos Subsídios de Transporte", que está atualmente em desenvolvimento.

Imagem 25. O Explorador dos Subsídios de Transporte (La Nación)

Começamos calculando quanto as empresas de ônibus recebem todos os meses do governo. Para isso, analisamos os dados publicados no site do Departamento de Transporte, onde foram publicados mais de 400 PDFs contendo relatórios mensais de pagamento para mais de 1.300 empresas desde 2006.

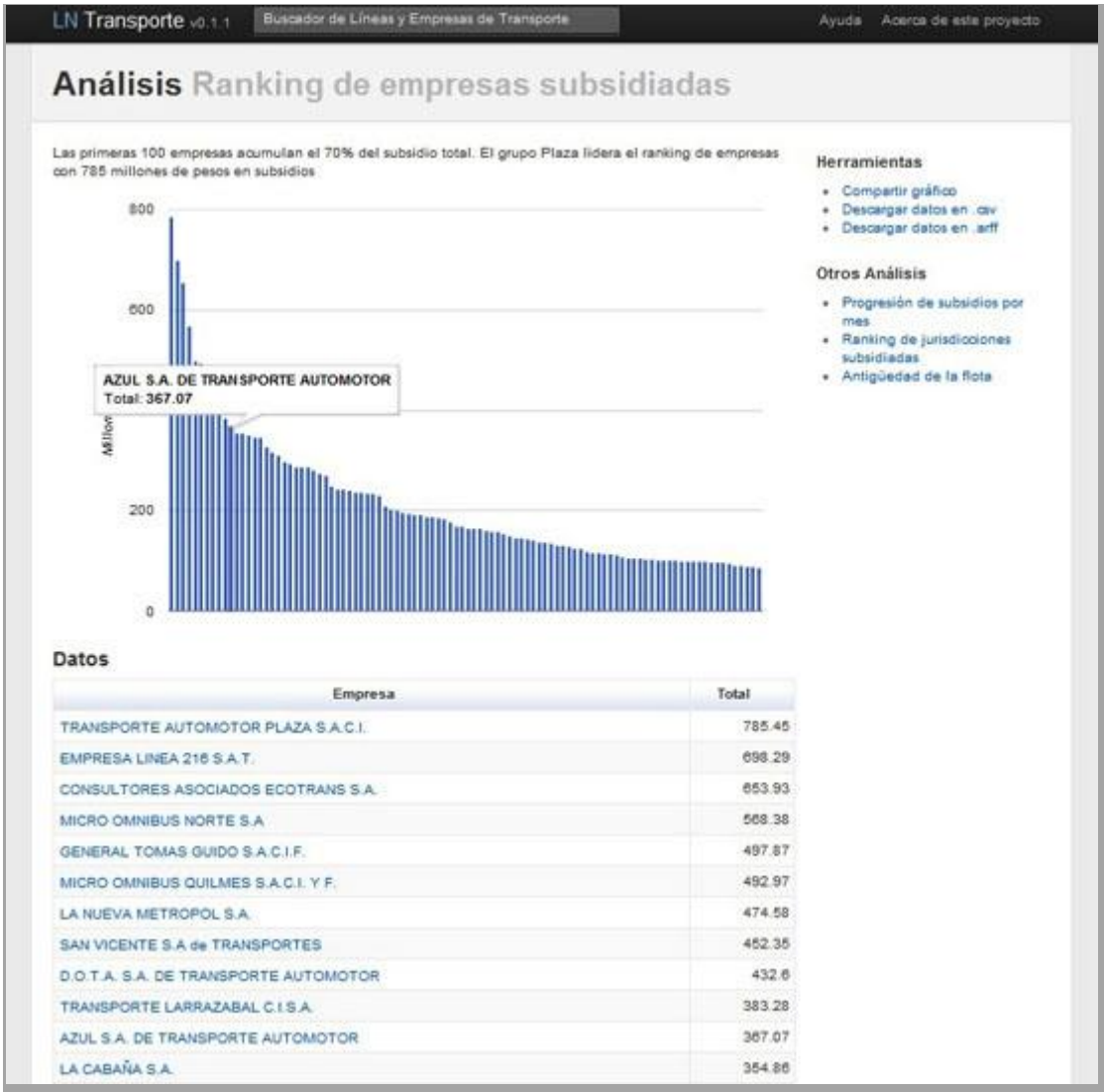


Imagem 26. Ranking de empresas de transporte subsidiadas (La Nación)

Trabalhamos com um programador sênior para desenvolver um software de extração de dados que automatizaria o download e a conversão dos arquivos PDFs do governo em arquivos de Excel e de banco de dados. Estamos usando a base de dados criada, com mais de 285 mil registros, para investigações e visualizações, tanto no impresso quanto online. Além disso, deixamos esses dados disponíveis em um formato interpretável por máquinas para qualquer argentino que quiser reusá-los e compartilhá-los.

O próximo passo foi identificar quanto custava em média a manutenção mensal dos veículos de transporte público. Fomos a outro site governamental, o da Comisión Nacional de Regulación del Transporte (CNRT, ou Comissão Nacional para Regulação do Transporte), responsável por regular o transporte na Argentina. Neste site, encontramos uma lista de empresas que detinham

juntas 9.000 veículos. Desenvolvemos um algoritmo que nos permitiu conciliar os nomes das empresas de ônibus e cruzar os dois conjuntos de dados.

Para avançar, precisávamos do número de registro de cada veículo. No site da CNRT, encontramos uma lista de ônibus por linha, por companhia, e com suas respectivas placas. As placas na Argentina são compostas de letras e números que correspondem à sua idade. Por exemplo, meu carro tem o número IDF234, onde o "I" corresponde ao período Março-Abril de 2011. Fizemos uma engenharia reversa das placas pertencentes a todas as companhias para saber a idade média da frota de cada uma. O objetivo foi mostrar quanto dinheiro vai para cada empresa e comparar os montantes tendo como base a idade de seus veículos.

No meio deste processo, o conteúdo dos PDFs divulgados pelo governo contendo os dados que precisávamos misteriosamente mudou, apesar das URLs e nomes dos arquivos continuarem os mesmos. Alguns PDFs agora estavam sem a coluna "totais", o que torna impossível cruzar os totais do período investigado completo, 2002-2011.

Levamos este caso para uma hackatona organizada pelo Hack/Hackers em Boston, onde o programador Matt Perry generosamente criou o que chamamos de "PDF Spy" ("Espião de PDF"). Este aplicativo ganhou prêmio da categoria "Mais Intrigante" daquele evento. OPDF Spy aponta para uma página cheia de PDFs e verifica se o conteúdo dentro dos PDFs foi alterado. "Nunca se deixe enganar pela *transparência do governo* novamente", escreve Matt Perry.

Imagem 27. Comparação da idade da frota ao montante de dinheiro que as empresas recebem do governo (La Nación)

Quem trabalhou no projeto?

Uma equipe de 7 jornalistas, programadores e um designer de interação trabalhou nesta investigação por 13 meses.

As habilidades necessárias para este projeto foram:

- Jornalistas com conhecimento sobre o funcionamento dos subsídios para o sistema público de transporte e quais os riscos envolvidos; conhecimento sobre o mercado de empresas de ônibus.
- Um programador com habilidade em extração de dados (scraping), análise e normalização de informações, e capaz ainda de converter PDFs em planilhas de Excel.
- Um estatístico para conduzir a análise de dados e os diferentes cálculos.
- Um designer para produzir visualizações de dados interativas.

Quais ferramentas usamos?

Usamos VBasic for applications, macros de Excel, Tableau Public, e a Junar Open Data Plataform, além de Ruby on Rails, a API de gráficos do Google, e Mysql para o Explorador de Subsídios.

O projeto teve grande impacto. Tivemos dezenas de milhares de exibições no site e a investigação ganhou destaque na primeira página da versão impressa do La Nación.

O sucesso desse primeiro projeto de jornalismo de dados nos ajudou internamente a montar uma operação de dados para reportagens investigativas e prestar serviço ao público. Isto resultou no Data.lanacion.com.ar, uma plataforma onde publicamos dados de vários assuntos de interesse público em formato interpretável por máquina.

— Angélica Peralta Ramos, La Nación (Argentina)

Jornalistas de dados cidadãos

As grandes redações não são as únicas que podem trabalhar em histórias baseadas em dados. As mesmas habilidades que são úteis para o jornalista de dados também podem ajudar repórteres cidadãos a acessar informações sobre a região onde vivem e transformá-las em matérias.

Essa foi a principal motivação do projeto de mídia cidadã Amigos de Januária, apoiado pelaRising Voices, da Global Voices Online, e pela organização Artigo 19. Entre setembro e outubro de 2011, um grupo de jovens moradores da cidade de Januária, no norte de Minas Gerais, uma das regiões mais pobres do Brasil, teve aulas sobre técnicas básicas de jornalismo e monitoramento do orçamento público municipal. Eles também aprenderam como preencher formulários de pedidos de acesso à informação e como acessar bases de dados oficiais na internet.

Januária, uma cidade com cerca de 65 mil habitantes, é conhecida também pelo fracasso de seus políticos. Ao longo de três mandatos municipais, teve sete prefeitos diferentes. A maior parte foi removida do cargo devido a denúncias que apontavam má condução da administração municipal, incluindo envolvimento em casos de corrupção.

Cidades pequenas como Januária não atraem a atenção da mídia, que tende a se focar em capitais e outros municípios de maior porte. No entanto, existe espaço para que os moradores dessas localidades ajudem a monitorar a administração pública, já que conhecem os problemas enfrentados pela sua comunidade melhor do que ninguém. Tendo a internet como uma importante aliada, eles podem acessar de forma mais fácil e rápida informações como orçamento municipal e outros dados locais.

Imagem 28. O projeto de jornalismo cidadão "Amigos de Januária" ensina habilidades fundamentais para transformar cidadãos em jornalistas de dados

Depois de participar de doze aulas, alguns dos repórteres cidadãos de Januária começaram a acessar dados públicos sobre a cidade e a produzir matérias. Soraia Amorim, por exemplo, uma jornalista cidadã de 22 anos, descobriu que o número oficial de médicos que constava na folha de pagamento do município divergia da realidade na área da saúde que ela conhecia. Para escrever sua matéria, Soraia acessou dados de saúde disponíveis online no site do Sistema Único de Saúde (SUS), que mostravam que Januária deveria ter 71 médicos, em diversas especialidades.

No entanto, esse número não correspondia com o que Soraia sabia sobre a disponibilidade desses profissionais na cidade. Os moradores estavam sempre reclamando sobre a falta de médicos na rede pública e alguns precisavam viajar para cidades vizinhas para serem atendidos. Soraia então entrevistou uma mulher que tinha sofrido um acidente de moto recentemente e não encontrou assistência no hospital de Januária, porque não havia nenhum médico disponível. A repórter cidadã também falou com a Secretaria Municipal de Saúde, que admitiu que havia menos médicos na cidade do que o número da base de dados do SUS.

Essas descobertas iniciais levantam muitas questões sobre as possíveis razões para a divergência entre os dados e a realidade de Januária. Uma delas é que os dados do SUS estão errados, o que poderia indicar que há um problema na qualidade das informações de saúde do Brasil. Outra é que Januária estaria informando dados errados para o SUS. Ambas as hipóteses precisariam de uma apuração mais aprofundada. No entanto, a matéria de Soraia é uma importante parte dessa cadeia, já que ilumina uma inconsistência e pode encorajar outras pessoas na cidade a investigar mais o caso.

"Eu costumava viver na zona rural e terminei o Ensino Médio com dificuldade", diz Soraia. "Quando as pessoas me perguntavam o que eu queria ser, eu sempre dizia que queria ser jornalista. Mas eu imaginava que era praticamente impossível devido ao mundo onde eu vivia". Depois de participar do projeto Amigos de Januária, Soraia acredita que o acesso a dados públicos é uma importante ferramenta para mudar a realidade da sua cidade. "Eu me sinto capaz de ajudar a mudar minha cidade, meu país, o mundo", conta, animada.

Alysson Montiériton, de 20 anos, é outro jornalista cidadão que participou do projeto e usou dados públicos para produzir uma matéria. Na primeira aula do projeto, quando os repórteres cidadãos foram para as ruas da cidade para procurar por assuntos que poderiam se transformar em matérias, Alysson decidiu escrever sobre um semáforo quebrado. Localizado em um cruzamento importante de Januária, ele estava quebrado desde o começo daquele ano. Depois de aprender como procurar dados na internet, o jovem repórter buscou o número de veículos que existem em Januária e o valor pago em impostos por quem tem carro. Na sua matéria, escreveu:

"A situação em Januária fica pior por causa ao grande número de veículos na cidade. De acordo com o IBGE, Januária tinha 13.771 veículos (entre os quais 7.979 eram motos) em 2010. (...) Os moradores da cidade acreditam que o atraso em arrumar o semáforo não é resultado da falta de recursos. De acordo com a Secretaria do Tesouro de Minas Gerais, a cidade recebeu R$ 470 mil em taxas de veículos em 2010."

Ao ter acesso aos dados, Alysson pôde mostrar que Januária tinha muitos veículos (quase um para cada cinco habitantes) e que um semáforo quebrado em um cruzamento movimentado poderia colocar muitas pessoas em perigo. Além disso, ele pode revelar o volume de recursos recebidos pela cidade em pagamento de impostos pelos proprietários de automóveis e, baseado nessa

informação, questionar se o dinheiro não seria suficiente para consertar o semáforo, oferecendo mais segurança para motoristas e pedestres.

Apesar das histórias escritas por Soraia e Alysson serem muito simples, elas mostram que os dados também podem ser usados por repórteres cidadãos. Não é preciso estar em uma grande redação e ser cercado de especialistas para usar dados em matérias jornalísticas. Depois de apenas doze aulas, Soraia e Alysson, nenhum deles com treinamento anterior em jornalismo, puderam trabalhar em matérias baseadas em dados e escrever textos interessantes sobre a realidade local de Januária. Além disso, as duas matérias mostram que os dados podem ser úteis inclusive em pequena escala. Mostram que também existem informações valiosas em pequenas bases de dados, não apenas nas gigantescas.

— Amanda Rossi, Amigos de Januária

O Grande Quadro com o Resultado das Eleições

Resultados de eleições são excelentes oportunidades, para qualquer veículo de imprensa, de se contar histórias de forma visual. Por muitos anos deixamos passar essa oportunidade, mas, em 2008, decidimos mudar isso junto com a editoria de infografia.

Queríamos mostrar os resultados de maneira a contar uma história, mas sem que parecesse apenas um amontoado de números em uma tabela ou em um mapa. Nas eleições anteriores, foi exatamente o que fizemos.

Não que haja algo errado com um grande apanhado de números, ou o que chamo de "estilo CNN" de tabelas, tabelas e mais tabelas. Isso funciona porque dá ao leitor exatamente aquilo que ele quer saber: quem ganhou.

E o perigo de estragar algo que não está propriamente errado é significativo. Ao criarmos algo radicalmente diferente e nos afastarmos do que as pessoas normalmente esperam, poderíamos tornar as coisas mais confusas, ao invés de simplificar.

No fim, Shan Carter, da editoria de infografia, trouxe a resposta exata, o que acabamos por chamar de "o grande quadro". Quando vi os primeiros esboços, foi um desses momentos de literalmente se levar as mãos à cabeça.

Era exatamente o que precisávamos.

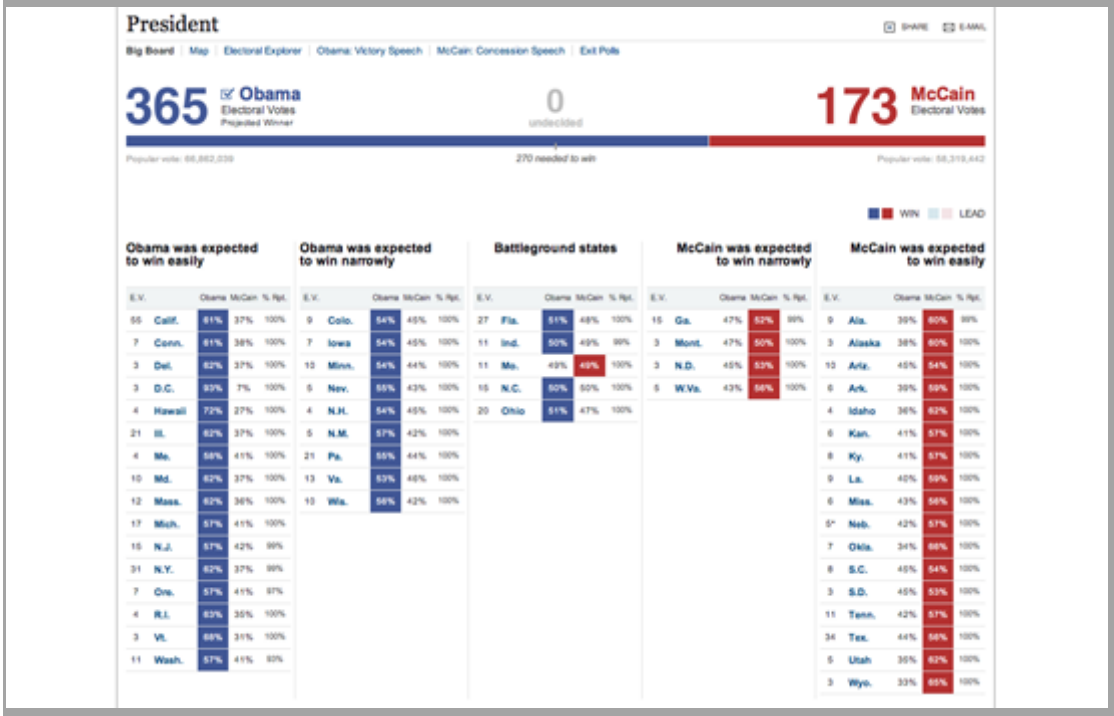


Imagem 29. O grande quadro com os resultados das eleições (New York Times)

O que faz disso uma bela peça de jornalismo visual? Para começar, os olhos do leitor são logo atraídos para a grande barra que mostra no alto os votos do colégio eleitoral, o que podemos chamar no contexto jornalístico de *lide*. Conta exatamente o que o leitor quer saber, e o faz rapidamente, com simplicidade e sem nenhum ruído visual.

Em seguida, o leitor é conduzido ao agrupamento dos estados americanos em cinco colunas diferentes, divididos de acordo com a avaliação do New York Times de quão inclinado um estado estava por um ou por outro candidato. E justamente na coluna central vem o que chamaríamos no jargão jornalístico de *olho gráfico*, onde explicamos por que Obama ganhou. A peça interativa torna o fato cristalino: Obama venceu em todos os estados onde sua vitória era esperada e em quatro dos estados indecisos.

Para mim, essa arquitetura com cinco colunas é um exemplo de como o jornalismo visual difere de outras formas de design. Idealmente, uma peça memorável de jornalismo visual será ao mesmo tempo bela e informativa. Mas ao optar entre a notícia ou a estética, o jornalismo deve pender para o lado da história. E enquanto esse layout pode não ser aquele que um designer purista escolheria para apresentar esses dados, ele entrega a notícia muito, muito bem.

E, por fim, como qualquer ferramenta interativa na web, ela convida o leitor a aprofundar a leitura. Há detalhes como porcentagens de votação em cada estado e o número de votos no colégio eleitoral, enquanto as porcentagens são deliberadamente exibidas com menos destaque, para não competir com os pontos principais da história.

Tudo isso faz do "grande quadro" um bela peça de jornalismo visual que delineia com perfeição a velha e boa pirâmide invertida.

— Aron Pilhofer, New York Times

Apurando o preço da água via crowdsourcing

Desde março de 2011, informações sobre a tarifa da água em toda a França são reunidas por meio de uma experiência de crowdsourcing. Em apenas 4 meses, mais de 5 mil pessoas indignadas com o controle corporativo do mercado de recursos hídricos tomaram o tempo de verificar sua conta de água, digitalizá-la e enviá-la ao projeto Prix de l'Eau (Preço da Água). O resultado é uma investigação sem precedentes que congregou geeks, ONGs e a mídia tradicional para ampliar a transparência sobre o abastecimento de água.

Imagem 21. O Preço da Água (Fondation France Liberté)

O mercado de abastecimento de água francês é formado por 10 mil clientes (cidades que compram água para distribuir aos contribuintes) e um punhado de companhias prestadoras do serviço. O equilíbrio de forças neste oligopólio é distorcido em favor das corporações, que algumas vezes cobram preços diferentes de cidades vizinhas!

A ONG francesa France Libertés tem lidado com questões relacionadas aos recursos hídricos no mundo inteiro nos últimos 25 anos. Agora, a entidade se foca em aprimorar a transparência do mercado francês e em colaborar com cidadãos e prefeitos, que negociam os acordos de abastecimento. O governo francês decidiu enfrentar o problema dois anos atrás, com um censo nacional do preço e qualidade da água. Até agora, apenas 3% dos dados necessários foram coletados. Para acelerar o processo, France Libertés resolveu envolver diretamente os cidadãos.

Em conjunto com a equipe OWNI, eu desenvolvi uma interface de crowdsourcing na qual os usuários podem incluir cópias digitalizadas de suas contas de água e inserir o preço pago no website prixdeleau.fr. Nos últimos quatro meses, 8,5 mil pessoas se inscreveram e mais de 5 mil contas foram enviadas e validadas.

Embora os resultados não permitam uma análise perfeita da situação do mercado, eles mostraram a partes interessadas, como as agências de supervisão de recursos hídricos, que havia uma preocupação popular genuína com o preço da água. Num primeiro momento, eles estavam céticos quanto à questão da transparência, mas, ao longo da operação, foram se juntando à France Libertés em sua luta contra a obscuridade e as más práticas comerciais. O que a imprensa pode aprender com isso?

Parcerias com ONGs

As ONGs demandam grandes volumes de dados para o desenvolvimento de estudos que subsidiem suas políticas. Essas entidades muitas vezes estão mais dispostas financiar uma coleta de dados do que um executivo da área de jornalismo.

Usuários podem oferecer dados brutos

O crowdsourcing funciona melhor quando os usuários realizam tarefas de coleta ou limpeza de dados.

Peça a fonte

Nós refletimos sobre a necessidade de pedir aos usuários uma cópia da conta original, pensando que isso poderia afastar alguns deles (especialmente porque nossa audiência era mais idosa do que a média). Ainda que pedir a conta original possa ter feito com que alguns desistissem, os dados ganharam mais credibilidade.

Crie um mecanismo de validação

Nós criamos um sistema de pontuação e um mecanismo de revisão por pares para verificar as contribuições. Isso se mostrou complicado demais para os usuários, que tinham poucos incentivos para realizar visitas frequentes ao website. O sistema, todavia, foi usado pela equipe da France Libertés, da qual cerca de 10 funcionários se motivaram a trabalhar com o sistema de pontos.

Seja simples

Nós construímos um mecanismo de envio automático de mensagens, para que os usuários pudessem solicitar dados sobre o preço da água pela Lei de Acesso à Informação com alguns poucos cliques. Apesar de inovadora e bem planejada, essa funcionalidade não gerou um grande retorno (apenas 100 requisições foram enviadas).

Mire na sua audiência

A France Libertés se associou à revista especializada em direito do consumidor *60 Millions de Consommateurs*, que incentivou muito sua comunidade a se envolver. Foi o par perfeito para uma operação como essa.

Escolha com cuidado seus indicadores de sucesso

O projeto angariou apenas 45 mil visitantes em quatro meses, o equivalente a 15 minutos de tráfego no nytimes.com. O que realmente importa é que um em cada cinco se inscreveram e um em cada dez se deram o trabalho de digitalizar e enviar sua conta de água.

— Nicolas Kayser-Bril, Journalism++

Coletando dados

Então você está pronto para começar o seu primeiro projeto de jornalismo de dados. E agora? Primeiro você precisa de alguns dados. Esta seção mostra onde encontrá-los na web, como solicitá-los usando as leis de acesso à informação, como usar a técnica de scraping para extrai-los de fontes não estruturadas e como usar crowdsourcing para montar suas próprias bases de dados com a ajuda dos leitores. Por fim, falaremos sobre o que a lei diz a respeito da reprodução de bases de dados de terceiros e como usar ferramentas simples para permitir que outros republiquem as informações.

O que há neste capítulo?

Guia rápido para o trabalho de campo

Procurando dados sobre um assunto ou área em particular? Não tem certeza se existem ou onde encontrá-los? Não sabe por onde começar? Nesta seção vamos ver como iniciar a busca por dados públicos em fontes da web.

Tornando sua busca mais eficiente

Apesar de nem sempre serem fáceis de serem achadas, muitas bases de dados na web são indexadas por mecanismos de busca, intencionalmente ou não. Algumas dicas:

- Quando estiver buscando dados, não esqueça de incluir tanto termos de busca relativos ao conteúdo quanto ao formato ou à fonte onde espera encontrá-los. O Google e outros buscadores permitem pesquisar por formato de arquivo. É possível buscar, por exemplo, apenas planilhas (inserindo "filetype:XLS filetype:CSV"), dados geocodificados ("filetype:shp"), ou bancos de dados ("filetype:MDB, filetype:SQL, filetype:DB"). Você pode até mesmo procurar por arquivos PDF ("filetype:pdf").
- Também é possível pesquisar pela parte de uma URL. Ao inserir "inurl:downloads filetype:xls", o Google tentará buscar todos os arquivos Excel que têm "downloads" em seu endereço (se encontrar um download, vale a pena checar por outros resultados na mesma pasta daquele servidor). Também é possível limitar a busca a resultados em apenas um domínio, colocando "site:agency.gov", por exemplo.
- Outra dica é não buscar o conteúdo diretamente, mas sim os lugares em que podem estar disponíveis dados em massa. Por exemplo, "site:agency.gov Directory Listing" pode retornar várias listas geradas automaticamente pelo servidor com acesso fácil aos dados brutos, enquanto "site:agency.gov Database Download" buscará apenas aquelas listas criadas intencionalmente para serem encontradas.

Indo direto à fonte

A primeira dica ao buscar dados de instituições públicas é tentar ir direto a quem detém os dados. Claro que se pode também fazer uma solicitação usando a lei de acesso à informação, mas o processo demora. É provável que você receba uma resposta de que os dados não estão no formato que você pediu, ou

que o órgão público usa um software proprietário que não permite a extração dos dados no formato solicitado. Mas, se consigo chegar à pessoa que cuida dos dados naquela instituição, posso questioná-la sobre as informações que ela têm e em que formato. Posso descobrir antes o que preciso fazer para solicitar as informações e ser bem sucedido. Os obstáculos dessa abordagem? Frequentemente, é difícil chegar a essas pessoas, pois os assessores de imprensa vão querer tomar a frente nesse contato. Nesses casos, o melhor é tentar marcar uma ligação em conferência ou, até melhor, um encontro cara a cara entre o assessor, o guru dos dados e você. Dá pra fazer isso de forma que seja difícil para eles dizer não. Diga que não quer dar mais trabalho a eles. Algo como "não quero criar um transtorno ou enviar um pedido muito abrangente, e uma reunião me ajudaria a entender qual a melhor forma de conseguir o que preciso."

Se esse método não funcionar, a alternativa é perguntar primeiro que layout de informações (record layout) e dicionário de dados (documento que mostra uma espécie de índice de dados) eles usam para, só então, fazer o pedido. Algumas vezes também pergunto como eles armazenam os dados e em qual sistema. Dessa forma, posso pesquisar de que maneira as informações podem ser exportadas antes de fazer a solicitação.

Para encerrar, minha história de maior sucesso aconteceu quanto trabalhava para um pequeno jornal em Montana. Precisava de dados sobre o condado local e fui informado que eles não poderiam ser exportados do servidor. Pesquisei um pouco, e me ofereci para ir até lá e ajudar. Trabalhei diretamente com o responsável pelos dados, escrevemos um pequeno script, e gravamos as informações em um disquete (isso foi há bastante tempo). Eu tinha meus dados e o condado está agora apto a fornecê-los a quem solicite. Eles também precisavam extrair os dados de vez em quando para uso próprio mas não entendiam completamente o sistema, então foi bom para ambos.

— Cheryl Philips, The Seattle Times

Navegando em sites e serviços de dados

Nos últimos anos, vários portais, hubs e outros sites especificamente dedicados a dados apareceram na web. São bons locais para se familiarizar com os diferentes formatos que existem por aí. Se você é principiante, deve dar uma olhada em:

Imagem 1. datacatalogs.org (Open Knowledge Foundation)

Portais de dados públicos oficiais

A disposição do governo em divulgar bases de dados varia de país para país. Um volume crescente de países está lançando portais de dados (inspirados no norte-americano data.gov e no britânico data.gov.uk) para promover o uso comercial e cívico das informações. Um índice global atualizado desses portais pode ser encontrado emdatacatalogs.org.

The Data Hub

Site coletivo administrado pela Open Knowledge Foundation que torna mais fácil procurar, compartilhar e reutilizar fontes abertas, especialmente de maneiras automatizadas.

ScraperWiki

Ferramenta online para "facilitar a extração de pedaços úteis de dados, de maneira que possam ser reutilizados por outros aplicativos, ou vasculhados por jornalistas e pesquisadores". A maioria dos "scrapers" (códigos para extrair dados específicos de um site) e suas bases de dados são públicos e podem ser reutilizados.

Portais de dados do Banco Mundial e das Nações Unidas

Fornecem indicadores confiáveis de todos os países, frequentemente com histórico de vários anos.

Infochimps e DataMarket

Startups com comunidades em torno do compartilhamento e venda de dados.

Freebase

Iniciativa ligada ao Google que fornece "uma base de dados com curadoria coletiva de pessoas, lugares e coisas."

Dados de pesquisas

Existem vários agregadores nacionais e temáticos de dados de pesquisas, como o UK Data Archive. Muitas bases têm acesso gratuito, mas outras exigem assinatura, ou não podem ser reutilizadas ou redistribuídas sem permissão.

Acessando dados de arquivos impressos

Logo após a divulgação pelo Wikileaks dos documentos das forças armadas dos Estados Unidos sobre as guerras do Afeganistão e Iraque, decidimos usar esse conceito para celebrar o 50º aniversário da Guerra da Argélia publicando o Algerian War Diaries. Digitalizamos os documentos do exército francês na Argélia, que estão disponíveis no arquivo do Ministério da Guerra em Paris, mas em papel. Enviamos jornalistas e estudantes para fotografar os papeis. Tentamos escanear usando um scanner portátil Canon P-150, mas não funcionou porque os arquivos estavam grampeados.

No fim das contas, reunimos cerca de 10.000 páginas em poucas semanas. Rodamos um software de reconhecimento de texto (ABBYY FineReader), mas o resultado foi ruim. Além disso, o ministro negou arbitrariamente acesso aos documentos mais interessantes e proibiu a republicação de arquivos que podiam ser fotografados livremente no local, então decidimos que não valia o risco e suspendemos o projeto.

— Nicolas Kayser-Bril, Journalism++

Pergunte a um fórum

Pesquise respostas já publicadas ou faça uma pergunta em Get The Data ou Quora. GetTheData é um forum de perguntas e respostas em que você pode levantar questões como onde encontrar dados sobre um determinado tema, como consultar e obter uma fonte específica, que ferramentas de visualização usar, como limpar os dados, ou como consegui-los em um formato que dê para trabalhar.

Pergunte a uma lista de e-mail

Listas de e-mail combinam a sabedoria de toda uma comunidade sobre um determinado tópico. Para jornalistas de dados, as listas Data-Driven Journalism e NICAR-L são excelentes pontos de partida. Ambas estão cheias de geeks envolvidos em Reportagens com Auxílio de Computador (RAC). É provável que alguém já tenha trabalhado em uma reportagem como a sua, e tenha uma ideia de por onde começar, ou até mesmo os dados que está procurando. Você também pode tentar o Projeto Wombat, "uma lista de discussão para perguntas de referência difíceis", pesquisar as várias listas da Open Knowledge Foundation, no theInfo, ou fazer buscas pelo tópico que está interessado.

Entre para o Hacks/Hackers

Hacks/Hackers é uma organização internacional de cunho popular em franca expansão com dezenas de ramificações e milhares de membros. Sua missão é criar uma rede de jornalistas ("hacks") e aficionados por tecnologia ("hackers") que repensam o futuro da mídia e da informação. Com uma rede tão ampla, você tem grandes chances de encontrar alguém que saiba onde procurar a informação que você está correndo atrás.

Pergunte a um especialista

Professores, funcionários públicos, e pessoal da indústria normalmente sabem onde procurar. Ligue para eles. Mande um e-mail. Aborde-os em eventos. Apareça em seus escritórios. Peça com jeito. "Estou fazendo uma reportagem sobre X. Onde posso encontrá-lo? Sabe quem pode ter essa informação?"

Estude a Tecnologia da Informação usada pelo governo

É bom entender o contexto tecnológico e administrativo em que são mantidas as informações governamentais quando se está buscando alguma base de dados. Seja CORDIS, COINS ou THOMAS, os sistemas se tornam mais úteis na medida em que você entende um pouco o propósito para o qual foram criados.

Encontre os fluxogramas das organizações e procure por orgãos/unidades que tenham função interdepartamental (por exemplo:

Serviços de TI, comunicação), e explore seus sites. Muitos dados são armazenados ao mesmo tempo por vários departamentos e, enquanto uns os tratam como jóias da coroa, outros podem liberá-los tranquilamente.

Procure por infográficos dinâmicos nos sites governamentais. Frequentemente, funcionam a partir de bases de dados estruturadas/APIs que podem ser usadas de outras maneiras (por exemplo, tabelas de vôo, aplicativos de Java com a previsão do tempo).

Varrendo dados telefônicos

Há alguns meses, quis analisar os dados de ligações telefônicas do governador do Texas Rick Perry, então candidato à presidência. Era o resultado de uma longa espera após um pedido pelos registros. Os dados chegaram em 120 páginas impressas com a qualidade de um fax. Era uma empreitada que exigia a tabulação e a limpeza dos dados, seguida do cruzamento com o API das White Pages (equivalente norte-americano das Páginas Amarelas) para fazer uma busca a partir dos números de telefone.

Combinando os nomes com os dados eleitorais federais e estaduais, descobrimos que Perry ligou para doadores de campanha usando telefones do governo, uma prática mal vista que levantou dúvidas sobre suas ligações com um comitê de arrecadação independente.

— Jack Gillum, Associated Press

Procure de novo

Quando estiver mais informado sobre o assunto, procure novamente usando frases e combinações improváveis de palavras que você tenha encontrado desde a última busca. Você pode ter um pouco mais de sorte com os mecanismos de busca!

Faça um pedido pela Lei de Acesso à Informação

Se você acredita que um órgão governamental tem as informações que precisa, um pedido usando a Lei de Acesso à Informação pode ser a melhor ferramenta. Na próxima seção, você saberá como fazer para dar entrada em uma solicitação.

— Brian Boyer (Chicago Tribune), John Keefe (WNYC), Friedrich Lindenberg (Open Knowledge Foundation), Jane Park (Creative Commons), Chrys Wu (Hacks/Hackers)

Quando falha a lei

Depois de ler um artigo acadêmico explicando que a a publicação dos resultados de inspeções sanitárias em restaurantes reduziu o número de doenças relacionadas à comida em Los Angeles, pedi à vigilância sanitária parisiense a lista de inspeções. Seguindo o procedimento da Lei de Acesso à informação francesa, aguardei 30 dias por uma resposta negativa, e então recorri à comissão de acesso aos dados públicos (CADA, em francês), que legisla sobre a legitimidade dos pedidos feitos por meio da lei. A CADA aceitou meu pedido e ordenou que liberassem os dados. Responderam pedindo mais dois meses de prazo e a CADA aceitou. Dois meses depois, nada foi feito.

Tentei conseguir o apoio de conhecidos (e ricos) defensores da abertura de dados públicos para recorrer à Justiça (o que custaria 5.000 euros e era vitória certa com o apoio da CADA), mas eles ficaram com medo de comprometer suas relações com os programas oficiais de open data. Esse é apenas um exemplo, entre vários, de descaso do governo francês pela lei e em que programas oficiais não fazem nada para ajudar iniciativas populares de acesso aos dados.

— Nicolas Kayser-Bril, Journalism++

Seu Direito aos Dados

Antes de fazer uma solicitação por Lei de Acesso à informação, você deve checar para ver se os dados que está procurando já estão disponíveis — ou se já foram solicitados por outras pessoas. O capítulo anterior traz algumas sugestões sobre onde você pode procurar. Se isso não adiantou, veja algumas dicas que podem ser úteis para fazer a solicitação de maneira mais eficiente:

Planeje com antecedência para economizar tempo

Considere fazer uma solicitação formal sempre que precisar procurar informações. É melhor não esperar esgotar todas as outras possibilidades. Você vai economizar tempo se fizer a solicitação no início de sua pesquisa e se mantiver outras maneiras de investigação em paralelo. Conte com atrasos: às vezes, órgãos públicos demoram para processar as solicitações.

Verifique as regras sobre taxas

Antes de dar início ao pedido formal, verifique as se há tarifas cobradas para pedir ou receber informações. Dessa forma, se um funcionário público solicitar dinheiro, você saberá quais são os seus direitos. Lembre-se de dizer em sua solicitação que você prefere que a informação seja enviada em arquivos eletrônicos para evitar custos de cópia e envio.

Saiba os seus direitos

Descubra quais são os seus direitos antes de começar, assim você saberá o que as autoridades públicas estão ou não obrigadas a fazer. Por exemplo, grande parte das leis de acesso informação delimita um tempo para que as autoridades respondam a pedidos. Ao redor do mundo, a média estabelecida pela maioria das leis é de alguns dias a até um mês. Tenha certeza qual é o caso antes de realizar a solicitação e anote a data quando você realizá-la.

Os governos não são obrigados a processar dados para você, mas deveriam prover todas as informações que possuem. Se forem dados que eles precisam ter para realizar suas competências legais, certamente deveriam fornecê-las a você.

Diga que você conhece os seus direitos

Geralmente, a legislação não requisita que você mencione a lei de acesso à informação ou a lei de liberdade de informação, mas mencionar é

recomendado porque demonstra que você tem conhecimento dos seus direitos legais e provavelmente vai incentivar que seu requerimento seja atendido conforme a lei. Para solicitações à União Europeia, o melhor é mencionar especificamente a Regulamentação 1049/2001.

Seja simples

Em todos os países, é melhor começar com uma simples solicitação de informação e, assim que você conseguir o dado inicial, adicionar mais perguntas. Dessa maneira, você não corre o risco da instituição pública solicitar mais prazo alegando ser um "pedido complexo".

Mantenha o foco

Um pedido a um departamento da autoridade pública provavelmente será respondido mais rapidamente do que um que necessite de uma pesquisa por toda a instituição. Uma solicitação que envolva a consulta da instituição a terceiros (por exemplo, uma empresa privada que possa saber a resposta, ou outro governo que seja, de certa forma, afetado pela informação) pode demorar muito tempo. Seja persistente.

Pense dentro dos arquivos

Tente descobrir quais dados estão organizados. Por exemplo, se você conseguir uma cópia em branco do formulário que a polícia preenche após acidentes de trânsito, saberá quais informações eles mantêm ou não sobre acidentes de carro.

Seja específico

Antes de enviar a sua solicitação, reflita: ela está de alguma forma ambígua? Isso é particularmente importante se você está pensando em comparar dados de diferentes órgãos públicos. Por exemplo, se você pedir informações sobre "os três últimos anos", alguns órgãos vão enviar informações dos três últimos anos do calendário e, outros, dados dos três últimos anos fiscais, o que vai tornar impossível uma comparação direta. Se você decidir ocultar a sua solicitação real em uma mais genérica, deve fazer seu pedido de maneira mais ampla, para que inclua a informação que você quer, mas não tão vasta que a torne obscura ou que desencoraje a resposta. Pedidos claros e específicos tendem a conseguir respostas mais rápidas e melhores.

Envie vários pedidos

Se você não tem certeza para qual órgão direcionar seu pedido, não há nada que o impeça de fazer solicitações a dois, três ou mais órgãos ao mesmo tempo. Em alguns casos, cada um deles dará uma resposta diferente, o que pode, na verdade, ser útil ao fornecer uma ideia mais completa das informações disponíveis do assunto que você está apurando.

Faça solicitações internacionais

Cada vez mais, as solicitações podem ser feitas de maneira eletrônica, não importa onde você mora. Se você não vive no país onde quer fazer a solicitação, uma alternativa é enviar o pedido para a embaixada, que vai encaminhá-lo ao órgão público competente para respondê-lo. Primeiro, você precisará verificar com a embaixada se ela realiza esse tipo de ação — talvez a equipe não terá sido treinada sobre as questões de direito à informação e, se for o caso, é mais seguro enviar o pedido diretamente para o órgão público.

Faça um teste

Se você está pensando em enviar o mesmo pedido para várias autoridades públicas, comece enviando um rascunho do pedido para algumas delas como um exercício piloto. Isso vai demonstrar se você está utilizando a terminologia correta para obter o material que deseja e se obter respostas para as suas perguntas é algo possível. Então, caso seja necessário, você pode revisar o pedido antes de enviá-lo a outros órgãos.

Antecipe as exceções

Se você acha que podem haver exceções para o pedido que está fazendo, quando estiver preparando as perguntas, separe a questão possivelmente problemática das demais e envie dois pedidos separadamente. Assim, você evita que as outras questões não deixem de ser respondidas por conta de uma exceção.

Solicite acesso aos arquivos

Se você vive próximo de onde a informação está guardada (por exemplo, na capital onde os documentos são armazenados), também pode solicitar checar os documentos originais. Isso pode ser útil quando estiver pesquisando informações contidas em um grande número de documentos que você gostaria de dar uma olhada. Esse tipo de consulta

deve ser gratuita e deve ser agendada em um horário razoável e conveniente a você.

Mantenha uma cópia!

Faça a sua solicitação por escrito e mantenha uma cópia para que você possa, no futuro, comprovar que seu pedido foi enviado, caso precise apelar devido a uma ausência de resposta. Isso também fornecerá provas de que você fez a solicitação, caso você pretenda escrever uma reportagem a respeito do processo.

Torne público

Acelere o recebimento de respostas tornando público que você realizou uma solicitação: escrever ou divulgar uma reportagem contando que a solicitação foi enviada pode colocar alguma pressão na instituição pública para processar e responder o pedido. Você pode atualizar as informações assim que conseguir respostas — ou, se o seu deadline expirar e não houver respostas, você também pode fazer do descaso uma matéroa. Agir dessa maneira tem o benefício extra de ensinar aos funcionários públicos sobre o direito de acesso à informação e como funciona na prática.

Há também diversos excelentes serviços que você pode utilizar para realizar a sua solicitação e qualquer pedido posterior, disponíveis para consulta pública na internet, tais como What Do They Know? para órgãos do Reino Unido, Frag den Staat para órgãos alemães, e Ask the EU para instituições da União Europeia. O projeto Alaveteli está ajudando a prover serviços semelhantes para dezenas de países ao redor do mundo.

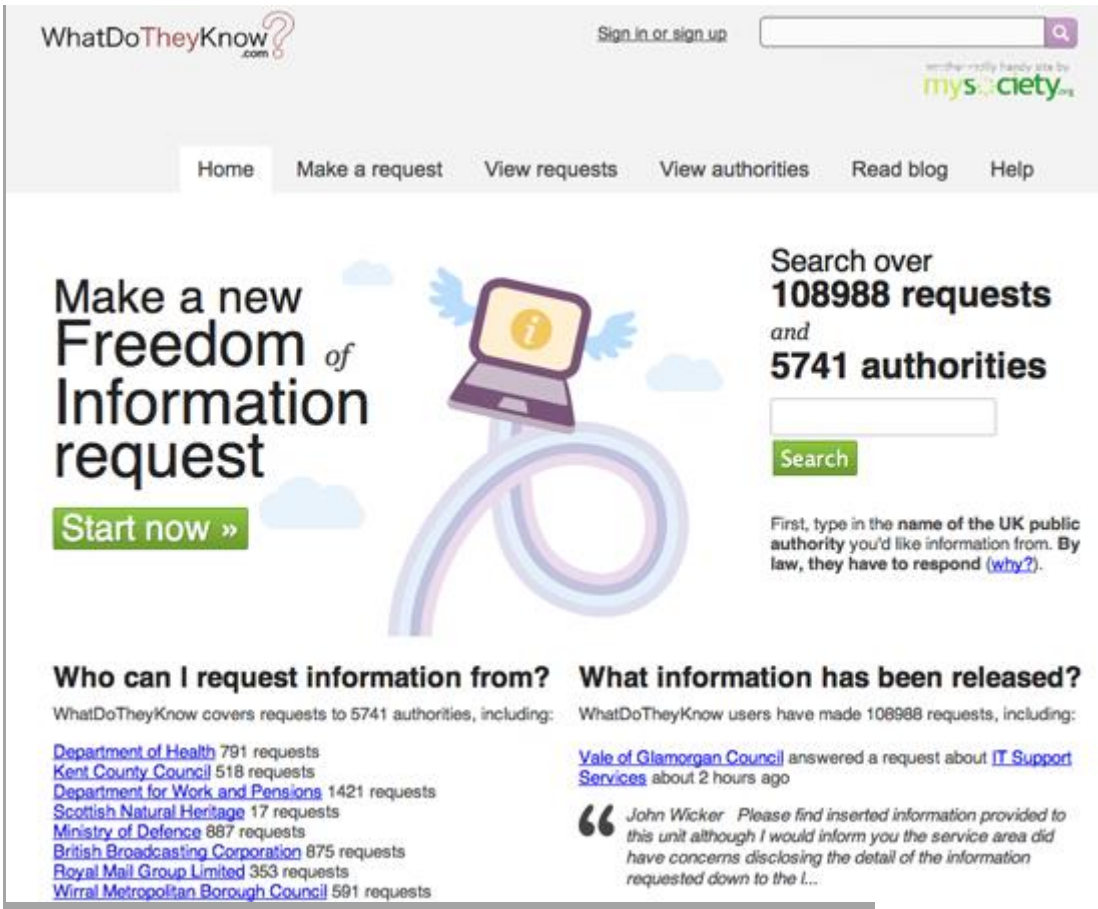


Imagem 2. What Do They Know? (My Society)

Envolva colegas

Se os seus colegas são céticos em relação a pedidos de acesso à informação, uma das melhores maneiras de convencê-los é escrever uma reportagem baseada em dados que você conseguiu utilizando a lei. Mencionar que fez uso da lei numa transmissão de rádio ou TV também é recomendado para a consciência do público em relação aos seus direitos.

Solicite por dados brutos

Se você quer analisar, explorar ou mexer nos dados usando um computador, deve pedir claramente por dados em um formato eletrônico e que possam ser tabulados. Você deve deixar especificar, por exemplo, que está pedindo informações orçamentárias em um formato "compatível para análise por um programa de contabilidade". Você também deve, de maneira clara, solicitar por informação em formato desagregado ou granular. Você pode ler mais a respeito neste relatório.

Organizações isentas das leis de acesso à informação

Você deve se informar sobre ONGs, empresas privadas, organizações religiosas e/ou outras instituições não obrigadas a divulgar documentos sob as leis de acesso à informação. No entanto, é possível encontrar dados sobre elas pedindo a órgãos públicos cobertos pelas leis. Por exemplo, você pode solicitar a um ministério se eles financiaram ou lidaram com uma empresa privada ou ONG específicas e pedir documentos. Se precisar de ajuda extra para solicitações baseadas nas leis de acesso, você pode também consultar o Kit de ferramentas de vazamentos legais para jornalistas.

— Helen Darbishire (Access Info Europe), Djordje Padejski (Knight Journalism Fellow, Stanford University), Martin Rosenbaum (BBC), e Fabrizio Scrollini (London School of Economics and Political Science)

Usando a Lei de Acesso à Informação para Entender Gastos

Já usei a lei de maneiras diferentes para ajudar a cobrir a COINS, a maior base de dados do Governo do Reino Unido para gastos, orçamentos e informações financeiras. No início de 2010, George Osborne afirmava que, caso ele se tornasse um chanceler, iria divulgar a base de dados COINS para promover maior transparência no Tesouro. Na época, me pareceu uma boa ideia investigar os dados e a estrutura da COINS, então enviei alguns pedidos baseados na Lei de Acesso à informação; um requisitando o esquema do banco de dados, um pedindo as instruções que os funcionários do Tesouro recebem quando vão trabalhar no COINS, e um pelo contrato do Tesouro com o provedor da base de dados. Todos eles resultaram na publicação de informações úteis. Também solicitei todos os códigos de despesas presentes na base de dados, que também foram publicados. Tudo isso ajudou a entender a COINS quando George Osborne efetivamente se tornou chanceler em maio de 2010, e publicou a base de dados em junho. Os dados da COINS foram usados em diversos sites incentivando o público a investigá-los —

incluindo OpenSpending.org e o site do The Guardian Coins Data Explorer.

Após a realização de mais investigações, parecia que uma grande parte do banco de dados não estava sendo divulgada: a Whole of Government Accounts (WGA, ou Contabilidade Total do Governo), que incluía 1.500 tipos de contas relacionadas a órgãos financiados com verba pública. Utilizei a lei de acesso para solicitar os dados do WGA de 2008 e 2009, mas sem sucesso. Solicitei o relatório feito pelo escritório de auditoria do WGA - que eu esperava que fosse explicar os motivos pelos quais o WGA não estava em condições de ser divulgado. Isso também foi recusado.

Em dezembro de 2011, a WGA foi divulgada nos dados COINS. No entanto, eu queria ter certeza que havia elementos suficientes para ver todo o conjunto de contas para cada um dos 1.500 órgãos incluídos no WGA. O que me levou à segunda maneira de utilizar a lei: garantir que os dados divulgados sob a agenda de transparência do Reino Unido estavam bem explicados e informavam o que deveriam. Enviei uma solicitação baseada na lei pedindo o grupo inteiro de contas para todos os órgãos públicos incluídos no WGA.

— Lisa Evans, the Guardian

Lei de Acesso à Informação no Brasil: Um longo caminho a percorrer

Como se sabe, a Constituição brasileira garante o direito de se requisitar informação do Estado no artigo 5º, inciso XXXIII e, também, o dever de os agentes públicos darem publicidade a seus atos (art. 37, caput). Nunca foi unânime a opinião de que se precisaria regulamentar esses dispositivos por meio de uma legislação específica. Com efeito, uma vez que a Constituição garante deveres e direitos, não deveria haver necessidade de elaborar ulteriormente a questão – além de criar outros problemas. Foi a constatação da ineficácia dos preceitos constitucionais na vida prática das relações entre o Estado e a sociedade que levou alguns dos céticos iniciais a mudarem de lado.

O principal problema trazido pela regulamentação efetuada pela lei nº 12.527/2011 foi ter levado a figura da "informação sigilosa" às três esferas e três poderes. Antes da lei, a noção jurídica do sigilo só existia para informações detidas pela administração pública federal. Depois da lei, qualquer estado, município, Tribunal de Contas, ente legislativo e assim por diante passou a gozar da prerrogativa de definir – sempre arbitrariamente – que tais ou quais tipos de informações seriam sigilosas. Por exemplo, o Tribunal de Contas da União estabeleceu que informações sobre gastos incorridos pelos gabinetes de seus ministros são sigilosas.

Esse gênero de oportunismo da opacidade está sendo praticado em todos os cantos do país. Como a nova legislação define que cada poder, em cada esfera, define seu próprio mecanismo de recurso contra negativas de prestação de informação, o que acaba por ocorrer é que o mesmo indivíduo que definiu que determinado tipo de dado deve permanecer secreto é aquele que dá a palavra final a qualquer recurso.

Mesmo entes que constituem poderes autônomos, como é o caso dos Tribunais de Contas e do Ministério Público (o primeiro, parte do Legislativo, e o segundo, do Executivo), meramente definem que isto ou aquilo é sigiloso e fica tudo por isso mesmo.

Nos municípios brasileiros e em boa parte dos estados, em que não há contraditório político relevante, a situação é idêntica. Sem sofrer contestação de ninguém, e como agora a lei lhes faculta o direito de definir arbitrariamente o que é sigiloso e o que não é, os respectivos chefes de Executivo praticam, agora

escudados na lei, exatamente o que antes praticavam em contradição com a Constituição.

Como o Ministério Público tem lavado as mãos em relação ao assunto, nessas áreas é como se a lei de acesso a informação não existisse. Isso assim permanecerá por muito tempo, essencialmente porque o motivo não é jurídico ou legal, mas econômico.

A regulamentação promovida pela lei satisfaz a uma condição necessária para a melhor circulação de informação. Tal condição, contudo, está longe de ser suficiente para atingir esse objetivo.

A lei estabelece o que pode, ou seja, condições sobre a oferta de informações: famílias de dados que devem ser tornados públicos por todos os órgãos do Estado, prazos para a prestação de informações que sejam solicitadas e a criação de organismos que recebam recursos de solicitantes caso informações requisitadas sejam recusadas ou não sejam fornecidas.

Ocorre que a regulamentação da oferta de qualquer coisa não cria demanda. Exceto no que tange a obrigatoriedade de publicação de certos dados relativos à execução orçamentária (mas mesmo assim o enforcement depende bastante da presença de quem vigie o assunto e reclame do eventual descumprimento), é óbvio que a consequência pretendida pela lei só ocorrerá se houver procura por informação.

Só isso poderia suprir a condição suficiente: a presença de uma demanda contínua e crescente por informação de qualidade e profundidade cada vez maiores. Não é o que acontece na maior parte do Brasil.

Em qualquer país, os demandantes por informação do Estado são, pela ordem: o setor privado; a imprensa; organizações não governamentais; acadêmicos; cidadãos. Evidentemente, cada um desses grupos procura informação porque tem algum interesse ou motivação. Quando as condições são desfavoráveis para o desenvolvimento de interesses, não há por que buscar informação.

É possível ver isso claramente nas diferenças entre as cobranças que se fazem a órgãos das três esferas administrativas. Os órgãos federais dos três poderes são os mais procurados. Os estados recebem demandas em grau variável conforme a região. E os municípios basicamente não recebem demandas.

A disparidade tem evidente origem no grau de desenvolvimento de cada lugar. Os estados mais pobres recebem menos demandas do que os mais ricos e os municípios, cuja imensa maioria é muito pobre, passam ao largo da questão.

É fácil entender por que as coisas se dão desse modo. Fechando a atenção sobre os municípios, dados da Secretaria do Tesouro do Ministério da Fazenda dão conta de que, em mais de 80% deles, os orçamentos dependem em alguma medida de repasses da União e dos estados. Desses, metade, ou cerca de 40% da totalidade das 5.653 municipalidades do país, dependem desses repasses em mais de 90% de seus orçamentos. Eles praticamente não arrecadam impostos locais (ISS, IPTU e outros).

A virtual ausência de arrecadação decorre da inexistência de atividade econômica robusta. Se não há criação de riqueza, não há competição entre empresas (não há empresas), entre capital e trabalho (não há capital nem trabalho) e, portanto, o contraditório político, quando existe, dá-se em torno das conveniências das micro-oligarquias locais. A totalidade da população depende da Prefeitura para sobreviver. Nessas condições, não há por que esperar que alguém formule demandas dirigidas à municipalidade.

A eventual imprensa que exista nesses lugares, quando não pertence aos oligarcas municipais, não pode sobreviver de anúncios (pois não há empresas que anunciem), subsistindo de favores da Prefeitura e dos governos estaduais, que assim adquirem apoio político. Ou seja, não se pode esperar dessa imprensa que aja criticamente em relação aos governantes.

Quanto às ONGs locais, quando existem (e existem às centenas de milhares, conforme o IBGE) servem para executar políticas públicas, sendo ingênuo esperar que nelas se desenvolva qualquer espécie de atitude crítica em relação à Prefeitura ao governo estadual ou aos demais poderes.

(O poder Legislativo seria um demandante importante de informação, não fosse o fato de ser ele comensal do poder, cooptado que é pelo mecanismo deletério do loteamento da administração pública entre os partidos políticos que formam a “base” do prefeito, do governador, do presidente da República.)

No final das contas, portanto, não há ninguém nesses lugares quem se anime a provocar a municipalidade na busca de informação.

A mesma situação de carência de demanda afeta boa parte dos estados do país, e pelo mesmo motivo: o subdesenvolvimento é incompatível com a formulação de demandas por informação.

Observe-se que a constatação da pobreza da demanda antecede a promulgação da lei de acesso a informação. Embora de modo desigual, a esfera federal brasileira, bem como diversos estados, produzem há muitos anos uma grande quantidade de dados sobre assuntos variados. O aproveitamento dessa informação pelos atores esperados (ONGs, jornais etc.) tem sido muito pequeno.

Há múltiplas razões para isso. A imprensa nacional que de fato demanda informação é constituída basicamente de três jornais diários e duas revistas semanais (deixando de lado os meios eletrônicos, cuja pauta não é normalmente "investigativa"). Entre as ONGs, das muitíssimas que há no país resta um punhado, contado nos dedos de uma mão, que se dedica a buscar e processar dados públicos para atingir seus objetivos institucionais.

A academia, por sua vez, opera com maturação lenta e sua produção tem repercussão pública limitada. Por fim, cidadãos privados não fazem demandas estruturadas.

Dado esse quadro de carência generalizada, não se deve esperar que a regulamentação do acesso a informação resulte em um salto significativo na qualidade do monitoramento do Estado. Os progressos que se possam esperar serão lentos, dar-se-ão primordialmente na esfera federal e secundariamente nos estados e municípios mais ricos. Os mais pobres permanecerão com os mesmos fluxos de informação deficientes que os afetavam antes da promulgação da lei.

— Claudio Weber Abramo, Transparência Brasil

Pedidos de informação funcionam. Vamos usá-los!

Usar a legislação de acesso à informação - ou fazer 'wobbing', como alguns chamam - é uma excelente opção. Mas exige método e, muitas vezes, persistência. Abaixo mostro três exemplos sobre os pontos fortes e os desafios do wobbing retirados do meu trabalho como jornalista investigativo.

Nota da tradução: wobbing é um neologismo, uma gíria surgida entre jornalistas holandeses para usar a lei de acesso a informação.

Estudo de caso 1: Subsídios agrícolas

Todos os anos, a União Europeia paga quase 60 bilhões de euros aos fazendeiros e ao setor agrícola. Todos os anos. Isso acontece desde o final dos anos 1950 e o argumento político é que os subsídios ajudam os agricultores mais pobres. No entanto, uma descoberta com base na lei de acesso à informação na Dinamarca em 2004 indicou que esta era apenas uma desculpa. Os pequenos agricultores estavam com dificuldades, como tantas vezes reclamaram, e, na realidade, a maior parte do dinheiro foi para um pequeno número de grandes proprietários de terra e para a agroindústria. Obviamente, eu queria descobrir se isso era um padrão na Europa.

No verão de 2004, pedi os dados à Comissão Europeia. Todos os anos, em fevereiro, a Comissão recebe os dados dos países membros. Na informações, estão quem se candidata para receber o financiamento da União Europeia, quanto os beneficiários conseguem, e se pegam os recursos para cultivar a terra, desenvolver a região deles ou para exportar leite em pó. A Comissão recebia as estatísticas como arquivos CSV em um CD. Uma grande quantidade de dados, mas, em princípio, fácil de trabalhar. Isto é, se você conseguisse por as mãos neles.

A Comissão recusou-se a divulgar os dados. O principal argumento era de que eles estavam dentro de um banco de dados e não poderiam ser recuperados sem um extenso trabalho. Uma explicação que o Ombudsman Europeu considerou como *má administração*. Você pode encontrar todos os documentos sobre este caso no site wobbing.eu. Mas não tínhamos tempo a perder com questões legais. Queríamos os dados.

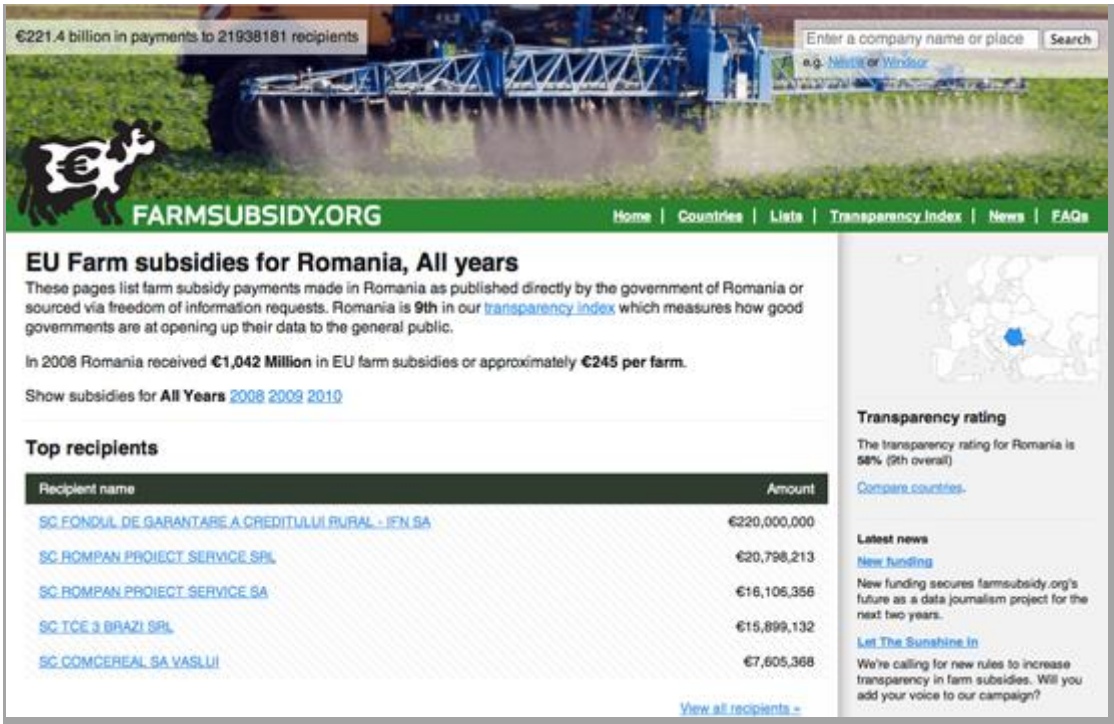


Imagem 3. O site sobre subsídios agrícolas (Farmsubsidy.org)

Assim, nos juntamos com parceiros em toda a Europa para obter os dados país por país. Colegas ingleses, suecos e holandeses conseguiram as informações em 2005. Finlândia, Polônia, Portugal, regiões da Espanha, Eslovênia e outros países abriram os dados também. Mesmo na Alemanha, onde é difícil usar a lei de acesso, obtive informações na província da Renânia do Norte-Westfalia em 2007. Tive de ir até o Tribunal de Justiça para obter os dados, mas isso resultou em alguns artigos legais na revista Stern.

Coincidência a Dinamarca e o Reino Unido terem sido os primeiros a abrir os dados? Não necessariamente. Naquela época, os subsídios agrícolas estavam sendo contestados na Organização Mundial do Comércio (OMC). Dinamarca e Reino Unido estão entre os países mais liberais da Europa, portanto, pode ser que ventos políticos tenham soprado na direção da transparência naqueles países.

A história não parou por aí; para mais episódios e para obter os dados, vejafarmsubsidy.org.

Lição: use e abuse das leis de informação. Há uma fabulosa diversidade de leis do tipo na Europa – e diferentes países podem ter diferentes interesses políticos em épocas diferentes. Pode-se tirar vantagem daí.

Conheça seus direitos

Você deve se preocupar sobre diretos autorais e e outras licenças ao publicar dados? Embora seja sempre bom checar com a equipe jurídica da sua publicação, vale regra geral: se os dados são publicados pelo governo, não se deve pedir nem perdão nem permissão; se forem publicados por uma organização que não faz dinheiro vendendo os dados, não há muito com o que se preocupar; se forem publicados por uma organização que faz dinheiro com a venda dos dados, então você definitivamente deve pedir permissão.

— Simon Rogers, the Guardian

Estudo de caso 2: Efeitos colaterais

Todos somos cobaias quando se trata de tomar remédio. As drogas podem ter efeitos colaterais. Nós sabemos: pesamos os benefícios e riscos potenciais e tomamos uma decisão. Infelizmente, nem sempre estamos bem informados para tomar essa decisão.

Quando adolescentes tomam uma pílula contra espinhas, eles esperam uma pele macia – e não um súbito mau humor. Mas foi exatamente isso que aconteceu com um medicamento: os jovens se tornaram depressivos e até mesmo suicidas depois de tomá-lo. A informação sobre o perigo deste efeito colateral — uma história óbvia para jornalistas — não estava facilmente disponível.

Há dados sobre efeitos colaterais. Os fabricantes têm de entregar regularmente para as autoridades de saúde informações sobre efeitos colaterais observados. Esses dados são mantidos por autoridades nacionais ou europeias depois que a droga é permitida no mercado.

O primeiro furo novamente veio da Dinamarca, da esfera federal. Durante uma investigação sobre o tema envolvendo uma equipe de dinamarqueses, holandeses e belgas, a Holanda também liberou seus dados. Outro exemplo de uso de leis de acesso à informação: ajudou bastante no caso chamar a atenção das autoridades holandesas para o fato de que os dados estavam acessíveis na Dinamarca.

Mas a história era verdadeira: na Europa, havia jovens suicidas e, infelizmente, também suicídios em vários países como resultado do medicamento. Jornalistas, pesquisadores e a família de uma jovem vítima estavam fazendo de tudo para ter acesso a essa informação. O Ombudsman europeu ajudou a pressionar por transparência na Agência Europeia de Medicamentos, e ao que parece, ele foi bem-sucedido. Então, os jornalistas puderam se debruçar sobre

os dados. Somos todos cobaias, como um pesquisador colocou, ou os mecanismos de controle são sólidos?

Lições: Não aceite um 'não' como resposta quando se trata de transparência. Seja persistente e siga a história todo o tempo. As coisas bem podem mudar com melhor acesso às informações mais adiante.

Estudo de caso 3: Mortes por causa do contrabando

Fatos da história recente podem ser extremamente dolorosos para populações inteiras, especialmente após de guerras e em tempos de transição. Dessa forma, como os jornalistas podem conseguir dados concretos para uma investigação sobre isso, quando — por exemplo — os beneficiários da guerra ocorrida na década passada estão agora no poder? Esta foi a tarefa de uma equipe de jornalistas da Eslovênia, Croácia e Bósnia .

A equipe começou a investigar o comércio de armas na ex-Iugoslávia durante um embargo da ONU no início de 1990. A base do trabalho foram documentos de inquéritos parlamentares sobre o assunto. Para documentar as rotas dos embarques e compreender a estrutura do comércio, o transporte teve de ser rastreado pela numeração dos navios nos portos e por placas de caminhões.

Comissões parlamentares eslovenas tinham realizado investigações sobre quem havia lucrado ilegalmente com a Guerra dos Balcãs, mas nunca chegou a uma conclusão. Mesmo assim, havia uma trilha extremamente valiosa de documentos revelados, incluindo 6 mil páginas que a equipe eslovena obteve por meio de um pedido de acesso à informação.

Neste caso, os dados tinham de ser extraídos dos documentos e classificados em bancos de dados. Aprimorando os dados com informações adicionais, análises e pesquisas, eles foram capazes de mapear numerosas rotas de comércio ilegal de armas.

A equipe foi bem-sucedida e os resultados são únicos e já garantiram ao time o primeiro prêmio deles. O mais importante é que a história importa para toda a região e pode bem ser melhorada por jornalistas de outros países pelos quais as cargas mortíferas passaram.

Lições: Dê visibilidade para matéria-prima que considerar boa, mesmo se você encontrá-la em lugares inesperados, e combine esse material com dados públicos existentes e acessíveis.

— Brigitte Alfter, Journalismfund.eu

Lei de acesso à informação com amigos

Muitos países dos Balcãs têm problemas com corrupção no governo, especialmente quando se trata de prestação de contas. Durante vários meses, em 2009, um grupo de jornalistas sérvios do Centre for Investigative Reporting, de Belgrado, vinham pedindo por leis de acesso diferentes tipos de documentos de mais de 30 municípios. Antes disso, quase nada estava acessível ao público. A ideia era obter os registros públicos originais e colocar os dados em planilhas, possibilitando executar verificações básicas e comparações entre os municípios e também obter uma noção de gastos máximos e mínimos.

Eram indicadores básicos como orçamento, despesas regulares e especiais, salários de autoridades, despesas de viagem, número de funcionários, despesas de telefone celular, gastos com ajuda de custo, valores de contratos públicos, etc. Foi a primeira vez que repórteres pediram esses tipo de informação.

O resultado foi uma base de dados abrangente que revela vários dados maquiados, malfeitos e casos de corrupção. Uma lista dos prefeitos mais bem pagos indicou que alguns deles estavam recebendo mais dinheiro do que o presidente sérvio. Muitos outros funcionários estavam recebendo rendimentos excessivos, com gigantescos reembolsos de viagens e de ajudas de custo. Nossos dados sobre contratos públicos, obtidos com dificuldade, ajudaram a revelar uma bagunça oficial.

Mais de 150 reportagens foram produzidas usando a base de dados e muitas delas foram aproveitadas pela mídia sérvia local e nacional. Nós aprendemos que comparar os registros com os dados de governos semelhantes pode mostrar desvios e lançar luz sobre prováveis casos de corrupção. Despesas exageradas e incomuns podem ser detectadas somente pela comparação.

— Djordje Padejski, Knight Journalism Fellow, Stanford University

Ultrapassando Obstáculos para obter Informação

Você tentou de tudo e ainda não conseguiu obter os dados. Encontrou eles na web, mas não há nenhuma opção para baixá-los e não foi possível copiar e colá-los. Não se preocupe, talvez ainda haja uma maneira de obter esses dados. Por exemplo, você pode:

- Obter os dados através de APIs web, interfaces providas por bases de dados e por várias aplicações web modernas (incluindo Twitter, Facebook, dentre outras). Essa é uma maneira fantástica de acessar tanto dados do governo ou dados privados quanto dados de sites de mídias sociais.
- Extrair as informações de arquivos PDF. Isso é muito difícil, pois o PDF é uma linguagem para impressoras e não possui muita informação sobre a estrutura dos dados exibidos. Mostrar como retirar informações de PDFs está além do escopo deste livro, mas existem algumas ferramentas e tutoriais que podem ajudá-lo.
- Extrair informações de telas dos sites (scraping). Consiste em extrair automaticamente conteúdo estruturado de uma página com o auxílio de um utilitário de captura ou programando um código. Embora esse método seja muito poderoso e possa ser usado em diferentes ocasiões, ele requer um certo nível de conhecimento sobre como a web funciona.

Diante de todas essas opções, não esqueça das mais simples: vale investir tempo buscando arquivos com dados já em formatos interpretáveis por máquinas ou até mesmo entrar em contato com a instituição que cuida dos dados que você deseja. Neste capítulo mostraremos um exemplo básico de como extrair dados (scraping) de uma páginas feita em HTML.

O que são Dados Legíveis por Máquinas?

O objetivo da maioria desses métodos é obter acesso a dados legíveis por máquinas. São dados criados para serem processados por computadores, em vez de serem apresentados a um ser humano. A estrutura desses dados está relacionada à informação que eles representam e não na maneira como são eventualmente exibidos. Exemplos incluem arquivos CSV, XML, JSON e outros arquivos do Excel, enquanto formatos como documentos do Word, páginas HTML, e arquivos PDF estão mais relacionados à apresentação visual da informação. O PDF, por exemplo, é em uma linguagem que conversa

diretamente com impressoras; ela se preocupa com o posicionamento de pontos e linhas numa página em vez de se focar na distinção entre as letras.

Captura de sites web: para quê?

Você visita um site, vê uma tabela interessante e tenta copiá-la para o Excel para acrescentar dados ou simplesmente guardá-la. Só que isso muitas vezes não funciona, ou a tabela que você quer está espalhada por várias páginas. Como copiar manualmente pode se tornar um trabalho tedioso, pode fazer sentido automatizar o trabalho escrevendo um pouco de código.

A vantagem deste tipo de captura é que você pode fazê-la em praticamente qualquer site, de previsões do tempo a gastos do governo, mesmo que o site não ofereça nenhuma API de acesso aos dados brutos.

O que é possível capturar

Existem limites para o que você consegue capturar por código. Alguns fatores podem dificultar o processo:

- Código HTML mal formado ou informação não estruturada (por exemplo, sites governamentais antigos).
- Sistemas de autenticação feitos para barrar acessos automatizados (por exemplo códigos CAPTCHA e paywalls).
- Sistemas baseados em sessão que usam cookies para rastrear a navegação do usuário.
- Ausência de listagens completas ou de possibilidade de realizar buscas usando caracteres curingas.
- Bloqueio, por parte dos administradores dos sites, de acessos em massa aos dados.

Pode haver também limitações legais: alguns países reconhecem direitos autorais sobre as bases de dados, podendo limitar o reuso da informação online. Às vezes você até pode ignorar essa licença — dependendo de onde você more, pode ter direitos especiais como jornalista. Capturar dados governamentais disponíveis na internet normalmente é legal, mas talvez seja o caso de confirmar antes de publicá-los. Organizações privadas e certas ONGs costumam ser menos tolerantes e talvez possam alegar que você está "sabotando" os sistemas deles. Outras informações podem infringir a privacidade de indivíduos e, dessa forma, violar as leis de privacidade de dados ou a ética profissional.

Correção, Captura, Compilação, Limpeza

O desafio relacionado a maioria dos dados do Reino Unido não é tê-los publicados, mas sim tê-los em um formato útil. Um monte de dados sobre gastos de viagens, bens dos membros do parlamento e de ocorrências de lobby são publicados em formatos difíceis de serem analisados.

Para algumas informações, só resta um árduo trabalho: combinar dúzias de arquivos curtos de Excel, por exemplo, é a única maneira de criar listas detalhadas sobre reuniões ministeriais no Reino Unido. Mas para outras informações, fazer a captura de telas de sites pode ser incrivelmente útil.

Usar serviços como o do site ScraperWiki para obter ajuda de programadores na produção programas que capturem registros como os dos bens dos membros do parlamento pode poupar metade do nosso trabalho: ao fim, conseguimos todos esse dados em uma única planilha, prontos para iniciar o trabalho de análise e limpeza.

Serviços como esse (ou ferramentas como o Outwit Hub) são de grande ajuda a jornalistas que precisam compilar dados desorganizados mas não conseguem programar sozinhos.

— James Ball, the Guardian

Ferramentas que ajudam na captura

Há vários programas que podem ser usados para extrair informações em massa de um site. Dependendo do seu browser, ferramentas como Readability (que ajudam a extrair texto de uma página) ou DownThemAll (que permite que você baixe vários arquivos de uma única vez) ajudarão a automatizar tarefas tediosas. Já o Scraper extension do Chrome foi criado especificamente para extrair tabelas de sites. Extensões como o FireBug permitem acompanhar exatamente como um site é construído e quais comunicações acontecem entre o navegador e o servidor.

ScraperWiki é uma página que permite que você codifique programas de captura em várias linguagens de programação diferentes, incluindo Python, Ruby e PHP. Se quiser começar a criar programas de captura sem armar um ambiente de programação no seu computador, esse é o caminho. Outros serviços, como o Google Spreadsheets e o Yahoo! Pipes também ajudam a fazer capturas de alguns sites.

Como um programa de captura (scraper) funciona?

Web scrapers geralmente são pequenos pedaços de código escritos em uma linguagem de programação como Python, Ruby ou PHP. A linguagem certa é uma questão de qual comunidade você tem acesso: se existe alguém na sua redação já trabalhando numa dessas linguagens, então faz sentido adotar a mesma linguagem.

Embora algumas das ferramentas mencionadas anteriormente sejam úteis para começar, a real complexidade envolvida em fazer capturas está em mirar as páginas certas e os elementos certos dentro dessas páginas para extrair a informação desejada. Essas tarefas não estão relacionadas a programação, mas ao entendimento das estruturas do site e do seu banco de dados.

Quando você abre um site, seu navegador irá quase sempre recorrer a duas tecnologias: HTTP, para se comunicar com o servidor e requisitar um recurso específico, como documentos, imagens ou vídeos; e HTML, a linguagem na qual os sites são criados.

A anatomia de uma webpage

Qualquer página HTML está estruturada como uma hierarquia de caixas (definidas pelas "tags" HTML). Uma caixa maior irá conter várias caixas menores — por exemplo, uma tabela possui várias divisões menores: linhas e células. Há vários tags realizando diferentes funções — algumas produzem caixas — outras tabelas, imagens ou links. Tags também podem ter propriedades adicionais (ex: podem ser identificadores únicos) e pertencer a grupos chamados "classes", que fazem com que seja possível mirar e capturar elementos individuais dentro de um documento. Assim, selecionar os elementos apropriados e extrair seu conteúdo é um ponto chave ao escrever um programa de captura.

Visualizando elementos em uma página web, tudo pode ser quebrado em caixas dentro de caixas.

Para fazer a captura, você precisará aprender um pouco sobre diferentes elementos que podem estar em um documento HTML. Por exemplo, o elemento `<table>` abrange uma tabela inteira, que tem uma `<tr>` (linha de tabela) que por sua vez contém `<td>` (dados da tabela) para cada célula. O elemento mais comum que você irá encontrar é o `<div>`, que basicamente significa qualquer bloco de conteúdo. A maneira mais fácil de se habituar com

esses elementos é usando uma developer toolbar no seu navegador: ela permite que, ao deixar o cursor do mouse sobre qualquer parte da página web, você veja o código por trás daquele elemento.

Tags trabalham marcando o início e o término de uma unidade. Por exemplo `<em>` *signifca o início de pedaço de texto que foi enfatizado com o estilo itálico e* `</em>` significa o final dessa seção. Fácil.

Um exemplo: Capturando Incidentes Nucelares com Python

NEWS é o portal da Agência de Energia Atômica Internacional (AIEA) para incidentes radioativos (e um forte candidato a membro do clube dos títulos estranhos!). A página lista incidentes em um site simples de estilo parecido ao de um blog que pode ser facilmente capturado.

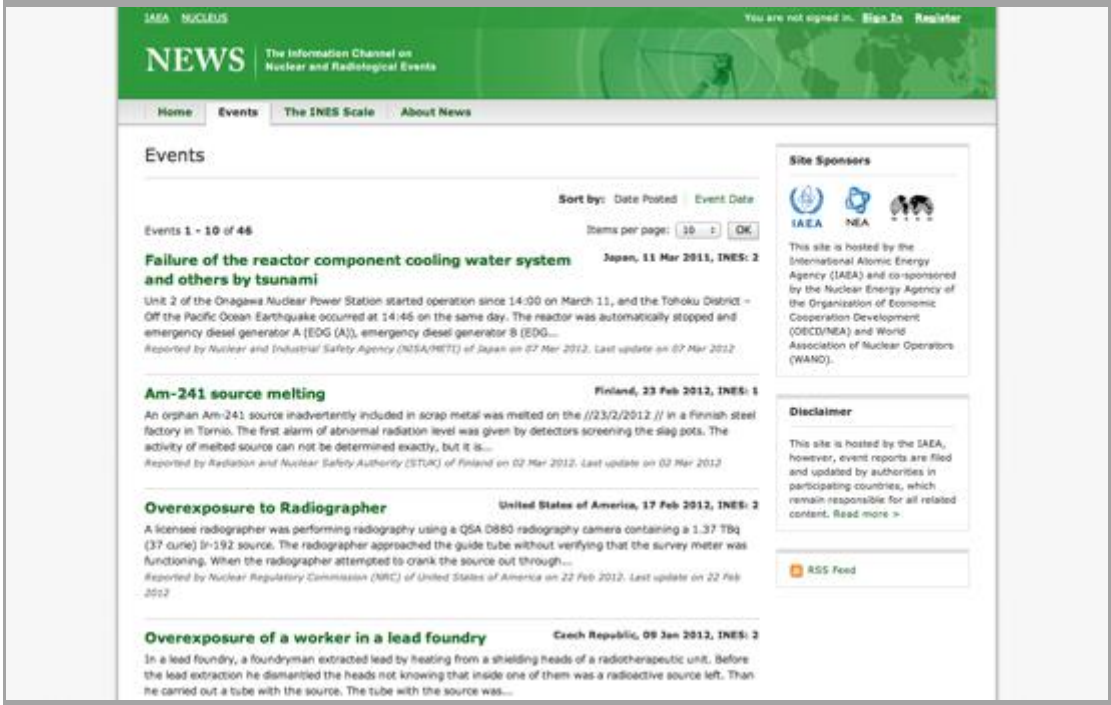


Imagem 4. O portal da Agência de Energia Atômica Internacional (AIEA) (news.iaea.org)

Para começar, crie um novo programa de captura (scraper) em linguagem Python noScraperWiki e você será apresentado a uma área de texto vazia, com excessão de alguns códigos prontos de suporte. Em uma outra janela do navegador, abra o site da AIEA e abra a developer bar do seu navegador. No view "Elements" tente localizar o elemento HTML para um dos itens de notícias. A barra developer bar ajuda você a conectar elementos na página web com seu código HTML relacionado.

Uma investigação nessa página irá revelar que os títulos são elementos `<h4>` dentro de uma `<table>`. Cada evento é uma linha `<tr>`, que também contém uma descrição e uma data. Se quisermos extrair os títulos de todos os eventos, devemos encontrar uma maneira de selecionar cada linha na tabela sequencialmente, enquanto copiamos o texto.

Para transformar esse processo em código, precisamos tomar conhecimento nós mesmos de todos os passos envolvidos. Para se ter uma ideia dos passos requeridos, vamos jogar um jogo: na janela do seu ScraperWiki, tente você mesmo escrever instruções individuais para cada coisa que você fará ao escrever o programa de captura, como passos de uma receita (ponha antes cada linha com um sinal de # para dizer ao Python que ela não se trata de um código) Por exemplo:

```
# Procure por todas as linhas na tabela

# Não deve ultrapassar o lado esquerdo.
```

Tente ser o mais preciso que puder e não assuma que o programa sabe alguma coisa sobre a página que você está tentando capturar.

Tendo escrito algum pseudo código, vamos compará-lo a esse código essencial para o seu primeiro capturador:

```
import scraperwiki

from lxml import html
```

Nessa primeira frase, nós estamos importando funcionalidades existentes de bibliotecas — trechos de código previamente escritos. `scraperwiki` nos dará a habilidade para baixar sites web, enquanto `lxml` é uma ferramenta para a análise estrutural de documentos HTML. Boa notícia: Se você está escrevendo um programa de captura em Python com o ScraperWiki, essas duas linhas de código sempre serão as mesmas.

```
url = "http://www-news.iaea.org/EventList.aspx"

doc_text = scraperwiki.scrape(url)

doc = html.fromstring(doc_text)
```

Em seguida, o código cria uma variável: `url`, que indicará sempre o endereço da página da AIEA. Isso diz ao programa de captura que queremos prestar atenção a esse fator. Observe que a URL está entre aspas pois não faz parte do código do programa mas se trata apenas de uma *string*, uma sequência de caracteres.

Em seguida nós usamos a variável `url` como entrada para uma função, `scraperwiki.scrape`. Uma função irá fornecer algum trabalho definido — nesse caso, ela irá baixar a página web. Quando terminada, ela irá associar a sua saída a alguma outra variável, `doc_text`. `doc_text` irá agora armazenar o texto do site web; não na forma visual que você vê no navegador, mas o código fonte, incluindo as tags. Como esse formulário não é muito fácil de analisar, usaremos uma outra função, `html.fromstring`, para gerar um representação especial onde podemos facilmente atingir os elementos que queremos, o chamado modelo de objetos de documento (DOM).

```
for row in doc.cssselect("#tblEvents tr"):

link_in_header = row.cssselect("h4 a").pop()

event_title = link_in_header.text

print event_title
```

Neste passo final, usamos o DOM para encontrar cada linha na tabela e extrair o título dos eventos de seu cabeçalho. Dois novos conceitos são usados: o "for... loop" (para cada vez que um evento ocorra disparar outro) e o elemento de seleção (`.cssselect`). O código for loop irá atravessar uma lista de itens, associar a cada um pseudônimo temporário (`row` nesse caso) e depois executar qualquer instrução para cada item.

O outro novo conceito, elemento de seleção, faz uso de uma linguagem especial para encontrar elementos dentro do documento. Seletores CSS são normalmente usados para adicionar informação de leiaute aos elementos HTML e podem ser usados para precisamente selecionar um elemento de uma página. Nesse caso (linha 6), estamos selecionando `#tblEvents tr`, no qual irá corresponder cada `<tr>` dentro de um elemento tabela com o ID `tblEvents` (o sinal # significa ID). Observe que isso irá retornar uma lista de elementos `<tr>`.

Isso pode ser visto na linha seguinte (linha 7) onde nós estamos aplicando ourto seletor para encontrar qualquer `<a>` (que é um hyperlink) dentro de um `<h4>` (um título). Aqui desejamos apenas olhar um único elemento (existe apenas um título por linha), então nós temos que colocá-lo para fora do topo da lista retornada pelo seletor com a função `.pop()`.

Observe que alguns elementos no DOM contêm texto (isto é, texto que não é parte de nenhuma linguagem de marcação), que podemos acessar usando a sintaxe `[elemento].text` conforme vemos na linha 8. Finalmente, na linha 9, estamos imprimindo o texto no console do ScraperWiki. Se você executar o seu

programa de captura, a janela menor deverá iniciar a listagem dos nomes dos eventos do site web da IAEA.

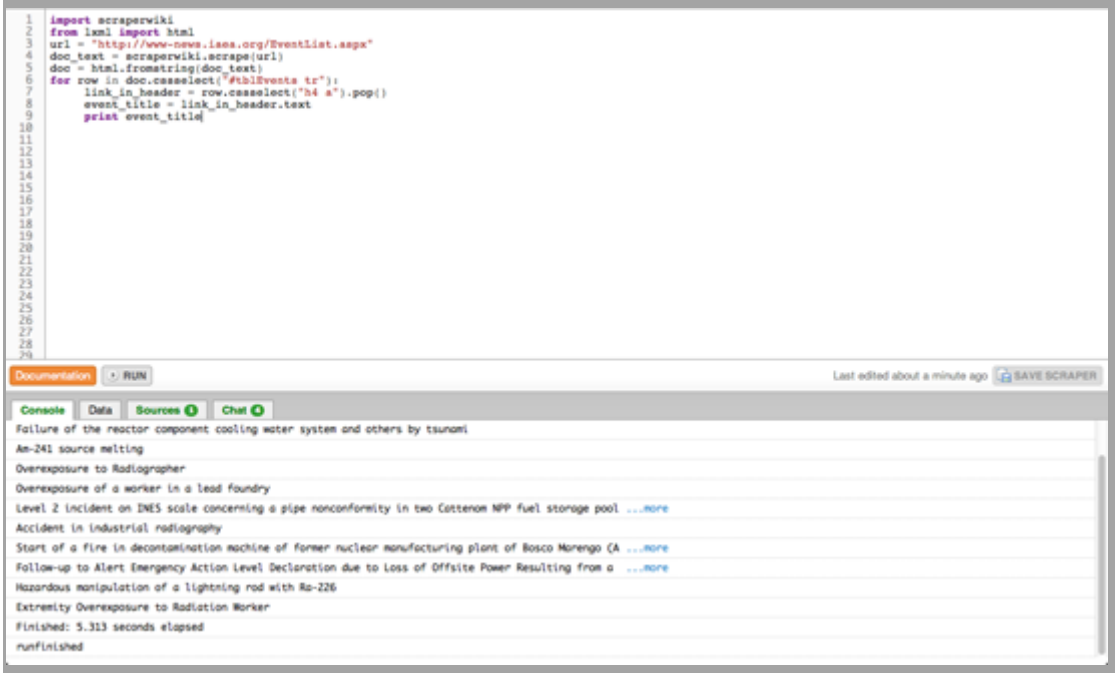


Imagem 5. Uma captura em ação (ScraperWiki)

Agora você pode ver um programa básico de captura operando: ele baixa a página web, a transforma em DOM, e em seguida permite que você possa selecionar e extrair certos conteúdos. Após ter essa noção básica, você pode tentar e resolver alguns dos problemas restantes usando o ScraperWiki e a documentação do Python:

- Você consegue encontrar o endereço do link em cada título de evento?
- Você consegue selecionar a pequena caixa que contém a data e o local usando o nome da classe CSS e extrair o texto do elemento?
- O ScraperWiki disponibiliza um pequeno banco de dados para cada programa de captura para que você possa armazenar os resultados; copie o exemplo relevante da documentação do ScraperWiki e adapte-o para que ele grave os títulos, links e datas dos eventos.
- A lista de eventos possui várias páginas; você consegue capturar múltiplas páginas para pegar o histórico de eventos também?

Conforme você for tentando resolver esses desafios, dê uma olhada em torno do ScraperWiki: existem vários exemplos úteis nos programas de captura já feitos; frequentemente os dados também são bastantes interessantes. Dessa forma,

você não precisa começar o seu programa de captura do zero: escolha um que seja similar, crie uma cópia e adapte ao seu problema.

— Friedrich Lindenberg, Open Knowledge Foundation

Capturando uma base de dados pública

Alguns médicos franceses são livres para escolher suas próprias taxas, de forma que uma pessoa pode pagar entre 70 e 500 Euros por uma consulta de 30 minutos a um oncologista, por exemplo. Esses dados das taxas são legalmente públicos, mas a administração somente disponibiliza uma base de dados online de difícil navegação. Para mostrar uma boa visão das taxas dos médicos para o Le Monde, decidi capturar a base de dados inteira.

Aí é onde a diversão começa. O formulário de busca é uma aplicação Flash que redireciona para uma página HTML de resultados através de uma requisição POST. Com a ajuda de Nicolas Kayser-Bril, demorou um pouco para descobrir como a aplicação poderia usar uma terceira página como um passo "escondido" entre o formulário de busca e a página de resultado. Essa página foi de fato usada para armazenar um cookie com valores do formulário de busca que depois foram acessados pela página de resultados. Teria sido difícil pensar em um processo mais complicado, mas as opções da biblioteca cURL no PHP tornam fácil contornar os obstáculos, uma vez que você saiba onde eles estão! No final, domar a base de dados foi uma tarefa de 10 horas, mas valeu a pena.

— Alexandre Léchenet, Le Monde

A Web como uma Fonte de dados

Como você pode descobrir mais sobre algo que só existe na Internet? Se você está querendo saber mais sobre um endereço de email, site, imagem ou artigo da Wikipedia, neste capítulo eu o levarei através das ferramentas que irão dizer mais sobre o que está por trás deles.

Ferramentas Web

Primeiro, alguns serviços que você pode usar para descobrir mais sobre um site inteiro, em vez de sobre uma página em particular:

Whois

Se você for em whois.domaintools.com pode obter informações básicas de registro de qualquer site. Recentemente, alguns donos de sites optaram por registros privados, no qual escondem seus detalhes, mas em vários casos você poderá ver o endereço, email e número de telefone da pessoa que registrou o site. Você também pode entrar com um endereço IP e obter dados da organização ou do indivíduo dono do servidor. Isso é especialmente útil quando você está tentando investigar o uso abusivo ou malicioso de um serviço, já que a maioria dos sites gravam o endereço IP de todos que os acessam.

Blekko

O buscador Blekko oferece uma quantidade incomum de insights sobre estatísticas internas que coleta de sites conforme rastreia a Web. Se você digitar o nome do domínio seguido de "/seo", você receberá uma página com informações sobre aquela URL. (Nota da Tradução: quando esta dica foi escrita, o site oferecia esse serviço gratuitamente. Agora, é preciso pagar para obter esse tipo de análise). A primeira aba mostra que outros sites possuem links para o domínio em ordem de popularidade. Isso pode ser extremamente útil quando você está tentando entender que cobertura um site está recebendo, e se você quiser entender por que ele está com uma classificação alta nos resultados de busca do Google, já que ela se baseia em links de entrada. Na segunda, abaixo, nos diz quais outros websites estão rodando da mesma máquina. É comum golpistas e spammers construírem uma legitimidade falsa criando muitos sites que comentam e se ligam uns aos outros. Eles parecem domínios diferentes, e talvez até tenham registros diferentes, mas frequentemente vivem num

mesmo servidor por ser muito mais barato. Essas estatísticas dão a você uma visão sobre a estrutura de negócio escondida dos sites que você estiver pesquisando.

Imagem 6. O buscador Blekko (Blekko.com)

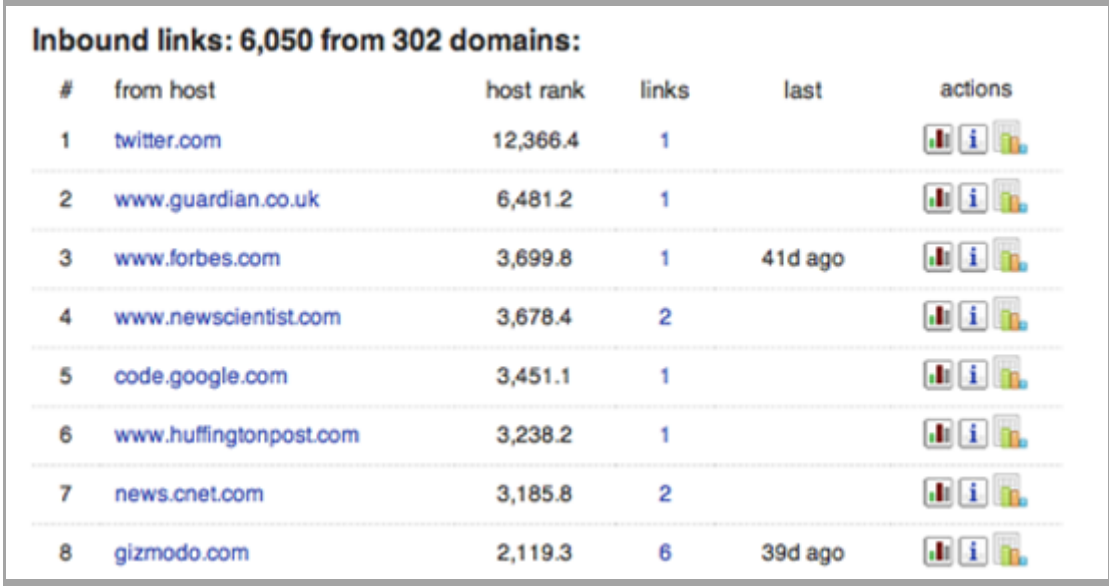


Imagem 7. Entendendo a popularidade na web: quem se liga a quem? Outra aba útil é a "Crawl stats", especialmente a seção "Cohosted with" ("Co-patrocinado por"). (Blekko.com)

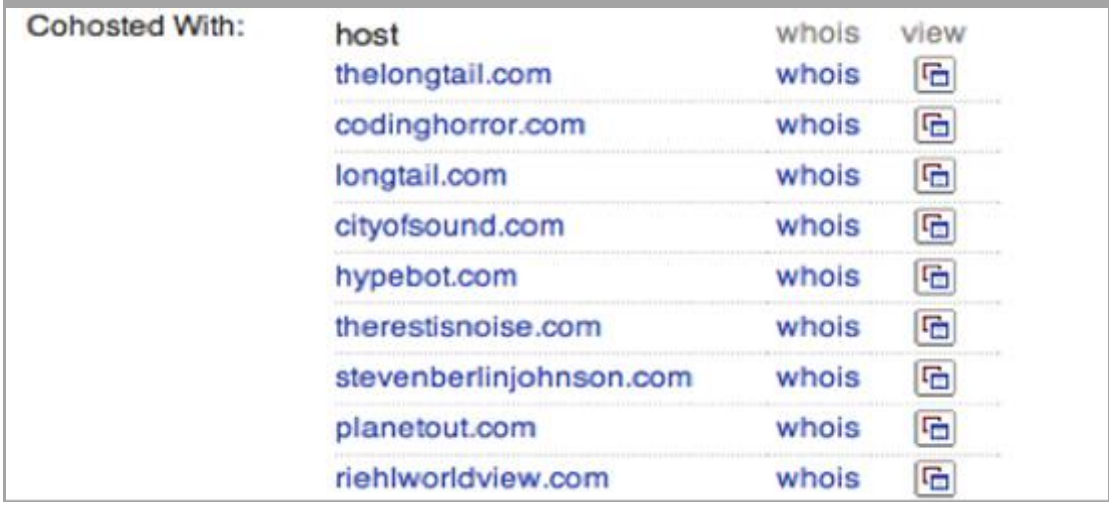


Imagem 8. Encontrando web spammers e scammers (Blekko.com)

Compete.com

Ao pesquisar um conjunto de consumidores americanos, o site compete.com obteve estatísticas de uso detalhadas da maioria dos sites, e tornou alguns detalhes básicos disponíveis de graça. Escolha a aba Site Profile e informe um domínio. Você irá em seguida ver um gráfico do tráfego do site no último ano, junto com números de quantas pessoas e quantas vezes o visitaram (como na segunda imagem abaixo). Como eles se baseiam em pesquisas, os números são apenas aproximados, mas eles se mostraram razoavelmente precisos quando pude compará-los com métricas internas. Parecem ser uma boa fonte para comparar dois sites. Embora os números absolutos possam ser imprecisos, ainda sim é uma boa representação da diferença relativa da popularidade. Como eles apenas pesquisam consumidores dos Estados Unidos, os dados podem ser ruins para sites de outros lugares.

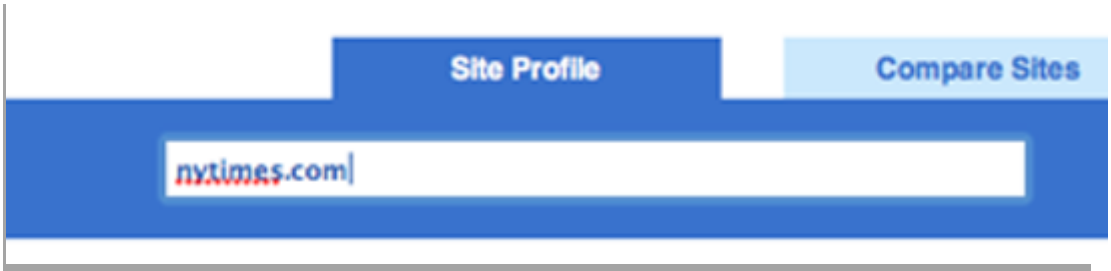


Imagem 9. Serviço de perfil do site Compete.com (Compete.com)

Imagem 10. O que está em voga? Qual é a demanda?: Hotspots na web (Compete.com)

Google

Uma função do Google que pode ser extremamente útil é a palavra-chave "site:". Se você adicionar "site:nomedapagina.com.br" à sua busca, o Google retornará apenas resultados dentro daquele site. Você pode ainda filtrar mais incluindo o prefixo das páginas que você está interessado, como por exemplo, "site:nomedapagina.com.br/secaox/", e você verá apenas resultados que correspondem àquele padrão. Isso é extremamente útil quando você estiver procurando por informações que os proprietários de um domínio tornaram públicas mas não tiveram interesse em divulgar.

Webpages, imagens e vídeos

Algumas vezes você está interessado nas atividades em torno de uma história em particular, e não no site inteiro. As ferramentas abaixo mostram a você diferentes ângulos de como as pessoas estão lendo, respondendo, copiando e compartilhando conteúdo na web.

Bit.ly

Sempre recorro ao bit.ly quando desejo saber como as pessoas estão compartilhando um determinado link. Para usá-lo, entre com a URL que você está interessado, clique em shorten. Vai aparecer o resultado e, abaixo dele, o número de "cliques", "saves" e "shares". Clicando na linha com essas informações, você cairá na página com as estatísticas completas. Essa combinação de dados de tráfego e conversações é muito útil quando estou tentando entender o porquê de um site ser tão popular, e quais são exatamente seus fãs.

Twitter

Quanto mais gente usa o serviço de micro-blogging, mais ele se torna útil como um indicador de como as pessoas estão compartilhando e falando sobre pedaços individuais de conteúdo. É extremamente simples descobrir conversas públicas sobre um link. Você apenas informa na caixa de busca a URL em que está interessado e clica em "more tweets" para visualizar o conjunto completo de resultados.

Cache do Google

Quando uma página se torna controversa, os publicadores podem despublicá-la ou alterá-la sem aviso. Se você suspeitar desse problema, recorra ao cache da página no Google. Ele mostra o site quando do último acesso feito pelo Google para indexação. Como a frequência com que o

Google faz esses acessos aumenta constantemente, suas chances são maiores se você tentar poucas horas após as suspeitas de mudança. Faça normalmente uma busca no Google com a URL desejada e, se você tiver sorte, haverá um link com o texto "cache" ao lado do título. Se ocorrer algum problema no carregamento quando você clicar, você pode alternar para uma versão somente com o texto da página. Copie e cole ou capture uma imagem da página rapidamente para guardar o estado anterior antes que o Google o atualize.

O Internet Archive da Wayback Machine

Se precisar saber como uma página foi alterada há meses ou anos, o Internet Archive possui um serviço chamado The Wayback Machine que periodicamente tira "fotos" de páginas da web mais populares. Se houver qualquer cópia da URL no site, ele mostrará um calendário para que você possa escolher a data a examinar. Em seguida apresentará uma versão da página mais ou menos como ela estava naquela data. Algumas vezes a página poderá estar sem estilos ou imagens, mas no geral dá para entender o foco do conteúdo da página na ocasião.

Visualizar o código fonte

Não são grandes as chances, mas geralmente desenvolvedores deixam comentários e outras dicas no código HTML de uma página. Para ver isso, você deve acionar a opção "Visualizar o código fonte" que exibirá o código HTML da página. Você não precisa entender o significado do código em si, apenas mantenha os olhos nos pedaços de texto espalhados. Mesmo que haja apenas notas de direito de cópia ou menções ao nome do autor, geralmente isso pode dar dicas importantes sobre a criação ou propósito da página.

TinEye

Às vezes você precisa saber a origem de uma imagem, mas sem uma legenda clara não existe uma maneira óbvia de se fazer isso com os mecanismos de busca tradicionais, como o Google. TinEye oferece um processo de "busca reversa", onde você fornece a imagem e ele encontra outras na web que parecem similares. Como ele usa reconhecimento de imagens para fazer a correspondência, isso funciona mesmo quando a imagem foi cortada, distorcida ou comprimida. Isso pode ser

extremamente efetivo quando você suspeitar que a imagem que está sendo passada como original ou nova estiver sendo mal representada.

YouTube

Se você clicar no ícone Estatísticas no lado direito inferior de qualquer vídeo, você pode obter um conjunto rico de informações sobre a audiência ao longo do tempo. Apesar de não mostrar tudo, é útil para entender em linhas gerais quem é a audiência daquele vídeo, de onde ela vem e quando.

Emails

Se você está pesquisando emails, vai querer saber mais detalhes sobre a identidade do remetente e sua localização. Não existe uma ferramenta pronta para ajudá-lo com isso, mas pode ser muito útil conhecer os campos escondidos no cabeçalho de toda mensagem. Eles funcionam como carimbos e podem revelar uma certa quantidade de informações sobre o remetente. Em particular, geralmente eles incluem o endereço IP da máquina da qual a mensagem foi enviada. Você pode em seguida usar o whois naquele endereço IP e descobrir que organização é dona daquela máquina. Se for alguma organização como a Comcast ou AT&T que fornecem conexões a consumidores, então você poderá visitar o MaxMind e obter sua localização aproximada.

Para ver esses cabeçalhos no Gmail, abra a mensagem e abra o menu próximo a "responder" no topo direito e escolha a opção "Mostrar original".

Você verá então uma nova página revelando o conteúdo escondido. Haverá algumas dezenas de linhas no início que são palavras seguidas por dois pontos. O endereço IP que você procura estará em um delas, mas o seu nome dependerá de como a mensagem foi enviada. Se for através do Hotmail, ela se chamará `X-Originating-IP:`, mas se for através do Outlook ou Yahoo será a primeira linha começando com `Received:`.

Buscando o endereço pelo Whois, por exemplo, ele me diz estar associado à Virgin Media, um Provedor de Acesso à Internet no Reino Unido, então eu uso o serviço de geolocalização do MaxMind para descobrir que ele está vindo da minha cidade natal

de Cambridge. Isso diz que estou razoavelmente confiante que de fato foram os meus pais que enviaram um email e não impostores.

Tendências

Se você estiver vasculhando um tópico mais abrangente, aqui estão algumas ferramentas que podem ajudá-lo com insights:

Wikipedia Article Traffic

Para saber como o interesse público em um tópico ou pessoa variou ao longo do tempo, você pode visualizar diariamente os números em qualquer página no Wikipedia em stats.grok.se. O site é um pouco grosseiro, mas é objetivo em mostrar a informação que precisa. Entre o nome que você está interessado para obter uma visão de tráfego mensal sobre a página. Isso exibirá um gráfico que mostra quantas páginas foram visualizadas cada dia do mês que você especificar. Infelizmente você apenas poderá ver um mês por vez, então terá que selecionar um novo mês e pesquisar de novo para ver mudanças de longo prazo.

Google Insights

Você pode ter uma visão clara dos hábitos de busca do público usando o Insights do Google. Escreva um conjunto de frases comuns de busca, como "Justin Bieber vs Lady Gaga", e você verá um gráfico do número relativo de buscas ao longo do tempo. Há várias opções para refinar a visualização de dados, desde restringir por áreas geográficas até obter mais detalhes sobre o tempo de acesso. A única desvantagem é a falta de valores absolutos — você obtém apenas porcentagens relativas, que podem ser difíceis de interpretar.

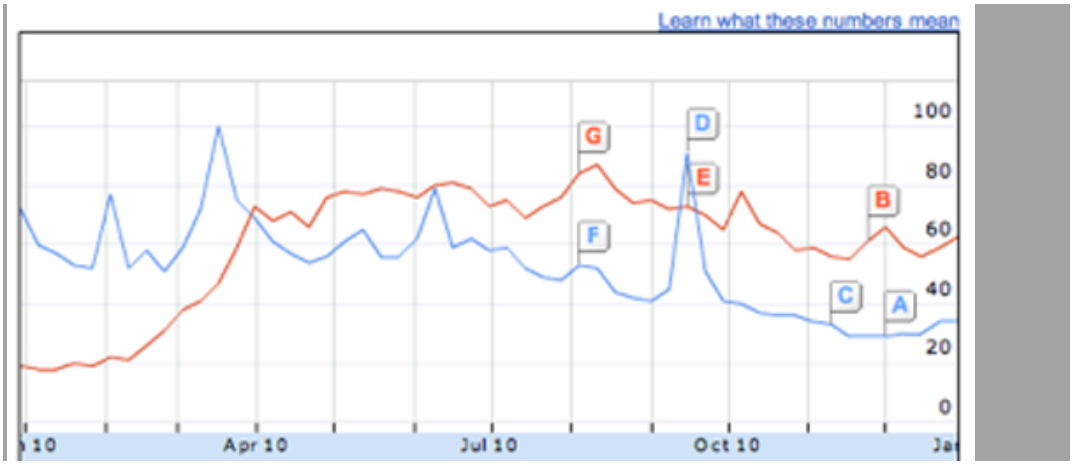


Imagem 11. Google Insights (Google)

O Crowdsourcing no Guardian Datablog

Crowdsourcing, de acordo com a Wipedia, é "um modelo de produção que utiliza a inteligência e os conhecimentos coletivos e voluntários espalhados pela internet para resolver problemas, criar conteúdo e soluções ou desenvolver novas tecnologias, assim como também para gerar fluxo de informação". O texto abaixo foi retirado de uma entrevista de Simon Rogers, onde ele conta como o Datablog usou o crowdsourcing na cobertura de escândalos com gastos de parlamentares, uso de drogas, e no caso dos documentos de Sarah Palin:

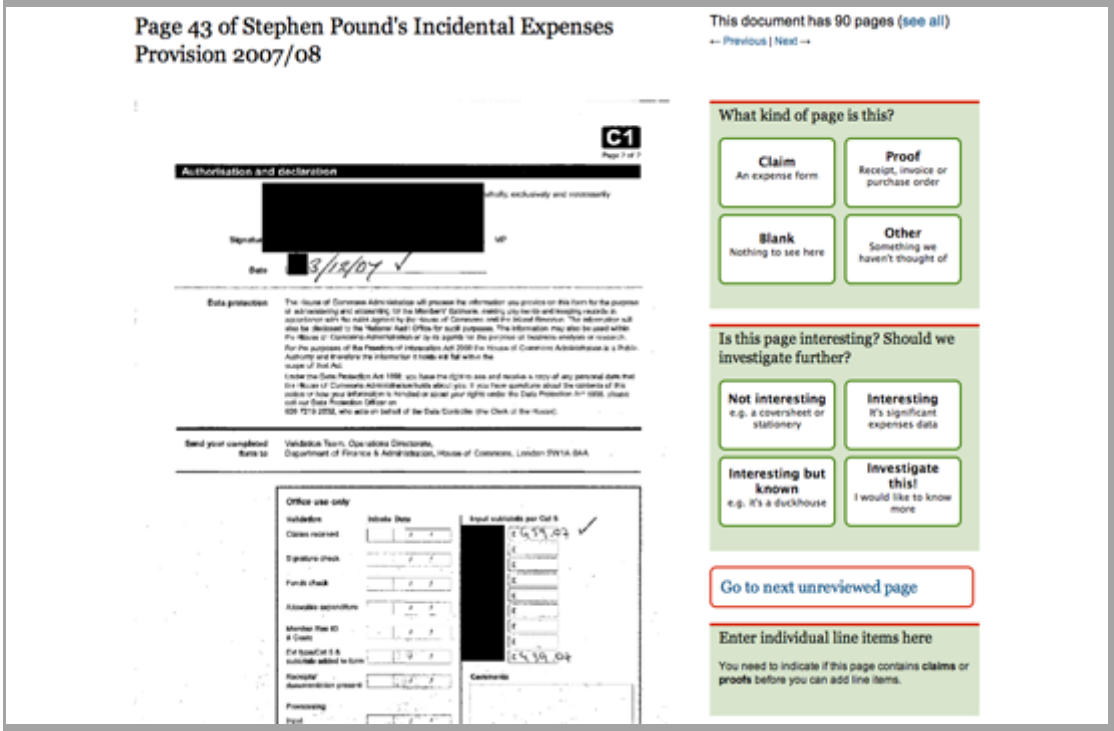


Imagem 12. Cópia já editada das despesas extras de Stephen Pound (the Guardian)

Algumas vezes, o repórter coleta um volume gigantesco de arquivos, estatísticas ou relatórios, tornando impossível que apenas uma pessoa processe tudo. Além disso, você pode encontrar material inacessível ou em formatos ruins que impossibilitem fazer muito com eles. É nestes casos que a colaboração em massa do crowdsourcing pode ajudar.

Uma coisa que o Guardian tem é muitos leitores, um monte de olhos atentos. Se há um projeto interessante no qual precisamos de muita leitura, então podemos chamá-los para nos ajudar. Foi o que fizemos com os Escândalo dos Gastos dos Membros do Parlamento. Nós tínhamos 450 mil documentos e pouquíssimo tempo para fazer qualquer coisa. Então, o que poderia ser melhor do que abrir a tarefa para nossos leitores?

O projeto dos Gastos Parlamentares gerou muitas delações de parlamentares. Mais até do que as histórias relacionadas aos dados originais. A ideia foi um sucesso incrível em termos de audiência. As pessoas realmente gostaram.

Nós estamos trabalhando com a revista de música MixMag sobre o uso de drogas, o que também tem sido fenomenal. Parece que será melhor do que a Pesquisa sobre Crime no Reino Unido em termos de quantas pessoas voltam, e isso é ótimo.

O que esses dois projetos possuem em comum é a ligação a questões com as quais as pessoas realmente se importam. Por isso, querem dedicar tempo a eles. Muitos dos nossos crowdsourcings dependem do trabalho de gente obcecada pelos temas. Com o caso dos Gastos dos parlamentares, tivemos um tráfego gigantesco no começo, mas a audiência foi morrendo. Mas ainda temos gente obssessiva repassando cada página disponibilizada e procurando por coisas erradas e notícias. Um único colaborador analisou 30 mil páginas. Eles sabem muita coisa.

Nós também usamos o recurso do crowdsourcing com os Documentos de Sarah Palin. Novamente, a colaboração em massa foi de grande ajuda para limpeza dos dados e para garimpar histórias.

Em termos de geração de conteúdo, o crowdsourcing funcionou realmente bem para nós. As pessoas gostaram e isso fez bem à imagem do Guardian. No entanto, para a extração de dados, nós não usamos tanto o recurso colaborativo.

Alguns dos projetos de crowdsourcing que fizemos e tiveram mais sucesso eram muito parecidos com enquetes tradicionais. Quando você pergunta para as pessoas sobre a experiência delas, sobre a vida delas, ou sobre o que estão fazendo, a colaboração é maior porque o público provavelmente não vai inventar nada disso. Eles vão falar o que sentem. Quando, porém, vamos pedir às pessoas praticamente para que façam nosso trabalho, precisamos achar uma estrutura de produção que permita confiar no que elas informaram.

Com relação à confiabilidade dos dados, acho que a fórmula do Old Weather é muito boa. Eles têm dez pessoas para cada entrada. É uma boa forma de garantir a precisão. No caso dos Gastos dos Deputados, tentamos minimizar o risco dos próprios parlamentares entrarem e editarem os conteúdos deles, para que parecessem melhor. Mas não é possível proteger-se continuamente contra esse tipo de ataque. Só podíamos conferir certas urls e verificar de onde vinham. Então, é um pouco mais complicado. Os dados que fomos extraindo não eram

sempre confiáveis. As matérias eram ótimas, mas não se produziam números brutos que podíamos usar com confiança.

Se tivesse que dar um conselho para os aspirantes ao jornalismo de dados que querem usar o crowdsourcing, encorajaria a fazer isso apenas nos casos em que os leitores realmente se importam e continuarão se importando, mesmo quando os dados pararem de render chamadas na capa. Além disso, se você torna a coisa parecida com um jogo, pode encorajar as pessoas a colaborar. Quando fizemos uma segunda vez o crowdsourcing sobre as despesas dos parlamentares, era como uma competição, com tarefas individuais para as pessoas cumprirem. Dar missões específicas a cada um realmente ajudou. E fez uma diferença enorme porque acho que, se você apenas apresenta para as pessoas uma montanha de informações e diz "cavoque isso", elas vão achar uma tarefa difícil e ingrata. Então, acho que tornar isso divertido é realmente importante.

— Marianne Bouchart, do Data Journalism Blog, entrevistando Simon Rogers, do the Guardian

Como o Datablog usou crowdsourcing para cobrir a compra de ingressos na Olimpíada

Creio que o projeto de crowdsourcing com a maior resposta foi um trabalho sobre a venda de ingressos na Olimpíada. Milhares de pessoas no Reino Unido tentaram comprar entradas para os Jogos de 2012 e houve muita indignação quando elas não os receberam. Teve gente que encomendou centenas de libras em ingressos e foi informada que não conseguiu nenhuma entrada. Mas ninguém sabia realmente se eram apenas alguns fazendo muito barulho quando, na verdade, uma maioria estaria satisfeita. Tentamos, então, encontrar uma maneira de descobrir isso.

Decidimos que, devido à ausência de dados de qualidade sobre o assunto, o melhor a fazer era perguntar. E achamos que não deveríamos tratar o assunto como algo muito sério, porque não teríamos uma amostragem representativa.

Criamos um formulário no Google e fizemos perguntas bem específicas. Era um longo formulário: perguntava o valor total da compra, quanto tinha sido debitado em seus cartões de crédito, para quais eventos eles pediram entradas, esse tipo de coisa.

How many Olympic tickets did you get? Here's our readers' results

We asked how you had fared in the London 2012 ticket ballot. Here's our analysis of the information you gave us

Tweet 21 Share 84 reddit this Comments (144)

The London 2012 Olympic stadium: will you be there? Will anyone you know? Photograph: Handout/Getty Images

How much money did London 2012 ticket buyers have to put on the line to stand even a 50:50 chance of getting at least one ticket? If the results submitted by Guardian readers are to be believed, at least £1,000.

Earlier this week, we asked readers of the Guardian London 2012 blog to let us know how their ticket-purchasing attempts had fared. By the end of the day – when this analysis was carried out – we'd had more than 5,000 responses.

Posted by James Ball Friday 3 June 2011 11.04 BST guardian.co.uk Article history

A larger | smaller

Sport Olympic tickets · Olympic Games 2012

More from London 2012 Olympics blog on

Sport Olympic tickets · Olympic Games 2012

More blogposts

Imagem 13. Quantos ingressos olímpicos você conseguiu?: os resultados dos leitores (The Guardian)

Colocamos uma pequena figura na página inicial do site e o formulário foi difundido rapidamente. Essa é uma das questões-chave: você não pode simplesmente pensar "O que eu quero saber para a minha reportagem?", você deve pensar "O que as pessoas querem me contar?". Só quando você acerta o que as pessoas querem falar naquele momento é que o crowdsourcing terá sucesso. O volume de respostas para este projeto, uma de nossas primeiras tentativas em crowdsourcing, foi imenso. Tivemos mil respostas em menos de uma hora e sete mil ao final do dia.

Portanto, obviamente, a esta altura levamos um pouco mais a sério a apresentação dos resultados. Inicialmente, não tínhamos ideia de quão satisfatório ele seria. Então, acrescentamos algumas considerações: os leitores do Guardian podem ser mais ricos que as outras pessoas, quem conseguiu

menos ingressos do que esperava pode estar mais disposto a falar com a gente, e coisas do tipo.

Não sabíamos qual seria o resultado. Descobrimos que cerca de metade das 7 mil pessoas que encomendaram ingressos e entraram em contato conosco não receberam nenhum. Apresentamos todas essas informações, e, porque muita gente havia participado na véspera, houve muito interesse nos resultados.

Semanas depois, o relatório oficial foi divulgado, e nossos números eram impressionantemente próximos. Eram quase exatos. Imagino que em parte por uma questão de sorte, mas também porque nós conseguimos que tanta gente nos respondesse.

Se você pergunta aos leitores sobre algo deste tipo no espaço de comentários do texto, estará limitado sobre o que poderá fazer com as respostas. Então, deve-se começar a pensar: "Qual a melhor ferramenta para o que eu quero saber?". É um espaço para comentários? Ou é construir um aplicativo? E se for construir um aplicativo, deve-se pensar: "Valerá a demora? E valerá investir os recursos necessários para fazê-lo?"

Neste caso lembramos dos formulários do Google, o Google Forms. Se alguém preenche o formulário, o resultado pode ser visto como uma linha em uma tabela. Mesmo se os dados ainda estiverem chegando, é possível abrir uma tabela e ver todos os resultados imediatamente.

Eu poderia ter tentado fazer o trabalho no Google, mas eu o baixei no Microsoft Excel e então fiz coisas como organizar do menor valor para o maior; também descobri espaços onde as pessoas escreveram os números (ao invés de pôr apenas os dígitos) do quanto eles gastaram e consertei tudo isso. Algumas pessoas usaram outras moedas, e as converti em libras. Tentei levar em conta todos os resultados e, em vez de excluir os inválidos, eu os arrumei — o que deu bastante trabalho.

Mas toda a análise ficou pronta em algumas horas, e eu descartei as respostas claramente tolas. Muitas pessoas decidiram mostrar que não tinham gasto nada com ingressos. O que pode parecer um pouco engraçado, mas tudo bem. Foram menos de cem respostas deste tipo em um total de mais de sete mil.

Também houve algumas dezenas que cadastraram cifras elevadas claramente falsas para tentar distorcer os resultados. Coisas como dez milhões de libras. Então, essa limpeza me deixou com um conjunto de dados com os quais eu

poderia trabalhar usando os princípios normais que usamos todos os dias. Eu fiz uma tabela dinâmica e calculei algumas médias. Este tipo de coisa.

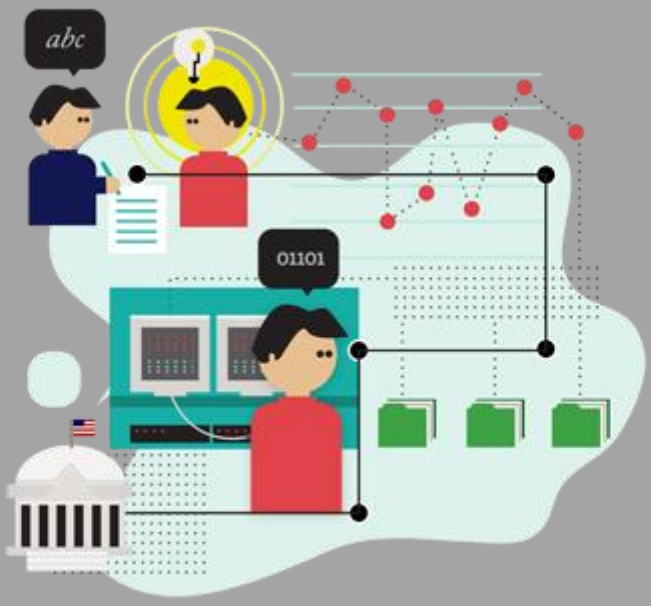
Nós não tínhamos nenhuma ideia das possibilidades do projeto, então éramos apenas eu e o editor do blog de Esportes trabalhando nisso. Juntamos nossas ideias e imaginamos que poderia ser divertido. Nós o fizemos, do começo ao fim, em 24 horas. Tivemos a ideia, bolamos alguma coisa na hora do almoço, colocamos no site, vimos que estava se tornando bem popular, o mantivemos na página de abertura do site o resto do dia, e apresentamos os resultados online na manhã seguinte.

Decidimos usar Google Docs porque ele te dá controle completo sobre os resultados. Eu não tive que usar as ferramentas de análise de mais ninguém. Posso colocá-los facilmente em um software de banco de dados ou em planilhas. Quando você usa programas especiais para pesquisas, geralmente fica restrito às ferramentas deles. Se fôssemos perguntar algo mais delicado, talvez tivéssemos pensado em programar. Mas geralmente é muito fácil pôr um formulário do Google no site do Guardian, e para o usuário é praticamente invisível o fato de estarmos usando tal formulário.

Nosso conselho sobre crowdsourcing é que você precisa querer perguntar coisas bem específicas. Faça questões de "múltipla escolha" tanto quanto possível. Tente obter alguns dados demográficos básicos daqueles a quem são dirigidas as questões, para checar se sua amostra não estará enviesada. Se você está perguntando sobre quantidades, tente especificar que o preenchimento deve ser feito com dígitos, em uma moeda específica, e detalhes como esse. Muitos não o farão, mas quanto mais você os guiar, melhor. E sempre, sempre, acrescente um espaço para comentários, porque muitas pessoas vão preencher os campos mas o que elas realmente querem é dar sua opinião sobre o caso. Principalmente numa reportagem que afetam consumidores ou sobre pessoas que estão injuriadas.

— Marianne Bouchart, Data Journalism Blog, entrevistando James Ball, The Guardian

Entendendo os Dados



O que fazer com os dados depois de consegui-los? Quais ferramentas usar? Esta seção pretende aprimorar seu conhecimento no tema, com dicas para trabalhar com números e estatísticas, e ideias para trabalhar com conjuntos de dados desorganizados, imperfeitos e em situação irregular. Vamos aprender como extrair histórias a partir de dados, ver quais são as melhores ferramentas, e como usar a visualização para conseguir insights sobre o tópico que você está analisando.

O que há neste capítulo?

Usando e compartilhando dados: a letra da lei, a letra miúda e a realidade

Nesta seção, vamos falar brevemente sobre a lei em relação a bancos de dados, além de explicar como você pode abrir seus dados usando licenças públicas já disponíveis e mecanismos legais. Não deixe nada do que vem a seguir diminuir seu entusiasmo pelo jornalismo guiado por dados. Restrições legais sobre dados na maior parte das vezes não vão atrapalhar e é fácil garantir que elas não sejam um obstáculo para que outros usem as informações que você publicou.

Obter dados nunca foi tão fácil. Antes da publicação generalizada de informações na web, mesmo que você houvesse identificado uma série de dados da qual precisasse, seria necessário pedir a quem quer que tivesse uma cópia para torná-la acessível a você, possivelmente usando papel, correios, ou uma visita pessoal. Agora, você faz seu computador pedir ao computador deles para lhe enviar uma cópia. Conceitualmente similar, mas agora o criador e quem divulgou os dados provavelmente não fazem ideia de que você fez o download de uma cópia.

E quanto ao download de dados por meio de um programa de captura de informações do site ("scraping"), os termos de serviço proíbem isso? Pense bem: seu navegador é só um programa. Será que os termos de serviço iriam permitir acesso aos dados apenas por determinados tipos de programas? Se você tem tempo livre e dinheiro o suficiente para ler documentação sobre isso no site e talvez pedir conselhos a um advogado, por favor, faça. Mas, no geral, apenas evite ser um babaca: se o seu programa sobrecarrega um site, sua rede pode muito bem ser impedida de acessá-lo — e você talvez mereça isso. Existe hoje um grande volume de boas práticas quanto a acessar e retirar dados por scraping da Web. Se planeja fazer isso, examinar exemplos em sites como o ScraperWiki vai ajudá-lo no início.

Uma vez que você tenha obtido os dados, você pode indagar, examinar, classificar, visualizar, correlacionar e realizar qualquer tipo de análise com sua cópia. Você pode publicar sua análise, citando a fonte. Você pode recorrer à palavra-de-ordem "fatos são livres" , mas talvez essa ideia só pegue de verdade entre aqueles que valorizam demais os direitos relacionados aos bancos de dados.

Mas e se você, sendo um bom ou candidato a bom jornalista de dados, pretende publicar não apenas a sua análise mas também as bases de dados que você

usou — e, quem sabe, complementou — durante a realização de sua análise? Ou talvez você esteja apenas fazendo curadoria dos dados e não tenha feito análise alguma (muito bom: o mundo precisa de curadores de dados). Se você está usando dados coletados por alguma outra entidade, pode haver um obstáculo. (Se o seu banco de dados foi totalmente montado por você, leia o próximo parágrafo de qualquer forma, como motivação para adotar práticas de compartilhamento.)

Se o detentor dos direitos autorais não tiver dado permissão para usar uma obra (ou caso a obra esteja em domínio público, mas com limitações ao uso que você quer fazer dela) e mesmo assim você continuar, o proprietário do copyright pode forçá-lo a parar. Embora os "fatos sejam livres", é possível que coleções de fatos sejam restringidas por direitos autorais. Em muitos lugares, simplesmente reunir uma base de dados, mesmo sem originalidade alguma, torna o banco de dados suscetível às leis de direitos autorais. Nos Estados Unidos, em particular, a tendência é exigir um nível mínimo de criatividade para que o copyright seja aplicável (Feist vs. Rural, uma disputa judicial sobre uma agenda de telefones, é o clássico americano se você quiser entender melhor). Entretanto, em outros locais também há "direitos sobre dados" separados dos direitos autorais (ainda assim, há muita sobreposição entre as leis). A mais conhecida dessas leis são os direitos sobre dados*sui generis* da União Européia. Se você estiver na Europa, pode ser uma boa ideia se certificar da permissão antes de publicar uma base de dados de outra entidade.

Obviamente, esse tipo de restrição não é a melhor maneira de desenvolver um ecossistema de jornalismo de dados. Elas não são boas para a sociedade em geral — antes mesmo dessas regras *sui generis* vigorarem, cientistas sociais e intelectuais já alertavam à União Europeia sobre o malefício dessas restrições. Pesquisas desde então mostram que eles estavam certos. Felizmente, como proprietário de um banco de dados, você pode remover dos seus dados esse tipo de restrição (desde que não contenha elementos que você não tem permissão para licenciar), bastando para isso conceder uma permissão prévia. Você pode fazer isso através de uma licença pública ou atribuindo o banco de dados ao domínio público. Assim como programadores liberam seu código sob licenças de open source para que outros possam criar a partir deles, você, como jornalista de dados, deve permitir a reprodução de sua coleção de dados e da sua análise. Há muitas razões para isso. Por exemplo, a sua audiência pode criar novas visualizações ou aplicativos para os quais você pode linkar — como o The

Guardian faz com seu grupo de visualização de dados no Flickr. Suas bases de dados podem ser combinadas a outras para dar a você e a seus leitores uma compreensão maior sobre um tema. Coisas que os outros fazem com seus dados podem dar pistas para novas reportagens, ou ideias de pauta, ou ideias para outros projetos guiados por dados. E certamente vão lhe trazer aplausos.

Imagem 14. Insígnias Open Data (Open Knowledge Foundation)

Quando alguém se dá conta de que liberar obras sob licenças públicas é uma necessidade, a questão se torna "qual licença"? Esse dilema em geral é respondido pelo projeto ou comunidade sobre cujo material você está trabalhando, ou para a qual você espera contribuir — use a mesma licença que eles. Se você precisar se aprofundar, comece pelas licenças livres e abertas — isto é, nas quais qualquer um tem permissão para qualquer tipo de uso (atribuição de crédito e compartilhamento pela mesma licença podem ser colocados como condições). O que as definições de Software Livre e Software Open Source fazem pelos programas, a Open Knowledge Definition faz para todo o conhecimento, inclusive bancos de dados: define o que torna uma obra aberta e quais permissões as licenças dão aos usuários.

Você pode acessar o site da Open Knowledge Definition para verificar o conjunto de licenças que pode ser usado. Em resumo, há três categorias de licenças abertas:

Domínio público

Servem como o máximo de permissão; não há condições impostas à reutilização da obra.

Permissão apenas com atribuição

Atribuir crédito ao autor é a única condição substancial imposta por estas licenças.

Licenças recíprocas, copyleft ou de igual compartilhamento

Exigem que as obras modificadas, se publicadas, devem ser compartilhadas pela mesma licença.

Se você está usando uma base de dados publicada por outra pessoa sob uma licença aberta, considere o parágrafo acima um guia rápido sobre como atender às condições de licenciamento. As licenças mais comuns, de Creative Commons, Open Data Commons e de vários governos, em geral oferecem um resumo que lhe permitirá identificar facilmente quais são as condições principais. Tipicamente, a licença estará numa página de onde a base de dados pode ser baixada (ou capturada por scraping, porque, é claro, páginas da Web podem conter conjuntos de dados) ou em algum lugar dentro da própria planilha, dependendo do formato. São estas marcações que você deve fazer também, quando abrir seus bancos de dados.

Voltando ao início, e se o banco de dados que você busca ainda não estiver online, ou estiver protegido por algum tipo de controle de acesso? Ao solicitar acesso para si mesmo, pense em pedir que os dados sejam abertos para o mundo inteiro reutilizar. Você pode até mesmo citar alguns exemplos de coisas incríveis que podem acontecer com os dados se eles forem

liberados.

É bom lembrar que privacidade e outras considerações podem ser necessárias no caso de alguns bancos de dados. Só porque ter o "open data" elimina barreiras técnicas e outras relacionadas ao copyright, não significa que você não precise seguir outras leis que se aplicam àquele conteúdo. Mas, como sempre, há muitos recursos e algumas proteções para jornalistas, caso seu bom senso o leve a investigar bancos de dados mais controversos.

Boa sorte! Mas o mais provável é que você precise dessa sorte em outras áreas do seu projeto, não no gerenciamento dos (baixos) riscos jurídicos.

— Mike Linksvayer, Creative Commons

Familiarizando-se com os dados em três passos

Da mesma forma como alfabetização refere-se à "habilidade de ler para obter conhecimento, escrever coerentemente, e pensar criticamente", ser alfabetizado em dados é ter a habilidade de consumir dados para o conhecimento, produzi-los de forma coerente, e pensá-los de forma crítica. Não é só entender de estatística, mas também sobre como trabalhar com grandes conjuntos de dados, como produzi-los, como conectar várias bases de dados e como interpretá-los.

Imagem 1. Aprofundando-se nos dados (foto: JDHancock)

A Poynter's News University oferece aulas de matemática on line para jornalistas, nas quais os repórteres recebem ajuda com conceitos como variações percentuais e médias. Curiosamente, esses conceitos são ensinados também perto das salas da Poynter, nas escolas da Flórida a alunos entre 10 e 11 anos, como mostra o currículo.

O fato de jornalistas precisarem de ajuda em questões relacionadas a um conteúdo ensinado antes do ensino médio mostra como as redações estão longe de serem alfabetizadas em dados. Isso é um problema. Como um jornalista de dados pode usar uma grande quantidade de dados sobre a mudança climática se não sabe o que significa "intervalo de confiança"? Como um repórter de dados

pode escrever um artigo sobre distribuição de renda se não consegue diferenciar média de mediana?

Um repórter não precisa ser graduado em estatística para se tornar mais eficiente com dados. Alguns poucos truques podem ajudar a conseguir um artigo muito melhor. Como diz o professor do Instituto Max Planck Gerd Gigerenzer, melhores ferramentas não vão levar à um melhor jornalismo se usadas sem conhecimento do assunto.

Mesmo se você não tem conhecimentos de matemática ou estatística, você pode se transformar facilmente em um jornalista de dados ocasional ao fazer 3 simples perguntas.

1. Como os dados foram coletados?

Crescimento do PIB espetacular

O jeito mais fácil de mostrar dados espetaculares é fabricá-los. Soa óbvio, mas dados sobre o Produto Interno Bruto (PIB) podem ser bem enganadores. O ex-embaixador do Reuino Unido no Uzbequistão Craig Murray diz em seu livro *Murder in Samarkand* que as taxas de crescimento no país asiático são objeto de intensas negociações entre o governo local e grupos internacionais. Ou seja, nada têm a ver com a economia.

O PIB é usado como o indicador número um porque os governos precisam dele para supervisionar sua principal fonte de renda — os impostos sobre o consumo. Quando um governo não é financiado por essas taxas, ou quando seu orçamento não é público, não há razão para coletar dados de PIB e pode se dar melhor perante aos eleitores fabricando esses dados.

Criminalidade sempre está em ascensão

"O crime na Espanha cresceu 3%", escreve o El Pais. Bruxelas vê o aumento de crimes de imigrantes ilegais e viciados em drogas, diz o RTL. Esse tipo de relatório baseado em estatísticas policiais é comum, mas não revela muito sobre violência.

Podemos confiar que, dentro da União Europeia, os dados não costumam ser adulterados. Mas policiais respondem a incentivos. Quando a performance está vinculada a uma métrica baseada em crimes solucionados, por exemplo, os policiais são incentivados a reportar ao máximo incidentes que não exigem investigação, como fumar maconha, por exemplo. Com mais crimes como esse, "fáceis de solucionar", a performance dos oficiais fica parecendo melhor. Isso

explica porque crimes ligados a dependentes de drogas na França quadruplicaram nos últimos 15 anos, enquanto o consumo continua constante.

O que você pode fazer

Em caso de dúvida sobre a credibilidade de um número, sempre faça a checagem, assim como você costuma fazer com uma citação de um político. No caso do Uzbequistão, um telefonema para alguém que viveu no país por um tempo é suficiente ("Será que o país triplicou o PIB desde 1995, como os dados oficiais indicam?").

Para os dados policiais, os sociólogos muitas vezes fazem estudos de vitimização, nos quais perguntam às pessoas se elas foram vítimas de crime. Estes estudos são muito menos voláteis do que os dados da polícia. Provavelmente por isso eles não viram manchete.

Outros testes permitem avaliar com precisão a credibilidade dos dados, tais como a Lei de Benford (conceito estatístico sobre a probabilidade de aparição de um número), mas nenhum irá substituir o seu senso crítico.

2. O que se pode aprender com os dados?

Risco de esclerose múltipla duplica quando se trabalha à noite

Certamente qualquer alemão em seu perfeito juízo iria parar de trabalhar em turnos noturnos depois de ler esta manchete. Mas o artigo não nos diz qual é o risco no fim das contas.

Considere um grupo de mil alemães. Apenas um deles vai desenvolver esclerose múltipla ao longo de sua vida. Agora, se cada um desses 1.000 alemães trabalhassem durante a noite, o número de doentes de esclerose múltipla iria saltar para 2. O risco adicional de desenvolver esclerose múltipla quando se trabalha de noite é de 1 em 1.000, não é 100%. Certamente esta informação é mais útil quando se pondera a possibilidade de assumir um trabalho nesse horário.

Em média, 1 de cada 15 europeus é totalmente analfabeto

O título acima parece assustador. Também é absolutamente verdadeiro. Entre os 500 milhões de europeus, 36 milhões provavelmente não sabem ler. Vale notar que 36 milhões de europeus também têm menos de 7 anos (dados do Eurostat).

Ao escrever sobre uma média, sempre pense "média de quê?" A população de referência é homogênea? Padrões de distribuição desigual explicam por que a maioria das pessoas dirige melhor que a média, por exemplo. Muitas pessoas têm zero ou apenas um acidente durante a sua vida. Alguns motoristas imprudentes tem um grande número, empurrando a média do número de acidentes para cima em comparação com o que a maioria das pessoas experimenta. O mesmo é verdade para a distribuição de renda: a maioria das pessoas ganha menos que a média.

O que você pode fazer

Sempre leve em consideração a distribuição e a taxa básica. Verificar a média e a mediana (número que separa a metade inferior e superior da amostra), assim como a moda (o valor mais frequente na distribuição), ajuda você a obter insights. Saber a ordem de grandeza torna a contextualização mais fácil, como no exemplo a esclerose múltipla. Finalmente, falar em frequências naturais (1 em cada 100) é a maneira mais fácil para os leitores entenderem do que percentuais (1%).

3. Quão confiável é a informação?

O problema do tamanho da amostra

"80% estão insatisfeitos com o sistema judicial", diz uma pesquisa que saiu no jornal Diário de Navarra. Como é possível saltar de 800 entrevistados para 46 milhões de espanhóis? Certamente os dados estão inflados, não? Não.

Ao pesquisar uma grande população (mais de alguns milhares), você raramente precisa de mais de mil participantes para alcançar uma margem de erro inferior a 3%. Isso significa que, para cada 20 vezes que você refizesse a pesquisa, 19 apontariam um resultado 3 pontos percentuais acima ou abaixo da distribuição real daquilo na população.

Beber muito chá reduz o risco de Acidente Vascular Cerebral (AVC)

Artigos sobre os benefícios de beber chá são comuns. Este do diário alemão Die Welt, que diz que o chá reduz o risco de infarto do miocárdio, não é exceção. Embora os efeitos do chá são seriamente estudados por alguns, muitos estudos deixam de levar em conta fatores de estilo de vida, como dieta, profissão ou esportes.

Na maioria dos países, o chá é uma bebida para as classes mais ricas, normalmente mais conscientes sobre a saúde. Se os pesquisadores não

controlarem para fatores de estilo de vida nos estudos sobre chás, eles não permitem dizer nada mais do que "pessoas ricas são mais saudáveis -- e provavelmente bebem chá."

O que você pode fazer

A matemática por trás de correlações e margens de erro nesses estudos está correta, pelo menos na maior parte do tempo. Mas se os pesquisadores não procurarem correlações (por exemplo, beber chá correlaciona-se com a prática de esportes), seus resultados são de pouco valor.

Como um jornalista, faz pouco sentido desafiar os resultados numéricos de um estudo, como o tamanho da amostra, ao menos que haja sérias dúvidas sobre isso. Entretanto, é fácil ver se pesquisadores não levaram em conta pequenos fatores relevantes.

— Nicolas Kayser-Bril, Jornalismo++

Dicas para Trabalhar com Números

- A melhor dica para lidar com dados é divertir-se. Eles podem parecer assustadores. Mas deixar-se intimidar não leva a lugar nenhum. Trate-os como algo a descobrir e explorar e veja como eles vão revelar segredos e histórias com uma facilidade surpreendente. Pense nisso como um exercício de imaginação. Seja criativo e imagine histórias que poderiam ser explicadas por aqueles dados, e coloque-as a prova. Perguntar "que outra história poderia explicar esse fenômeno?" é um modo prático de descobrir como números grandes ou ruins podem ter uma outra explicação que não esteja relacionada com o que você procurava.
- Não confunda ceticismo com relação aos dados com o cinismo. O ceticismo é bom; o cinismo simplesmente lava as mãos e abandona a situação. Se você acredita no jornalismo de dados (e claro que acredita, senão não estaria lendo este livro), precisa acreditar que os dados têm algo melhor para oferecer que mentiras ou fatos distorcidos para as manchetes. Quando usados de forma cuidadosa, os dados sempre nos dão um profundo conhecimento. Não precisamos ser cínicos nem ingênuos, mas precisamos estar alertas.
- Se eu disser que, durante a recessão, aumentou o consumo de álcool, você poderia me responder que é porque todo mundo está em depressão. Se eu disser que diminuiu, é porque todos estão quebrados. Em outras palavras, o que os números dizem não faz diferença para a interpretação que você está determinado a fazer — normalmente que as coisas são terríveis. O fato aqui é que, se você acreditar nos dados, tente deixá-los falar antes de submetê-los ao seu estado de espírito, crenças ou expectativas. Há tantos dados sobre tudo que, com frequência, você vai ser capaz de encontrar algo que confirme sua crença, seja ela qual for. Em outras palavras, o jornalismo de dados, pelo menos para mim, acrescenta pouco se você não tiver a mente aberta. Ele só será objetivo se você se esforçar para isso, e não simplesmente porque você está se baseando em números.
- Incerteza é OK. Nós damos aos números uma carga de autoridade e certeza. Mas, muitas vezes, a resposta é que não há resposta, ou que o resultado que temos é muito inexato. Penso que deveríamos assumir isso. Se soa como uma boa maneira de matar reportagens, eu diria que é uma ótima maneira de levantar novas questões. Com frequência há mais de uma forma legítima

de se interpretar os dados. Os números não têm que ser nem verdadeiros nem falsos.

- A investigação é uma matéria. O relato sobre como você tentou fazer a sua descoberta pode ser uma ótima peça jornalísitica, mostrando como você foi de uma prova à outra — e isso se aplica às evidências dos dados, nas quais é raro um número ser suficiente. Fontes diferentes oferecem novos ângulos, novas ideias e uma compreensão mais completa. Me pergunto se não estamos muito presos a uma vontade de nos mostrar como autoridades e apresentar uma resposta ao público — e, assim, deixamos passar a oportunidade de nos mostrar como detetives.
- As melhores perguntas são as mais antigas: este número é mesmo importante? De onde vem? São formas de refletir sobre os dados: o que fica de lado quando olhamos apenas um único número, as complicações da vida real, outras comparações, agrupamentos ou divisões geográficas que podem ser feitos. Em resumo, o contexto.

— Michael Blastland, jornalista freelance

Primeiros passos para trabalhar com dados

Pelo menos três conceitos fundamentais devem ser levados em conta na hora de iniciar um projetos de dados:

- A coleta dos dados deve começar com uma lista de perguntas que você quer responder.
- Dados muitas vezes estão bagunçados e precisam ser limpos.
- Bases de dados podem ter elementos não documentados.

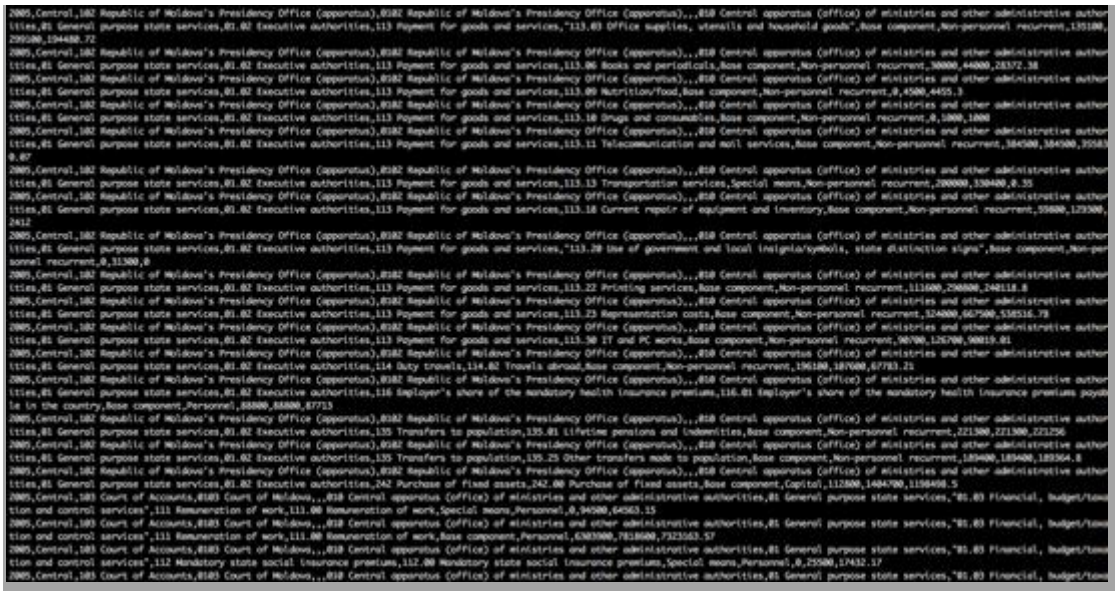


Imagem 2. Dados bagunçados

Saiba quais perguntas você quer responder

De certa forma, trabalhar com dados é como entrevistar uma pessoa. Você faz perguntas e faz com que eles revelem as respostas. Assim como uma fonte só pode informar sobre o que conhece, bases de dados só podem responder perguntas para as quais têm os registros certos e as variáveis adequadas. Isso significa que você deve avaliar com cuidado as perguntas para as quais busca resposta antes mesmo de coletar os dados. Basicamente, o trabalho é feito de trás para frente. Primeiro, liste o que você pretende demostrar em sua reportagem usando dados. Depois, decida quais variáveis você deve coletar e analisar para chegar a esse resultado.

Digamos que você queira fazer uma reportagem sobre o padrões de criminalidade em sua cidade abordando as horas e os dias em que são mais frequentes, assim como os locais onde cada tipo de delito acontece mais.

Você vai se dar conta que sua coleta de dados deve incluir o dia e horário que cada crime foi cometido, o tipo de delito (homicídio, furto, roubo, etc.), assim como o local em que ocorreu. Então, data, hora, tipo de delito e endereço são as variáveis mínimas que você precisa para responder suas perguntas.

Veja, porém, que há diversas questões potencialmente interessantes que essas quatro variáveis *não poderão* responder, como a cor e o gênero das vítimas, o valor roubado, ou quais policiais prendem o maior número de criminosos. Além disso, só será possível conseguir registros abrangendo um determinado período, como os últimos três anos. Isso significa que não será possível afirmar se os padrões de criminalidade mudaram ao longo da década, por exemplo. Essas perguntas podem estar fora do foco da matéria, e tudo bem. O que não dá certo é, durante a análise dos dados, decidir de repente que é necessário saber qual a porcentagem de crimes que resulta em prisões em cada parte da cidade.

Uma lição aqui é que normalmente é uma boa ideia pedir *todos* os registros e variáveis de um conjunto de dados, em vez de um recorte que permita apenas responder às perguntas imediatas da reportagem. (Na verdade, conseguir todos os dados pode ser mais barato do que um subconjunto, caso seja necessário pagar pelo trabalho de programação para selecionar uma parte específica.) Além disso, sempre é possível fazer o recorte por conta própria e ter o conteúdo completo permite responder novas questões que possam surgir e pode render novas ideias ou suítes para as matérias. Pode acontecer de a divulgação de algumas variáveis ser proibida por leis de confidencialidade, como nome de vítimas ou informantes. Ainda assim, uma base de dados parcial é melhor que nenhuma, desde que se saiba quais perguntas ela pode ou não responder.

Limpando Dados

Um dos maiores problemas do trabalho com bases de dados é que, frequentemente, as informações foram coletadas com propósitos burocráticos. O problema é que os padrões de exatidão e precisão adotados para cada finalidade são bastante diferentes.

Por exemplo, uma das principais funções da base de dados da justiça criminal é garantir que o réu João seja levado da cadeia ao juiz Silva na hora da audiência. Para alcançar esse objetivo, não importa muito se a data de nascimento do João está trocada ou se o endereço está escrito errado. De maneira geral, o sistema ainda pode usar esse registro imperfeito para levar o João para a corte do juiz Silva na hora marcada.

Mas erros assim podem atrapalhar tentativas de descobrir os padrões daquela base de dados. Por esse motivo, o primeiro grande esforço a ser feito após conseguir um conjunto de dados é examinar o quão bagunçado está e organizá-lo. Uma forma rápida e boa de fazer isso é criar tabelas de frequência das categorias, variáveis das quais se espera que haja um número pequeno de valores diferentes. (No Excel, por exemplo, isso pode ser feito usando filtros ou tabelas dinâmicas para cada categoria.)

Um exemplo simples é a categoria "gênero". Nesse campo pode haver uma mistura de valores como esses: Masculino, Feminino, M, F, 1, 0, MASC, FEM, etc., incluindo erros ortográficos, como "Femenino". Para fazer uma análise correta, é necessário adotar um padrão — como M ou F, talvez — e então mudar todas as variações para se adequar a essa norma. Outra base de dados que apresenta esse problema é a de financiamento de campanha dos Estados Unidos, que no campo Ocupação pode listar "Lawyer", "Attorney", "Atty", "Counsel", "Trial Lawyer", e uma série de variações do vocabulário norte-americano para definir advogados, além de erros de ortografia; mais uma vez, a saída é padronizar as ocupações, adotando uma lista menor de possibilidades.

A organização dos dados se torna ainda mais problemática quando se trabalha com nomes. Seriam "João Carlos Silva'*, "João Silva" "J.C. Silva" e "J. Carlos Silva*' a mesma pessoa? Talvez seja necessário verificar outras variáveis, como endereço e data de nascimento ou mesmo uma pesquisa mais aprofundada em outros registros. Ferramentas como o Google Refine, porém, podem tornar a arrumação dos dados mais rápida e menos tediosa.

Dados com sujeira

Graças a uma forte legislação para garantir o acesso a informações públicas nos Estados Unidos, conseguir dados aqui não é um problema tão grande como em outros países. Mas ainda precisamos enfrentar dados coletados com propósitos burocráticos e não analíticos. Frequentemente estão "sujos", ou seja, com valores não padronizados. Várias vezes recebi dados que não correspondiam ao suposto formato de arquivo e ao dicionário de dados que os acompanhavam. Algumas agências insistem em enviar dados em formatos inadequados como .pdf, que precisam ser convertidos. Problemas como esses fazem com que você fique grato quando recebe um conjunto de dados sem complicações.

— Steve Doig, Walter Cronkite School of Journalism, Arizona State University

Bases de dados podem ter elementos não documentados

A Pedra de Rosetta de qualquer base de dados é o chamado dicionário de dados. Tipicamente, esse arquivo (pode ser texto ou PDF ou mesmo uma planilha) dirá como os dados estão formatados, a ordem das variáveis, os nomes de cada variável e os tipos de dados de cada variável. Essa informação será usada para importar o arquivo de dados para o software de análise pretendido (Excel, Access, SPSS, Fusion Tables, qualquer das diversas variações de SQL, etc.)

Outro elemento fundamental do dicionário de dados é a explicação dos códigos usados pelas variáveis. Por exemplo, gênero pode ser codificado de forma que "1=Masculino" e "0=Feminino." Crimes podem ser codificados de acordo com os números oficiais de identificação de cada tipo de delito. Registros de tratamentos hospitalares podem usar qualquer uma das centenas de combinações padronizadas de 5 dígitos que identificam os diagnósticos do paciente. Sem o dicionário de dados, pode ser difícil ou até impossível analisar propriamente essas bases.

No entanto, mesmo com o dicionário de dados à mão, pode haver problemas. Foi o que aconteceu com repórteres do Miami Herald, na Flórida, quando faziam uma análise dos diferentes graus de punição aplicados por juízes a motoristas embriagados. Os jornalistas conseguiram os registros de condenações do sistema judicial e analisaram os números segundo três diferentes variáveis de sentença constantes do dicionário de dados: quantidade de tempo na prisão estadual ou federal, quantidade de tempo na cadeia local e valor da multa. Os números variavam bastante de acordo com o juiz, dando aos repórteres evidências para uma matéria sobre como alguns magistrados eram mais duros e outros lenientes.

Para cada juiz, porém, entre 1% e 2% dos casos não mostravam tempo de prisão, de cadeia ou valor de multa. Então, os gráficos mostrando os padrões de sentença mostravam uma pequena parcela de casos identificados como "sem punição". Quando a matéria e os gráficos foram publicados houve grita dos juízes que diziam que o jornal os estava acusando de desrespeitar a lei estadual que estabelecia como obrigatória a punição de motoristas embriagados.

Os repórteres voltaram então para o órgão do tribunal que havia produzido o arquivo com os dados e perguntaram o que havia causado o erro. Eles foram informados que os casos em questão envolvem réus primários indigentes. Normalmente, seriam multados, mas eles não tinham dinheiro. Por isso, os

juízes os sentenciavam a serviços comunitários, como limpar lixo nas ruas. No fim, souberam que a lei que obrigava os juízes a punir motoristas bêbados foi promulgada depois que a estrutura das bases de dados foi criada. Dessa forma, todos os funcionários do tribunal sabiam que, naqueles dados, os zeros nas variáveis prisão-cadeia-multa significavam serviços comunitários. No entanto, isso *não foi* apontado no dicionário de dados, e portanto obrigou o Herald a fazer uma correção.

A lição aprendida nesse caso é sempre perguntar a quem forneceu os dados se há algum elemento não documentado, sejam novos códigos ainda não incluídos no dicionários de dados, mudanças no formato, ou qualquer outra coisa. Além disso, sempre examine os resultados de sua análise e pergunte "Isso faz sentido?" Os repórteres do jornal The Herald tinham um prazo para fazer os gráficos e estavam tão concentrados na média de punições de cada juiz que não prestaram atenção nos poucos casos em que não havia punição. Eles deveriam ter se perguntado se fazia sentido todos os juízes violarem a lei estadual, ainda que em poucos casos.

— Steve Doig, Walter Cronkite School of Journalism, Arizona State University

Dados Misturados, Escondidos e Ausentes

Lembro de uma situação engraçada em que tentamos acesso os dados de subsídios agrícolas concedidos pela União Europeia à Hungria: estava tudo lá — mas em um PDF extremamente pesado e misturado com dados de subsídios nacionais. Nossos programadores precisaram trabalhar por *horas* até que os dados se tornassem úteis.

Também tivemos uma história interessante sobre os subsídios europeus à pesca, que todas as agências de pagamento dos 27 estados-membro são obrigadas a declarar. Aqui está um trecho retirado de uma reportagem que escrevemos sobre o assunto: "No Reino Unido, por exemplo, o formato dos dados varia da muito amigável busca em páginas HTML a resumos em PDF ou até listas de recebedores em vários formatos escondidas no fim de press releases. Tudo isso em apenas um estado-membro. Na Alemanha e na Bulgária, enquanto isso, se publicavam listas vazias. Os títulos apropriados estão lá mas sem nenhum dado."

— Brigitte Alfter, Journalismfund.eu

O pão de 32 libras

Uma matéria do jornal Wales sobre quanto o governo galês está gastando em produtos sem glúten trouxe na manchete uma informação de que estavam sendo pagos 32 libras (cerca de R$ 100) por um pão. No entanto, eram 11 pães que custaram 2,82 libras cada.

Os números, obtidos de uma resposta redigida pela Assembleia Galesa e num release de estatísticas do Serviço Nacional de Saúde (National Health Service, NHS) galês, foram listados como custo por item. Entretanto, não foi dada nenhuma definição adicional no dicionário de dados para explicar a que se refere um item ou qual seria sua unidade de medida.

O jornal assumiu que o dado tratava de uma unidade de pão, e não de um pacote com vários pães — o que era realmente. Ninguém, nem os parlamentares responsáveis pela resposta nem a assessoria de imprensa, levantou a questão sobre a quantidade até a segunda-feira depois que a história foi publicada.

Portanto, não assuma que as notas explicativas para os dados do governo vão ajudar a esclarecer as informações apresentadas ou que as pessoas responsáveis pelos dados vão perceber que eles não são claros, mesmo quando você lhes disser qual é sua suposição equivocada.

Geralmente os jornais querem boas manchetes, então, a menos que algo obviamente contradiga uma interpretação, é mais fácil ficar com o que traz um bom título em vez de verificar em detalhes e arriscar que a matéria caia, especialmente sob prazos apertados.

Imagem 3. Pão sem glúten custa aos contribuintes galeses £32 (WalesOnline).

No entanto, jornalistas têm a responsabilidade de checar afirmações ridículas mesmo que isso signifique que a matéria deixe de ser manchete.

— Claire Miller, WalesOnline

Comece com os dados e termine com uma reportagem

Para atrair leitores com jornalismo de dados, você tem de conseguir mostrar algum número na manchete que os faça sentar e prestar atenção. O texto deve ser fluido o suficiente para que não se note sua origem: um conjunto de dados. Pense sobre qual é o seu público enquanto o desenvolve e tente escrevê-lo de maneira empolgante.

Um exemplo disso pode ser encontrado em um projeto desenvolvido pelo Bureau of Investigative Journalism utilizando o Sistema de Transparência Financeira da Comissão Europeia. A reportagem foi construída aproximando-se o conjunto de dados de perguntas específicas.

Procuramos nos dados por palavras-chave como "coquetel", "golfe" e "dias ausentes". Isso nos permitiu determinar quanto a Comissão havia gasto nesses itens e nos trouxe diversas questões e caminhos a seguir.

No entanto, palavras-chave nem sempre trazem a resposta para o que você procura — às vezes, é preciso fazer uma pausa e pensar sobre o que realmente se está buscando. Durante esse projeto, também queríamos descobrir quanto os membros da comissão gastaram em viagens de jatinhos, mas como o conjunto de dados não continha os termos "jato particular", tivemos de conseguir o nome da prestadora de serviços de viagem. Assim que soubemos o nome da empresa, "Abelag", pudemos investigar os dados e descobrir quanto estava sendo gasto em serviços oferecidos pela companhia.

Com essa abordagem, tínhamos um objetivo claro na análise dos dados — encontrar um número para colocar na manchete. Os detalhes viriam em seguida.

Uma outra abordagem é procurar pelo que não deveria estar ali naqueles dados. Exemplo de como isso funciona é mostrado pelo projeto colaborativo EU Structural Funds (Fundos Estruturais da União Europeia), organizado pelo Financial Times em parceria com o Bureau of Investigative Journalism.

Analisamos os dados com base nas regras da própria Comissão em relação a empresas e associações que deveriam ser proibidas de receber fundos estruturais. Gastos com tabaco e produtores de tabaco estão proibidos, por exemplo.

Buscando por nomes de empresas, processadores e produtores de tabaco, encontramos informações que revelavam que a British American Tobacco havia recebido 1,5 milhão de euros para uma fábrica na Alemanha.

Como o financiamento não seguia as regras de despesas da Comissão, foi uma maneira rápida de encontrar uma boa história nos dados.

Você nunca sabe o que encontrará num conjunto de dados, então basta procurar. Você deve ser bastante ousado e essa abordagem geralmente funciona melhor quando tentamos identificar características óbvias por meio de filtragem (os maiores, os extremos, os mais comuns, etc.).

— Caelainn Barr, Citywire

Contando histórias com dados

Às vezes, o jornalismo de dados dá a impressão de que se limita à apresentação dos dados — como visualizações que transmitem de forma rápida as informações, ou bancos de dados interativos que permitem às pessoas pesquisar locais como ruas ou hospitais. Tudo isso é muito valioso, mas o jornalismo de dados também deve contar histórias. Mas quais são os tipos de histórias que você encontra em bases de dados? Com base em minha experiência na BBC, elaborei uma lista, ou "tipologia", de diferentes tipos de histórias vindas de dados.

Acho que ajuda ter esta lista sempre em mente. Não só quando você está analisando os dados, mas também no momento anterior, de coleta (seja pesquisando bancos de dados públicos ou fazendo pedido pela lei de acesso à informação).

Medição

A história mais simples é contar ou totalizar algo: "Prefeituras de todo o país gastaram tantos bilhões em clipes de papel no último ano." Frequentemente, é difícil saber se isso é muito ou pouco. Para isso, é necessário contexto, que pode ser obtido por meio de:

Proporção

"No último ano, as prefeituras gastaram dois terços de seu orçamento de papelaria em clipes de papel."

Comparação interna

"As prefeituras gastaram mais em clipes de papel do que enviando refeições a domicílio para idosos."

Comparação externa

"O gasto das prefeituras em clipes de papel ao longo do último ano foi o dobro do orçamento nacional de ajuda externa."

Também há outras formas de explorar os dados de maneira contextualizada ou comparativa:

Mudança ao longo do tempo

"Os gastos das prefeituras em clipes de papel triplicou ao longo dos últimos quatro anos."

"Tabelas de classificação"

São normalmente geográficas ou por instituição, e você precisa se certificar de que a base de comparação seja razoável (por exemplo, levando em consideração o tamanho da população local). "A prefeitura de Borsetshire gasta mais em clipes para cada funcionário do que qualquer outra autoridade local, cerca de quatro vezes a média nacional."

Ou você pode dividir os temas dos dados em grupos:
Análise por categorias

"Prefeituras administradas pelo Partido Roxo gastam 50% mais em clipes de papel do que as administradas pelo Partido Amarelo."

Ou você pode relacionar fatores numericamente:
Associação

"Prefeituras administradas por políticos que receberam doações de empresas de papelaria gastam mais em clipes de papel, com gastos médios de 100 libras para cada libra doada."

Mas, claro, sempre lembre que correlação e nexo de causalidade não são a mesma coisa.

Ao investigar gastos com clipes de papel, certifique-se de que se preocupou com a coleta das seguintes informações:

- Gastos totais para dar um contexto
- Relações geográficas/históricas/outras para fornecer dados comparativos
- Os dados adicionais que você precisa para dar credibilidade às comparações são justos, levando em conta o tamanho da população?
- Outros dados que proporcionem análises interessantes ao compará-los ou relacioná-los com os gastos

— Martin Rosenbaum, BBC

Jornalistas de dados comentam suas ferramentas preferidas

Psssss. Esse é o som da descompressão dos dados. E agora? O que você procura? E que ferramentas são necessárias para sair do lugar? Pedimos a alguns jornalistas de dados que comentassem como trabalham com as informações encontradas. Veja o que dizem:

No Guardian Datablog, fazemos questão de interagir com nossos leitores. Como eles têm a chance de ver e replicar nosso conteúdo com rapidez, acabam notando detalhes que nos escapam. Nesse contexto, quanto mais intuitivas forem as ferramentas de dados, melhor. Tentamos selecionar ferramentas acessíveis a qualquer um, sem que seja preciso pagar, aprender uma linguagem de programação ou fazer algum treinamento especial.

Por isso usamos bastante os produtos do Google. Todos os conjuntos de dados que organizamos e divulgamos são oferecidos ao público na forma de planilhas do Google. Qualquer pessoa que tenha uma conta no serviço pode baixar essas informações, elaborar seus próprios gráficos, organizá-las e criar tabelas dinâmicas.

Usamos o Google Fusion Tables para mapear os dados. Quando criamos mapas de calor, compartilhamos os arquivos em formato KML. Assim, os leitores podem baixá-los e construir seus próprios mapas de calor — talvez adicionando camadas extras ao original do Datablog. Outra característica interessante das ferramentas do Google é que funcionam bem nos diferentes dispositivos usados pelos leitores para acessar o blog, seja o computador, tablet ou smartphone.

Além do Google Spreadsheets e do Fusion, utilizamos outras duas ferramentas no nosso dia a dia. A primeira é o Tableau, para visualizar conjuntos de dados multidimensionais; e a segunda é o ManyEyes, para rápidas análises dos dados. Nenhuma delas é perfeita, por isso continuamos procurando outras ferramentas de visualização que agradem aos nossos leitores.

The Guardian— Lisa Evans

Será que algum dia vou ser programadora? Dificilmente! Certamente não acho que todos os repórteres precisam aprender a programar. Mas acredito que é muito útil ter uma noção geral do que é possível fazer, além, claro, de saber como conversar com programadores.

Se você está começando agora, vá com calma. Primeiro, é preciso convencer seus colegas e editores de que o jornalismo de dados vale a pena. Isso porque

pode gerar reportagens únicas e impossíveis de se obter de outras maneiras. Uma vez notado o valor dessa abordagem, você pode partir para matérias e projetos mais complexos.

Meu conselho é que, a princípio, você aprenda a usar o Excel e faça reportagens mais simples. Comece aos poucos e vá aprimorando a análise e o mapeamento de bases de dados. É possível fazer tanta coisa no Excel — uma ferramenta poderosa pouco explorada pela maioria das pessoas. Se puder, faça um curso de Excel para jornalistas, como o que é oferecido pelo Centre for Investigative Journalism.

Sobre a interpretação dos dados: leve com muita seriedade. Fique atento aos detalhes e questione seus resultados. Mantenha anotações sobre como você está processando os dados e faça uma cópia de documentos e arquivos originais. É fácil cometer erros. Sempre faço minhas análises duas ou três vezes praticamente do zero. E é ainda melhor se você pedir ao seu editor ou a alguma outra pessoa que analise os dados separadamente, de forma a comparar os resultados.

Financial Times— Cynthia O'Murchu

A habilidade de criar softwares complexos tão rapidamente quanto um repórter escreve uma reportagem é um fenômeno recente. Isso costumava levar muito mais tempo. As coisas mudaram graças ao surgimento de duas plataformas de desenvolvimento grátis e de código aberto: Django e Ruby on Rails, ambas lançadas em meados do ano 2000.

O Django, construído na linguagem de programação Python, foi desenvolvido por Adrian Holovaty ao lado de uma equipe que trabalhava na redação do Lawrence Journal-World em Lawrence, cidade do estado norte-americano do Kansas. O Ruby on Rails foi criado em Chicago por David Heinemeier Hansson e a 37Signals, empresa focada em aplicativos web.

As duas plataformas abordam o "padrão MVC" de arquitetura de software de formas diferentes, são excelentes e permitem construir rapidamente até mesmo aplicativos mais complexos. Elas eliminam parte do trabalho rudimentar ligado ao desenvolvimento de um programa. Coisas como criar e localizar itens da base de dados e relacionar URLs com códigos específicos do app já vêm incorporadas às plataformas, dispensando os programadores de escrever códigos para essas ações básicas.

Não há uma pesquisa sobre as equipes que desenvolvem aplicativos de notícias nos Estados Unidos, mas acredita-se que a maioria delas utiliza uma dessas plataformas para criar apps noticiosos que utilizam bases de dados. No ProPublica, usamos o Ruby on Rails.

O desenvolvimento de servidores web rápidos, como o Amazon Web Services, também derrubou parte das barreiras que tornavam lenta a criação de aplicativos.

Independentemente disso, usamos ferramentas bem comuns para trabalhar com dados: Google Refine e Excel para limpar os dados; SPSS e R para estatísticas; ArcGIS e QGIS para sistemas de informação geográfica (GIS); Git para gerenciar códigos-fonte; TextMate, Vim e Sublime Text para escrever códigos; e uma mistura de MySQL, PostgresSQL e SQL Server para bases de dados. Também criamos nossa própria plataforma JavaScript, chamada "Glass", que permite desenvolver rapidamente aplicativos interativos em Java.

ProPublica— Scott Klein

Às vezes, a melhor ferramenta é a mais simples — é fácil subestimar o poder de uma planilha. Usar um planilha quando tudo ainda era em DOS me permitiu entender uma complexa fórmula para o acordo de parceria entre os donos do clube de beisebol Texas Rangers — na época em que George W. Bush era um de seus principais sócios. Uma planilha me ajuda a perceber discrepâncias e erros nos cálculos. E ainda posso escrever scripts para limpeza dos dados e muito mais. É um item básico no pacote de ferramentas de um jornalista de dados.

Dito isso, minhas ferramentas favoritas são ainda mais poderosas — SPSS para análises estatísticas e programas de mapeamento para identificar padrões geograficamente.

The Seattle Times— Cheryl Phillips

Sou um grande fã do Python. Trata-se de uma maravilhosa linguagem de programação, de código aberto, fácil de ler e de escrever (você não precisa colocar um ponto e vírgula depois de cada linha). Mais importante, o Python tem uma grande base de usuários e conta com diversos plugins (chamados de pacotes) para qualquer coisa que você precise.

Considero o Django uma ferramenta pouco necessária para jornalistas de dados. Trata-se de uma plataforma web para a linguagem Python — uma ferramenta

para criar aplicativos grandes relacionados a bases de dados. O Django é definitivamente muito pesado para pequenos infográficos interativos.

Também uso o QGis, uma ferramenta de código aberto que oferece uma ampla gama de funções para jornalistas de dados que trabalham com informações geográficas. Se você precisa converter dados geoespaciais de um formato a outro, QGis é a ferramenta ideal por ser compatível com praticamente todos os formatos existentes (Shapefiles, KML, GeoJSON, etc.). Se for preciso cortar algumas regiões, o QGis também faz isso. E tem a vantagem de contar com uma grande comunidade de usuários que publicam tutoriais na internet.

O programa R foi criado essencialmente como uma ferramenta de visualização científica. É difícil encontrar algum método de visualização ou técnica de manipulação de dados que ele já não tenha. Trata-se de um universo próprio, a Meca da análise visual de dados. A desvantagem é que você precisa aprender (mais uma) linguagem de programação, pois o R tem uma linguagem própria. Mas depois de iniciada a escalada na curva de aprendizagem, não vai haver ferramenta mais poderosa do que esse software. Jornalistas de dados podem usar o R para analisar grandes conjuntos de dados que ultrapassem os limites do Excel (por exemplo, uma tabela com um milhão de linhas).

O que realmente é muito legal no R é a possibilidade de manter um "protocolo" exato do que você está fazendo com os dados — da leitura de um arquivo CSV à criação de gráficos. Se os dados mudarem, você recria o gráfico com um clique. Caso alguém fique curioso quanto à integridade do material, é só mostrar a fonte exata, que também permite que qualquer pessoa recrie o gráfico por conta própria (ou talvez identifique os erros que você cometeu).

A combinação NumPy + MatPlotLib é um modo de fazer a mesma coisa na linguagem Python. São dois exemplos de pacotes em Python usados para análises e visualização de dados, ambos são limitados a visualizações estáticas. Porém, não podem ser empregados na criação de gráficos interativos mais avançados.

Não uso o MapBox, mas ouvi falar que é uma boa ferramenta para criar mapas mais sofisticados baseados no OpenStreetMap. Ele permite, por exemplo, personalizar os estilos nos mapas (cores, textos, etc). E também tem um companheiro, chamado Leaflet, que é uma biblioteca JavaScript de alto nível para mapear que permite alternar rapidamente entre os fornecedores de mapas (OSM, MapBox, Google Maps, Bing, etc.).

RaphaelJS é uma biblioteca de visualização mais simples para trabalhar com formas básicas (círculos, linhas, texto). É possível fazer animações com os elementos, adicionar interações, entre outros recursos. Como não contém qualquer modelo para ser usado como base de gráficos, é preciso desenhar um conjunto de retângulos por conta própria.

Apesar disso, o lado bom do Raphael é que tudo o que você fizer nele funciona também no Internet Explorer. Esse não é o caso de muitas outras (maravilhosas) bibliotecas de visualização, como o d3. Infelizmente, muitos ainda usam o IE e nenhuma redação deveria ignorar esses 30% de seus usuários.

Além do RaphaelJS, há também a opção de criar alternativas em Flash para o IE. É basicamente o que o The New York Times está fazendo. Isso significa que você tem que desenvolver cada aplicativo duas vezes.

Ainda não estou convencido sobre o "melhor" processo para conciliar a visualização no Internet Explorer e em outros navegadores modernos. Muitas vezes, noto que os aplicativos em RaphaelJS são horrivelmente lentos no IE, mais ou menos dez vezes mais devagar do que rodam em Flash nos browsers modernos. Assim, alternativas em Flash podem ser a melhor opção se você quiser oferecer visualizações animadas e de alta qualidade para todos.

Open Knowledge Foundation— Gregor Aisch

Minha ferramenta mais útil é o Excel, que pode lidar com a maioria dos problemas de Reportagem com Auxílio de Computador (RAC) e tem a vantagem de ser fácil de aprender e estar disponível para a maioria dos repórteres. Quando preciso unir tabelas, costumo usar o Access, mas depois exporto o conteúdo de volta ao Excel para continuar o trabalho. Utilizo o ArcMap da ESRI para análises geográficas; uma ferramenta poderosa utilizada por agências que coletam dados geocodificados. O TextWrangler é bom para examinar dados textuais por meio de layouts peculiares e delimitados, e tem a opção de localizar e substituir expressões regulares. Quando técnicas estatísticas como a regressão linear são necessárias, utilizo o SPSS, que tem um menu intuitivo. Para trabalhos ainda mais pesados, por exemplo filtrar e programar conjuntos de dados com milhões de registros, recorro ao SAS.

Walter Cronkite School of Journalism— Steve Doig

Nossas ferramentas incluem Python e Django para hackear, capturar dados de páginas (scraping) e brincar com eles; e o PostGIS, QGIS e MapBox para criar mapas online mirabolantes. R e a dupla NumPy + MatPlotLib disputam nossa preferência para fazer a análise exploratória dos dados, embora nossa ferramenta predileta ainda está amadurecendo: CSVKit. De certa forma, tudo o que fazemos é desenvolvido na nuvem.

Chicago Tribune— Brian Boyer

No La Nación, costumamos utilizar:

- Excel para limpar, organizar e analisar os dados;
- Planilhas do Google para publicar os dados e conectá-los com serviços como o Google Fusion Tables e o Junar Open Data Platform;
- Junar para compartilhar dados ou incorporá-los a algum de nossos posts no blog;
- Tableau Public para visualizações interativas de dados;
- Qlikview, uma ferramenta de inteligência de negócios muito rápida para analisar e filtrar grandes conjuntos de dados;
- NitroPDF para converter PDFs para texto e arquivos de Excel; e
- Google Fusion Tables para visualizações de mapa.

La Nación (Argentina)— Angélica Peralta Ramos

Como o Transparência Hacker é uma comunidade sem preconceito técnico, usamos diversas ferramentas e linguagens de programação. Cada membro tem suas preferências, e essa grande variedade é ao mesmo tempo nossa força e nossa fraqueza. Alguns de nós estão produzindo uma distribuição de Linux da Transparência Hacker, que nos permitirá trabalhar com os dados em qualquer lugar. Esse kit tem algumas ferramentas e bibliotecas muito úteis, como: Refine, RStudio e o OpenOffice Calc (comumente subestimado pelos mais experientes, mas muito útil para trabalhos rápidos/pequenos). Também temos utilizado muito o Scraperwiki para fazer protótipos e salvar os resultados online.

> Para a visualização de dados e gráficos, gostamos de muitas ferramentas, como Python e NumPy, que são muito poderosas. Algumas pessoas na comunidade têm brincado com o R, mas ainda acho que bibliotecas JavaScript para plotagem de gráficos como d3, Flot e RaphaelJS acabam sendo empregadas na maioria dos nossos projetos. Por fim, temos feito muitas experiências com mapas, e o Tilemill tem se mostrado uma ferramenta de trabalho muito interessante.
>
> *Transparência Hacker*— Pedro Markun

Como mostramos os dados no Verdens Gang

Jornalismo é levar novas informações ao leitor o mais rápido possível. A forma mais rápida pode ser um vídeo, uma fotografia, um texto, um gráfico, uma tabela ou uma combinação de tudo isso. A respeito de visualizações, o objetivo deve ser o mesmo: informação rápida. Novas ferramentas de dados permitem aos jornalistas encontrar histórias com as quais eles não teriam contato de outra forma, assim como apresentá-las de novas maneiras. Aqui estão alguns exemplos de como nós apresentamos dados no jornal mais lido na Noruega, o Verdens Gang (VG).

Números

Esta história é baseada em dados do Instituto de Estatísticas Norueguês, dados de contribuintes e dados do monopólio nacional de loterias. No gráfico interativo abaixo, o leitor podia encontrar diferentes tipos de informação de cada municipalidade ou condado norueguês. A tabela mostra a porcentagem da renda gasta em jogos e foi construída usando-se o Access, Excel, MySql e Flash.

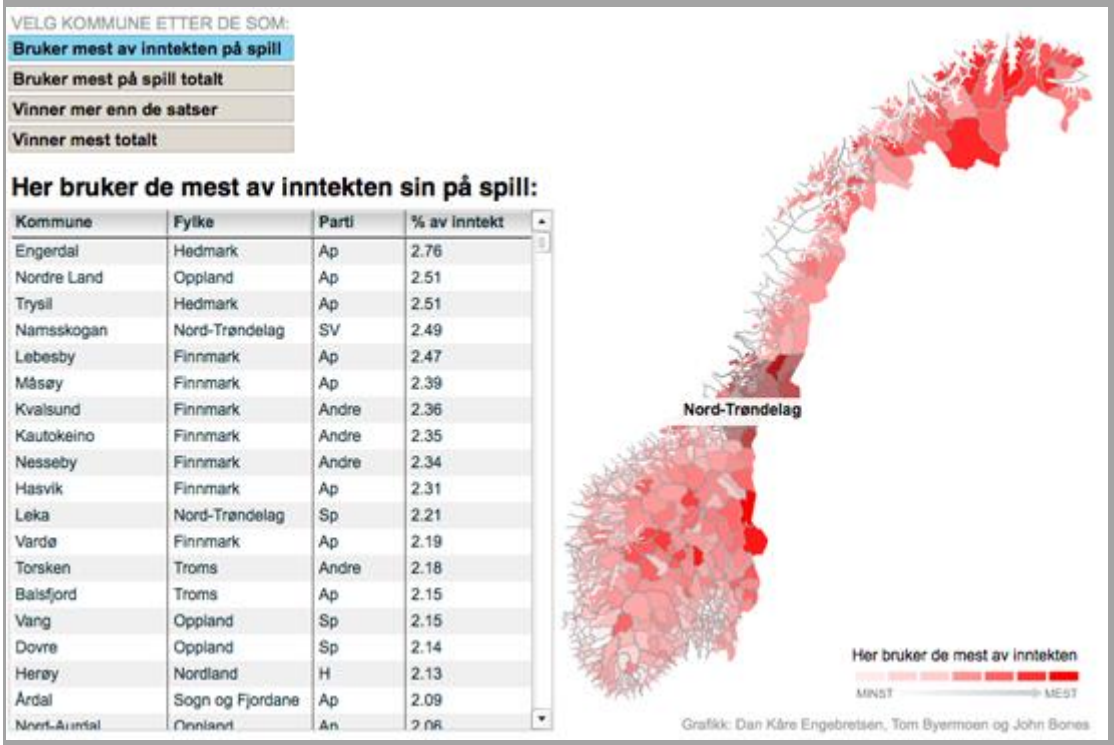


Imagem 23. Mapeando dados dos contribuintes e da Lotto (Verdens Gang)

Redes

Nós utilizamos análises de redes sociais para estudar as relações entre os 157 filhos e filhas das pessoas mais ricas da Noruega. Nossa investigação mostrou

que os herdeiros dos mais ricos da Noruega também herdaram as redes sociais dos seus pais. Ao todo, foram mais de 26.000 conexões, e os gráficos foram todos finalizados manualmente com o Photoshop. Usamos Access, Excel, Bloco de Notas e a ferramenta de análise de redes sociais Ucinet.

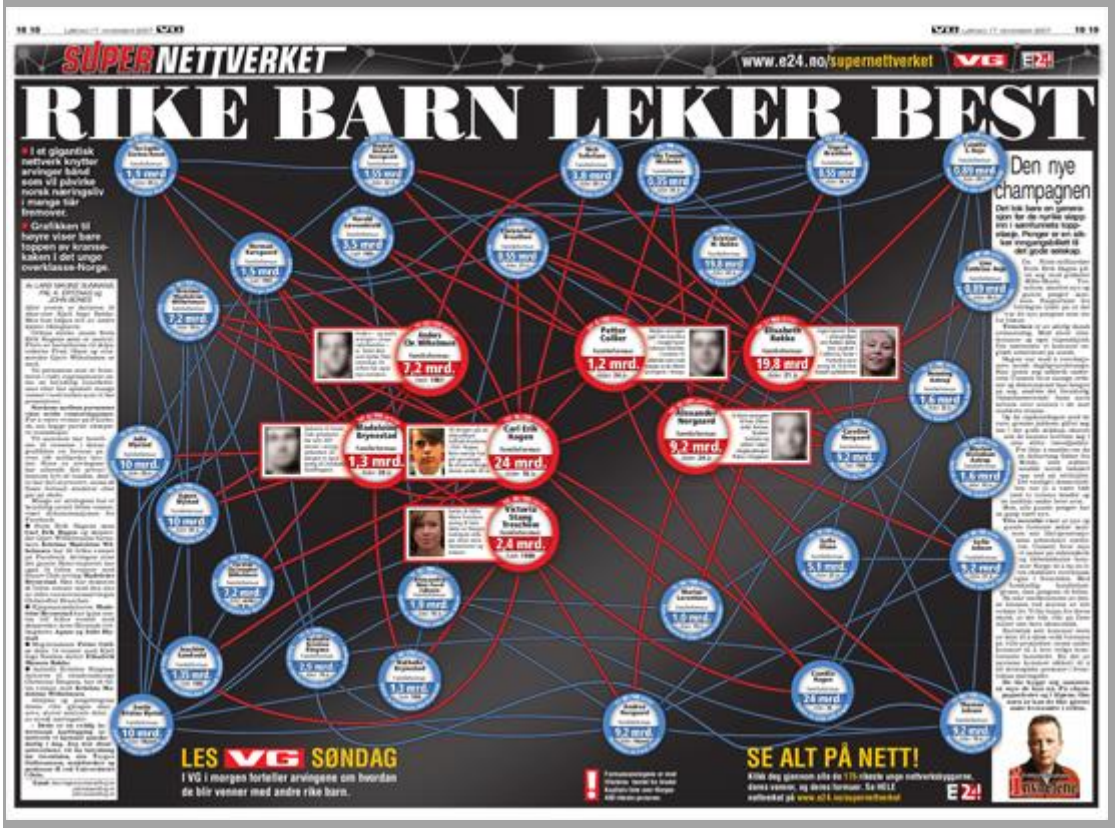


Imagem 24. Aves ricas de mesma plumagem voam juntas (Verdens Gang)

Mapas

Neste mapa de calor animado combinado com um gráfico simples de barras, você pode ver a incidência de crimes no centro de Oslo, hora a hora, no fim de semana, por vários meses. No mesmo mapa, é possível conferir o número de oficiais da polícia trabalhando ao mesmo tempo. Quando o crime está realmente acontecendo, a quantidade de policiais está no nível mais baixo. O mapa foi feito usando ArcView com Spatial Analyst.

Imagem 25. Mapa de calor animado (Verdens Gang)

Mineração de texto

Para esta visualização, fizemos mineração de dados (extração de padrões ocultos em bases de dados) nos discursos feitos por sete líderes de partidos noruegueses durante suas convenções partidárias. Todos os discursos foram analisados, e esses estudos forneceram ângulos para algumas reportagens. Cada reportagem foi relacionada a um gráfico e os leitores puderam explorar e conhecer melhor a linguagem dos políticos. Essa visualização foi feita usando Excel, Access, Flash e Illustrator. Se tivesse sido feito em 2012, teríamos feito o gráfico interativo em JavaScript.

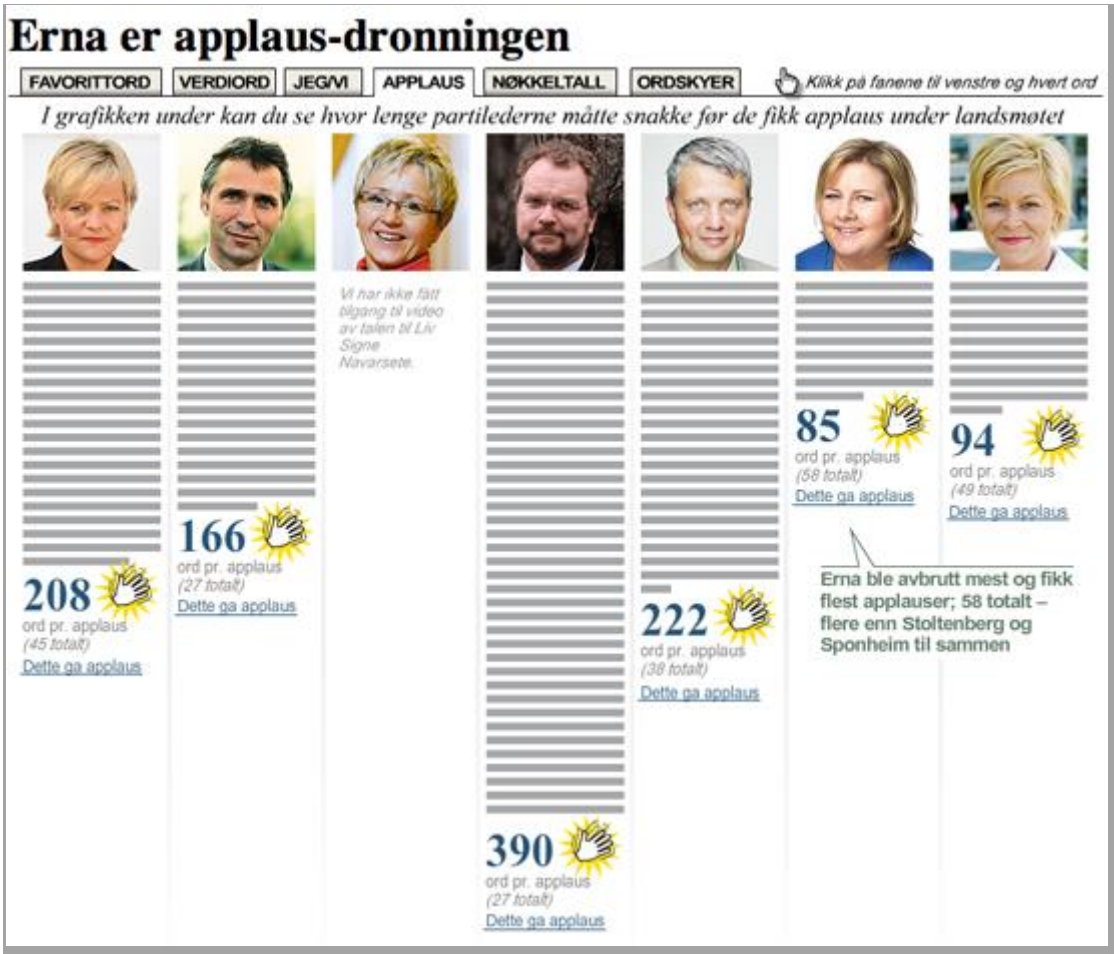


Imagem 26. Mineração de texto dos discursos de líderes partidários (Verdens Gang)

Conclusão

Quando precisamos visualizar uma reportagem? Na maioria das vezes não precisamos, mas há momentos em que queremos fazer isso para ajudar nossos leitores. Reportagens que contêm uma grande quantidade de dados geralmente precisam de visualização. No entanto, temos de ser críticos ao escolher que tipo de dados vamos apresentar. Conhecemos todos os detalhes quando informamos sobre algo, mas o que o leitor realmente precisa saber na reportagem? Talvez uma tabela seja suficiente, ou um gráfico simples mostrando uma evolução do ano A para o ano C. Ao trabalhar com jornalismo de dados, a questão não é necessariamente apresentar grandes quantidades de dados. É sobre jornalismo!

> Tem havido uma tendência clara nos últimos três anos para criar gráficos interativos e tabelas que permitem ao leitor se aprofundar em temas diferentes. Uma boa visualização é como uma boa fotografia. Você entende do que se trata só de olhar para ela por um momento ou dois. Quanto mais você olhar para a visualização, mais você a vê. A visualização é ruim quando o leitor não sabe por onde começar ou terminar, e quando a visualização está sobrecarregada de detalhes. Neste cenário, talvez um texto seja melhor, não?
>
> *— John Bones, Verdens Gang*

Dados públicos viram sociais

Os dados têm valor inestimável. O acesso a eles tem o potencial de jogar luz sobre diversos assuntos de uma forma que impulsiona resultados. No entanto, um mau tratamento dos dados pode colocar os fatos em uma estrutura que não comunica nada. Se não promover discussão ou proporcionar um entendimento contextualizado, os dados podem ter um valor limitado para o público.

A Nigéria voltou para a democracia em 1999, depois de longos anos de ditadura militar. Sondar os fatos por trás dos dados era uma afronta à autoridade e visto como uma tentativa de questionar a reputação da junta. A Lei de Segredos Oficiais levou os funcionários públicos a não compartilhar informações do governo. Mesmo 13 anos depois da volta da democracia, acessar dados públicos pode ser uma tarefa difícil. Quando se trata de informações sobre gastos públicos, por exemplo, é difícil passá-las de uma maneira clara para a maioria da audiência, que não conhece bem contabilidade financeira.

Com o aumento do número de celulares e de nigerianos online, vimos uma imensa oportunidade de usar tecnologias de visualização de dados para explicar e engajar as pessoas em torno às despesas públicas. Para isso, tínhamos que envolver os usuários em todas as plataformas, assim como chegar aos cidadãos por meio de ONGs. Lançamos o projeto BudgIT, que visa fazer dos dados públicos um objeto social, e construir um extensa rede que demande mudanças.

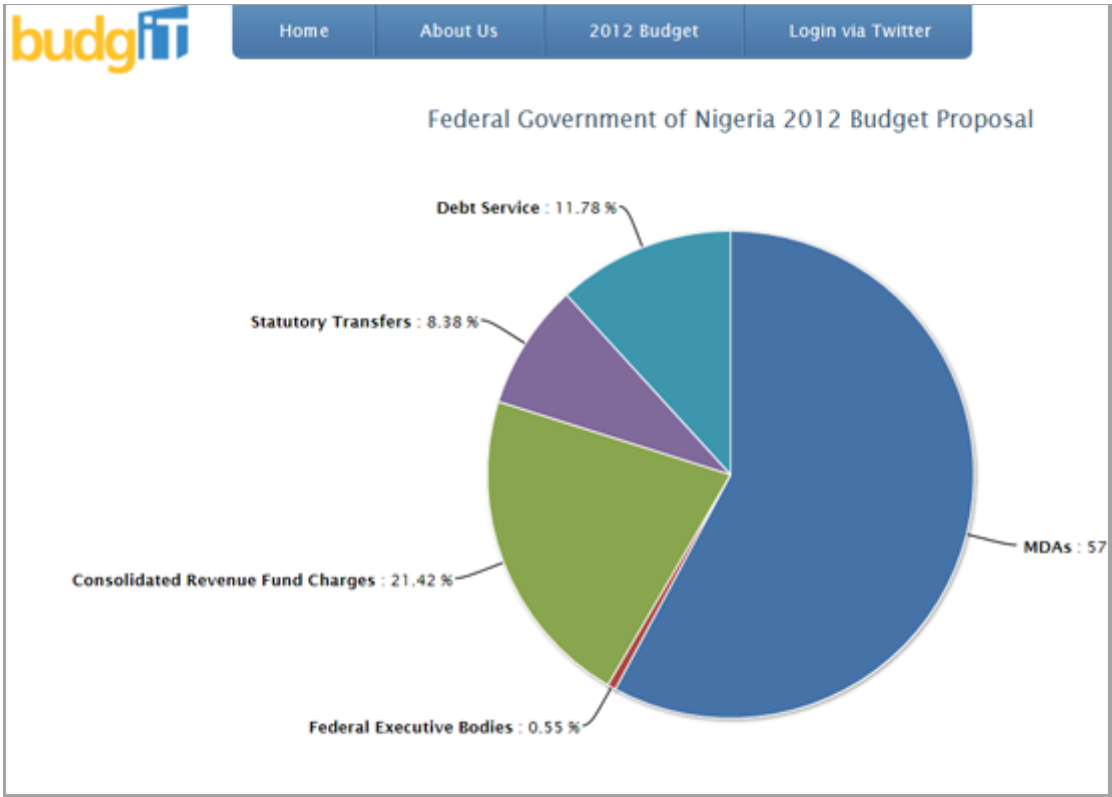


Imagem 27. Aplicativo The BudgIT (BudgIT Nigeria)

Para conseguir engajar os usuários, temos que entender o que eles querem. Com o que o cidadão nigeriano se preocupa? Onde eles veem uma lacuna de informação? Como podemos tornar os dados relevantes para suas vidas? O alvo imediato do BudgIT é o nigeriano de educação média, conectado a fóruns online e mídias sociais. Para competir pela atenção de usuários, temos que apresentar os dados de maneira breve e concisa. Após transmitir uma prévia dos dados na forma de um tweet ou infográfico, há oportunidade para um envolvimento sustentado. Isso pode ser feito por meio de uma experiência mais interativa, a fim de entregar aos usuários um contexto mais amplo.

Na visualização de dados, é importante entender o nível de compreensão que os usuários têm desse tipo de informação. Por mais bonitos e sofisticados que sejam, vimos que diagramas complexos e aplicativos interativos podem não ser ideais para uma comunicação efetiva com os nossos leitores. Uma boa visualização vai falar com o usuário por meio de um uma linguagem que ele entenda, assim como contar uma história com a qual ele sinta uma conexão imediata.

Conseguimos engajar mais de 10 mil nigerianos na questão do orçamento, e os dividimos em três categorias para dar a eles informações de maior valor. As categorias estão explicadas resumidamente abaixo:

Usuários ocasionais

Querem informação de forma simples e rápida. Estão interessados em ter uma ideia geral dos dados, não em análises detalhadas. Podemos atrai-los com tweets ou gráficos interativos.

Usuários ativos

Estimulam a discussão e usam os dados para melhorar seus conhecimentos sobre determinada área ou desafiar as suposições ligadas a eles. Para esses usuários, queremos oferecer mecanismos de feedback e a possibilidade de que compartilhem ideias com seus pares pelas redes sociais.

Consumidores massivos de dados

Querem dados brutos para visualização ou análise. Nós simplesmente damos a eles as informações que desejam.

Com o BudgIT, o engajamento do nosso usuário é baseado em:

Estimular discussões sobre tendências atuais

O BudgIT acompanha discussões online e offline e procura fornecer dados sobre os assuntos atuais. Por exemplo, durante as greves do setor de combustíveis de janeiro de 2012, houve agitação constante entre os manifestantes com relação à necessidade de reinstituir os subsídios ao combustível e reduzir gastos públicos exagerados e desnecessários. O BudgIT acompanhou o debate pelas mídias sociais e, em 36 trabalhosas horas, construiu um aplicativo que permite aos cidadãos reorganizar o orçamento nigeriano.

Bons mecanismos de feedback

Tentamos engajar os usuários por meio de canais de discussão e das redes sociais. Muitos querem conhecer as histórias ligadas aos dados, enquanto outros perguntam nossa opinião. Garantimos que nossas respostas expliquem apenas os fatos por trás dos dados, sem vínculos com visões pessoais ou políticas. Precisamos manter abertos os canais de feedback, responder ativamente a comentários e envolver a audiência criativamente para garantir que a comunidade construída ao redor dos dados se mantenha.

Tornar local

Para uma base de dados voltada a um grupo específico de usuários, o BudgIT tenta localizar ou adaptar seu conteúdo e promover um canal de discussão que se conecte às suas necessidades. Em particular, estamos interessados em engajar o público por meio de mensagens SMS.

Depois de publicar dados de gastos no site yourbudgit.com, chegamos aos cidadãos com a ajuda de várias ONGs. Também planejamos desenvolver uma rede participativa em que os cidadãos e instituições governamentais se encontrem em prefeituras para definir itens fundamentais do orçamento a serem priorizados.

O projeto teve cobertura de mídia local e estrangeira, da CP-Africa à BBC. Fizemos uma análise dos orçamentos de 2002-2011 para o setor de segurança para uma jornalista da AP, Yinka Ibukun. A maioria das organizações de mídia é composta por "usuários pesados de dados" e nos pede informações para usar em reportagens. Estamos planejando mais colaborações com jornalistas e organizações de notícias ao longo dos próximos meses.

— Oluseun Onigbinde, BudgIT Nigeria

Engajando pessoas nos seus dados

Tão importante quanto publicar dados é obter uma reação da audiência. Você é humano; vai cometer erros, perder coisas e ter ideias erradas de tempos em tempos. A sua audiência é um dos bens mais úteis que você tem. Ela pode verificar fatos e apontar outras coisas que não foram consideradas.

Engajar o público, no entanto, é complicado. Você está lidando com um grupo de pessoas condicionadas por anos de uso da internet, de navegação de site em site, e que deixam apenas um comentário sarcástico ao longo de suas caminhadas. Construir uma relação de confiança com seus usuários é crucial; eles precisam saber o que vão obter, como reagir e dar feedback ao que será ouvido.

Mas primeiro é preciso pensar no público que você tem, ou que deseja ter. O público que vai ser informado e informar por meio do tipo de dados com os quais você trabalha. Se a audiência está ligada a um setor particular, será necessário explorar formas de comunicação personalizadas. Existem organizações que você pode contatar para que ajudem na divulgação do material a um público mais amplo? Existem sites comunitários ou fóruns com os quais conversar? Há publicações comerciais especializadas que gostariam de ajudar na confecção de sua reportagem?

As redes sociais também são uma ferramenta importante. No entanto, mais uma vez, dependem do tipo de dados sobre a mesa. Se estiver trabalhando com estatísticas globais de transportes, por exemplo, vai ser complicado encontrar um grupo no Facebook ou no Twitter especialmente interessado nas suas atividades. Por outro lado, se estiver peneirando índices mundiais de corrupção ou de crimes locais, será mais fácil achar pessoas preocupadas com esses assuntos.

Quando se trata do Twitter, a melhor abordagem é entrar em contato com perfis de personalidades públicas, explicando brevemente a importância de seu trabalho e incluindo um link. Com sorte, eles retuitarão a mensagem aos seus leitores. Esta é uma ótima forma de aumentar a exposição do seu trabalho com um esforço mínimo — e sem atormentar as pessoas!

Depois de obter leitores para a sua página, pense em como eles vão interagir com seu trabalho. Claro, podem ler a história que você escreveu e ver mapas e infográficos. Mas é imensamente valioso oferecer também canais de resposta.

Mais que tudo, eles podem contribuir com ideias sobre o tema tratado, ajudando a definir as próximas tarefas do projeto de cobertura.

Primeiro, não precisa nem dizer que o ideal é publicar os dados brutos em suas reportagens. Você pode apresentar os dados em uma planilha CSV ou hospedá-los em outros serviços, como o Google Docs. Assim, você terá apenas uma versão dos dados e poderá atualizá-la a qualquer momento, por exemplo para corrigir possíveis erros. Se puder, a melhor alternativa é fazer as duas coisas. Permita que as pessoas acessem as informações brutas da sua reportagem da forma mais fácil possível.

Então, pense em outras formas de interagir com o público. Acompanhe as métricas que revelam quais partes de suas bases de dados estão conseguindo mais atenção — é provável que as áreas de maior tráfego digam algo sobre detalhes que você tenha perdido. Por exemplo, você pode não ter dado destaque para as estatísticas de pobreza da Islândia, mas se esses blocos recebem muitas visitas, é porque pode valer a pena estudá-los melhor.

Pense além da caixa de comentários. Você pode anexar comentários a células particulares de uma planilha? Ou a uma região específica de um infográfico? Enquanto a maioria dos sistemas de edição não permitem esse tipo de incorporação de informações, vale a pena avaliar essa possibilidade se estiver criando um material mais elaborado. Os benefícios que esse recurso pode trazer aos seus dados não podem ser subestimados.

Certifique-se de que os demais usuários também vejam esses comentários — em muitos casos, eles têm quase tanta importância quanto os dados originais, e se você mantiver essa informação somente para si, vai privar o público desse valor.

Finalmente, outras pessoas podem querer publicar seus próprios infográficos e histórias baseados nas mesmas fontes de dados. Por isso, pense em qual é a melhor forma de vinculá-los e alinhar o trabalho deles. Você também pode usar uma hashtag específica para o conjunto de dados. Ou, se ele for muito pictórico, compartilhe em um grupo do Flickr.

Também pode ser útil contar com uma via confidencial de compartilhamento de informações. Em alguns casos, algumas pessoas podem não se sentir seguras de fazer suas contribuições publicamente, ou mesmo não se sentir confortáveis nesse contexto. Elas podem preferir submeter informações por meio de um endereço de e-mail, ou até mesmo usar uma caixa de comentários anônimos.

> A coisa mais importante que você pode fazer com seus dados é divulgá-los da forma mais ampla e aberta possível. Permitir que os leitores verifiquem seu trabalho, encontrem erros e apontem detalhes perdidos que tornarão melhores tanto o seu jornalismo como a experiência do público.
>
> *— Duncan Geere, Wired.co.uk*

Comunicando os dados

Depois de observar bem os dados e decidir que eles rendem uma boa matéria, como você transmite tudo isso ao público? Esta seção começa com histórias curtas sobre como os jornalistas têm mostrado dados aos leitores — indo de infográficos e plataformas de dados abertos a links de download. Vamos examinar com mais detalhes como construir aplicativos de notícias e os prós e contras da visualização de dados. Finalmente, daremos uma olhada no que se pode fazer para engajar o público no seu projeto.

O que há neste capítulo?

Apresentando os dados ao público

Há muitas maneiras de apresentar dados ao público — da publicação de bancos de dados brutos em reportagens até a criação de belas visualizações e aplicativos interativos. Pedimos a alguns dos principais jornalistas de dados que dessem dicas sobre como fazer essa apresentação das informações.

Fazer ou não visualizações?

Há momentos em que os dados são a melhor opção para contar uma história, e não texto ou fotos. É por isso que expressões como "aplicativo de notícias" e "visualização de dados" viraram chavões em muitas redações. A atual safra de novas ferramentas e tecnologias (muitas vezes gratuitas) também desperta interesse na área. Elas ajudam até o jornalista com menos conhecimentos técnicos a transformar dados em narrativa visual.

Ferramentas como o Google Fusion Tables, Many Eyes, Tableau e Dipity facilitam muito a criação de mapas, tabelas, gráficos e até mesmo aplicativos de dados — algo até então limitado a especialistas. Com a facilidade de acesso às ferramentas, a questão não é mais se é possível transformar os dados numa visualização, mas quando se deve ou não fazê-lo.Uma visualização ruim de dados é pior do que nenhuma em muitos aspectos.

— Aron Pilhofer, New York Times

Usando animações gráficas

Desde que tenham um roteiro preciso, animações bem-cronometradas e explicadas claramente podem dar vida a números ou ideias complexas orientando o público pela reportagem. As palestras em vídeo de Hans Rosling são um bom exemplo de como os dados ganham vida na hora de contar uma história na tela. Concorde ou não com sua metodologia, também acho o trabalho da Economist, Índice de vulnerabilidade à revolução dos países árabes, um bom exemplo do uso do vídeo para ilustrar uma reportagem baseada em números. Você não iria, ou não deveria, apresentar o gráfico da Economist como uma imagem estática. Há muita coisa acontecendo. Mas vendo o passo a passo de como foi desenvolvido, você entende como e por que se chegou ao índice. Gráficos em movimento reforçam o que o público está ouvindo. Uma voz em off, com efeitos visuais explicativos, é um recurso muito poderoso e memorável ao contar uma história.

— Lulu Pinney, designer freelance de infográficos

Mostrando ao mundo

Nosso fluxo de trabalho geralmente começa com o Excel. É uma maneira fácil e rápida de se trabalhar os dados. Se identificamos informações valiosas, vamos à redação — temos a sorte de estar ao lado da redação principal do Guardian. Então, observamos como devemos visualizar ou exibir os dados na página, e escrevemos o post que os acompanhará. Quando estou escrevendo, geralmente tenho ao lado do editor de texto uma versão reduzida da planilha em questão. Muitas vezes, também faço partes da análise enquanto escrevo, com o fim de destacar coisas interessantes. Por fim, publico o post e gasto um pouco de tempo tuitando sobre o tema, garantindo que a história esteja presente em todos os canais necessários e seja enviada aos lugares certos

Metade do tráfego de alguns dos nossos posts vem do Twitter e do Facebook. Estamos muito orgulhosos com o fato de que a média de tempo gasto pelo usuário lendo um post do Datablog seja de 6 minutos. Em comparação com a média de 1 minuto do resto do site do The Guardian, é um tempo muito bom. É importante lembrar que o tempo gasto numa página é uma das principais métricas para analisar audiência.

Esses números ajudam a convencer nossos colegas sobre o valor do que estamos fazendo. Isso e as grandes reportagens de dados com as quais trabalhamos: COINS (banco de dados do Tesouro do Reino Unido), WikiLeaks e os protestos violentos que atingiram o país. Para os dados sobre gastos do sistema COINS, tivemos 5 a 6 repórteres especializados ajudando quando as informações foram liberadas pelo governo do Reino Unido. Também tivemos outra equipe de 5 a 6 profissionais quando a administração britânica liberou informações de gastos acima de 25 mil libras — incluindo repórteres renomados, como Polly Curtis. O projeto WikiLeaks também foi, obviamente, muito importante, cheio de histórias sobre o Iraque e o Afeganistão. Os protestos violentos e saques no país também merecem destaque, com mais de 550 mil acessos em dois dias.

Mas não se trata apenas de sucessos no curto prazo: ser uma fonte confiável de informações úteis também é importante. Tentamos ser o lugar onde você pode obter informações significativas sobre os temas que cobrimos.

— Simon Rogers, The Guardian

Publicando os dados

Muitas vezes, inserimos os dados no site com uma visualização e um formulário que permite o download das informações. Nossos leitores podem explorar os dados por trás das histórias, interagindo com a visualização ou utilizando os dados de outras maneiras. Por que isso é importante? Porque aumenta a transparência do Seattle Times. Dessa forma, mostramos aos leitores os mesmos dados usados para tirar conclusões importantes. E quem os utiliza? Nossos críticos, com certeza, bem como aqueles interessados na reportagem e em todas as suas ramificações. Ao tornar os dados disponíveis, incentivamos a colaboração destes mesmos críticos e dos leitores em geral para descobrir fatos que possivelmente deixamos passar e que outros fatores poderíamos ter explorado — participação valiosa na busca do jornalismo relevante.

— Cheryl Phillips, The Seattle Times

Tornando os dados acessíveis

Facilitar o acesso do público aos dados que usamos em nosso trabalho é a coisa certa a fazer por várias razões. Os leitores podem se certificar de que não torturamos os dados para chegar a conclusões injustas. Abrir nossos dados é, na tradição da ciência social, permitir que pesquisadores repliquem o nosso trabalho. Incentivar os leitores a estudarem os dados pode gerar dicas que viram outras reportagens com aqueles dados. Finalmente, os leitores interessados em seus dados são mais suscetíveis a sempre voltar ao site.

— Steve Doig, Faculdade de Jornalismo Walter Cronkite, Universidade do Estado do Arizona

Iniciando uma plataforma de dados abertos

No La Nación, a publicação de dados abertos é parte fundamental de nossas atividades na área de jornalismo de dados. Na Argentina, não há leis de liberdade de informação nem portais nacionais de dados, de modo que nos sentimos fortemente compelidos a oferecer aos leitores o acesso aos dados usados em nossas reportagens.

Por isso, publicamos dados brutos estruturados em nossa plataforma integrada Junar, bem como no Google Spreadsheets. Incentivamos explicitamente que outras pessoas reutilizem nossos dados, e explicamos um pouco sobre como fazer isso com documentos e tutoriais em vídeo.

Além disso, apresentamos alguns desses conjuntos de dados e visualizações em nosso blogNación Data. Fazemos isso para disseminar uma cultura de dados e as ferramentas de publicação de dados na Argentina, bem como para mostrar aos outros como os utilizamos e como eles podem reutilizá-los.

Desde que inauguramos a plataforma, em fevereiro de 2012, recebemos sugestões e ideias ligadas a bancos de dados, a maior parte vinda de pesquisadores acadêmicos e estudantes universitários que se sentem gratos cada vez que respondemos com uma solução ou uma base de dados específica. As pessoas também estão bastante engajadas em nossos dados no Tableau, e várias vezes conseguimos ser o item mais comentado e visto no serviço. Em 2011, tivemos 7 das 100 visualizações mais visitadas.

— Angélica Peralta Ramos, La Nación (Argentina)

Tornando os dados humanos

À medida que a discussão sobre os limites do big data ganha maiores proporções, algo importante tem sido esquecido — o elemento humano. Muitos de nós pensamos sobre dados como números que flutuam livremente, mas eles são medições de coisas tangíveis (e muitas vezes relacionadas aos seres humanos). Os dados estão amarrados às vidas de pessoas reais, e quando nos envolvemos com números, precisamos considerar os sistemas do mundo real de onde vieram.

Tome-se, por exemplo, dados de localização, que estão sendo coletados agora por milhões de dispositivos móveis. É fácil pensar sobre estes dados (números que representam latitude, longitude e tempo) como um "escape digital", mas eles são retirados de momentos de nossas narrativas pessoais. Mesmo que pareçam secos e clínicos quando vistos em uma planilha, ao permitirmos que as pessoas coloquem seus próprios dados em um mapa para reproduzi-los, elas experimentam uma espécie de lembrança, poderosa e humana.

No momento, os dados de localização são usados por terceiros — desenvolvedores de aplicativos, grandes marcas e anunciantes. Embora as empresas de telecomunicações armazenem os dados, a parte principal nesta equação — você — não tem nem acesso nem controle sobre essa informação. No grupo de Pesquisa e Desenvolvimento do New York Times, lançamos um projeto piloto chamado OpenPaths para permitir que o público explorasse seus próprios dados de localização e experimentasse o conceito de propriedade dos dados.

Afinal, as pessoas devem ter o controle destes números tão estreitamente ligados às suas próprias vidas e experiências.

Jornalistas prestam um serviço muito importante ao trazer à tona esta humanidade inerente aos dados. Ao fazer isso, eles têm o poder de mudar a compreensão do público — tanto em relação aos dados quanto em relação aos sistemas que geram essas informações.

— Jer Thorp, Artista residente de dados: The New York Times R&D Group

Dados abertos, código aberto, notícias abertas

O ano de 2012 pode ser considerado o ano das notícias abertas para o Guardian. Elas estão no coração de nossa ideologia editorial e emitem uma mensagem-chave na nossa marca atual. No meio de tudo isso, é claro que precisamos de um processo aberto para o jornalismo de dados. Este processo não só deve ser alimentado por dados abertos, mas também ser ativado por ferramentas abertas. Esperamos ser capazes de oferecer, para cada visualização que publicamos, o acesso tanto aos dados por trás dela como ao código que a alimenta.

Muitas das ferramentas usadas hoje na visualização são de código fechado. Outras vêm com licenças que proíbem o uso de dados derivados. Muitas vezes, as bibliotecas existentes de código aberto resolvem bem um único problema, mas não oferecem uma metodologia mais abrangente. De modo geral, esses fatores dificultam que certos trabalhos sejam usados como base para outros. Esse cenário fecha diálogos, em vez de iniciar. Para que isso aconteça, estamos desenvolvendo um grupo de ferramentas abertas para a produção de narrativas interativas — The Miso Project (@themisoproject).

Estamos discutindo este trabalho com uma série de outras organizações de mídia. O envolvimento da comunidade é importante para que se possa atingir o pleno potencial do software de código aberto. Se formos bem-sucedidos, vai prevalecer uma dinâmica diferente entre nossos leitores. As contribuições podem ir além de comentários — por exemplo, promovendo a correção de bugs ou a reutilização de dados de formas inesperadas.

— Alastair Dant, The Guardian

Adicione um link para download

Nos últimos anos, tenho trabalhado com alguns gigabytes de dados para projetos ou reportagens, de varreduras de tabelas datilografadas dos anos 1960

a pesquisas de informações nos 1,5 gigabytes de telegramas diplomáticos divulgados pelo WikiLeaks. Sempre foi difícil convencer os editores a publicarem sistematicamente os dados brutos em um formato aberto e acessível. Ignorando o problema, adicionei links do tipo "Faça o Download dos Dados" dentro das reportagens, direcionando os leitores diretamente para o arquivo de referência. O interesse de potenciais reutilizadores era muito, muito baixo. No entanto, os poucos casos de reutilização nos deram novas ideias ou estimularam conversas que fazem valer cada minuto extra dedicado a cada projeto!

— Nicolas Kayser-Bril, Journalism++

Conheça seu campo de atuação

Há uma grande diferença entre hackear por diversão e construir uma estrutura para obter escala e desempenho. Certifique-se de que você estabeleceu uma parceria com pessoas que têm o conjunto de habilidades adequadas para o seu projeto. Não se esqueça do design. Design de usabilidade, experiência de usuário e apresentação podem influenciar muito o sucesso de seu projeto.

— Chrys Wu, Hacks/Hackers

Como construir um aplicativo jornalístico

Aplicativos jornalísticos são janelas para os dados por trás das reportagens. Devem permitir pesquisas em bancos de dados, visualizações intuitivas ou ir além disso. Independentemente do formato, aplicativos encorajam os leitores a interagir com os dados em um contexto que tem significado para eles: olhar tendências de crime na região onde vivem, checar registros de recomendações do médico local ou procurar contribuições a um candidato político.

Mais que infográficos de alta tecnologia, os melhores aplicativos jornalísticos são produtos duráveis. Eles fogem do ciclo da notícia, frequentemente ajudando leitores a resolver problemas do mundo real ou respondendo a perguntas de uma maneira nova e original. Quando jornalistas do ProPublica queriam explorar a segurança das clínicas norte-americanas de diálise renal, eles desenvolveram um aplicativo que ajudou os usuários a checar essas informações em suas cidades. Oferecer um serviço tão importante e relevante cria um relacionamento com os usuários que vai muito além do que uma reportagem pode fazer sozinha.

Está aí o desafio e a promessa de desenvolver aplicativos jornalísticos de última geração: criar algo de valor durável. Se você é um desenvolvedor, qualquer discussão sobre como construir um bom aplicativo jornalístico deve começar com uma mentalidade de criação de um produto: manter o foco no usuário e trabalhar para potencializar seu investimento. Então, antes de começar a construir, é interessante perguntar a si mesmo três questões, discutidas nas próximas páginas.

Imagem 1. Rastreador de instalações de diálise (ProPublica)

Quem é meu público e do que ele precisa?

Aplicativos jornalísticos não servem à reportagem, mas, sim, ao usuário. Dependendo do projeto, esse usuário deve ser um paciente de diálise que quer saber sobre a segurança de sua clínica, ou até mesmo o dono de uma casa desavisado sobre o risco de terremoto próximo à sua residência. Não importa quem seja, qualquer discussão sobre elaborar um aplicativo deve começar com as pessoas que farão uso da ferramenta.

Um único aplicativo deve servir a diversos usuários. Por exemplo, um projeto chamadoCurbwise, do Omaha World-Herald, no Nebrasca, é direcionado a proprietários de casas que desconfiam ser sobretaxados, a moradores curiosos que têm interesse nos valores das propriedades vizinhas, e a corretores de imóveis que querem se informar sobre promoções recentes. Em cada um desses casos, o aplicativo encontra uma necessidade específica que faz com que os usuários continuem trabalhando com a ferramenta.

Proprietários, por exemplo, podem precisar de ajuda para juntar informações sobre os imóveis próximos para argumentar que suas taxas são injustamente altas perante as outras. Unir esses dados é algo complicado e consome tempo, problema que o Curbwise resolve ao compilar em um relatório amigável todas as informações necessárias para reclamar das taxas às autoridades locais. Curbwise vende esse relatório por US$ 20 e as pessoas pagam porque ele resolve um problema real em suas vidas.

Se seu aplicativo resolve um problema real, como faz o Curbwise, ou completa uma reportagem com uma visualização interessante, tenha sempre em mente as pessoas que farão uso dele. Concentre-se em desenhá-lo e criá-lo com base nas necessidades dos usuários.

Quando tempo devo gastar nisso?

Desenvolvedores na redação são como água no deserto: altamente procurados e em escassez. Criar aplicativos jornalísticos significa equilibrar as necessidades diárias de uma redação com o compromisso de longo prazo de construir bons produtos.

Digamos que seu editor venha com uma ideia: a Câmara Municipal vai ter uma votação na próxima semana sobre a demolição de várias propriedades históricas na cidade. Ele sugere que se faça um aplicativo simples que permita aos usuários ver os prédios em um mapa.

Como desenvolvedor, você tem algumas opções. Pode construir um bonito mapa com software personalizado ou usar ferramentas que já existem, como o Google Fusion Tables para mapear bibliotecas e terminar o trabalho em algumas horas. A primeira opção vai resultar em um aplicativo melhor, mas a segunda deve economizar tempo para outro trabalho com maiores chances de ter um impacto duradouro.

O fato de a reportagem permitir a elaboração de um lindo e complexo aplicativo não significa que você precise construi-lo. Equilibrar prioridades é crucial. O truque é lembrar que cada aplicativo desenvolvido tem um custo: no caso, outro importante aplicativo poderia ser produzido no lugar dele.

Como levar as coisas a um nível superior?

Produzir aplicativos jornalísticos de alta qualidade pode ser caro e demandar muito tempo. É por isso que vale a pena perguntar sobre o retorno. Como você transforma um trabalho de sucesso temporário em algo especial?

Criar um trabalho duradouro que transcenda o ciclo da notícia é um caminho. Mas também é possível criar uma ferramenta que economize o seu tempo (e abrir o código dela!), ou usar estatísticas avançadas de uso do aplicativo para saber mais sobre o comportamento do seu público.

Muitas organizações desenvolvem mapas com dados do Censo para mostrar mudanças demográficas em suas cidades. Mas quando a equipe do Chicago Tribune lançou umaplicativo próprio sobre o tema, eles levaram essa tarefa a

um outro nível. O time desenvolveu técnicas e ferramentas para construir esse tipo de mapa com mais rapidez. O material também foi liberado posteriormente para que outras organizações pudessem usá-lo.

Onde eu trabalho, no Center for Investigative Reporting, juntamos um simples banco de dados com possibilidade de pesquisa a um sofisticado quadro de rastreamento que nos permitiu aprender, entre outras coisas, quanto, nos nossos aplicativos jornalísticos, os usuários valorizavam a a exploração aleatória e a pesquisa focada.

Com o risco de soar sovina, sempre pense no retorno do investimento. Resolva um problema genérico; crie uma nova maneira de engajar usuários; abra os códigos de parte do trabalho; use métodos de análise para aprender mais sobre seus usuários; ou, até mesmo, encontre casos como o Curbwise, no qual parte do aplicativo desenvolvido pode gerar receita.

Embalando

O desenvolvimento de aplicativos jornalísticos percorreu um longo caminho em um curto período de tempo. Aplicativos jornalísticos 1.0 eram muito parecidos com infográficos 2.0: visualização interativa de dados misturada a banco de dados com possibilidade de busca, feitos para otimizar a narrativa da reportagem. Agora, muitos desses aplicativos podem ser feitos por repórteres que precisam cumprir prazos com ferramentas de código aberto, liberando os desenvolvedores para pensar projetos maiores.

Aplicativos jornalísticos 2.0 são a arte de combinar a narrativa e o serviço público do jornalismo ao desenvolvimento de um produto e à expertise da tecnologia mundial. O resultado, sem dúvida, será uma explosão de inovação sobre maneiras de tornar os dados relevantes, interessantes e especialmente úteis para o público — e, ao mesmo tempo, ajudar o jornalismo a fazer o mesmo.

— Chase Davis, Center for Investigative Reporting

Aplicativos jornalísticos no ProPublica

Um aplicativo jornalístico é um grande e interativo banco de dados que conta uma história. Encare o aplicativo como qualquer outra peça jornalística. A diferença é que ele usa código de programação em vez de palavras e fotos.

Ao mostrar dados importantes a cada um dos leitores, o aplicativo pode ajudar a compreender a notícia de forma particular, relevante para cada contexto. Ele é capaz de ajudar o leitor a compreender sua ligação pessoal com um fenômeno nacional amplo e associar o que o leitor conhece àquilo que não conhece, incentivando a compreensão de conceitos abstratos.

Tendemos a construir um aplicativo jornalístico quando temos um conjunto de dados (ou acreditamos que possamos adquirir um conjunto de dados) de abrangência nacional mas com granularidade suficiente para revelar detalhes importantes.

Um aplicativo jornalístico deve contar uma história e, como em qualquer boa história, precisa ter uma manchete, um subtítulo, um lide e um olho gráfico. Pode ser difícil distinguir esses conceitos numa ferramenta interativa, mas eles estão lá se você procurar com cuidado.

Ao mesmo tempo, um aplicativo noticioso deve ser um dínamo, precisa gerar mais pautas, mais investigação, mais reportagem. Os melhores aplicativos do ProPublica foram usados como base para séries de reportagens em jornais locais.

Veja, por exemplo, nosso aplicativo jornalístico Dólares para os Médicos. A ferramenta rastreou, pela primeira vez, os milhões de dólares que empresas farmacêuticas pagam a médicos por consultorias, palestras e por aí vai. A ferramenta permite procurar os seus próprios médicos e checar que pagamentos receberam da indústria farmacêutica. Jornalistas em outras redações também usaram a ferramenta. Mais de 125 redações, incluindo o The Boston Globe, o Chicago Tribune e o The St. Louis Post-Dispatch, publicaram reportagens investigativas sobre médicos nos seus estados a partir dos dados do Dólares para os Médicos.

Poucas dessas reportagens foram resultado de parceria formal. A maioria foi produzida de maneira indenpendente — em alguns casos, não tínhamos muito conhecimento de que a reportagem estava sendo preparada até a publicação. Como somos uma pequena organização mas temos abrangência nacional, esse

tipo de repercussão é crucial para nós. Não temos conhecimento local em 125 cidades diferentes, mas se nossos dados ajudarem repórteres que tenham fontes locais a contar histórias com impacto, estamos desempenhando nossa missão.

Um dos meus aplicativos jornalísticos preferidos é o Mapeando Los Angeles, do Los Angeles Times, que começou a mapear os muitos bairros da cidade de forma colaborativa. Até o lançamento do Mapeando L.A., não havia consenso sobre as fronteiras entre os bairros. Após o projeto, o L.A. Times tem usado esses bairros como esqueleto e base para belos exemplos de jornalismo de dados — como mostrar taxas de criminalidade em cada bairro, qualidade do ensino escolar, etc., algo que eles não conseguiriam fazer antes. Ou seja, além de ser amplo e específico ao mesmo tempo, o Mapeando L.A. também é um gerador de histórias: conta histórias pessoais de cada leitor.

Os recursos necessários para construir um aplicativo jornalístico podem variar bastante. O The New York Times tem dúzias de pessoas trabalhando em aplicativos e infográficos interativos. Mas o site Talking Points Memo fez um aplicativo de ponta para rastrear pesquisas eleitorais com apenas dois funcionários — e nenhum tinha diploma de ciências da computação.

Como a maior parte do desenvolvedores que trabalham em redações, usamos uma adaptação da metodologia Agile para construir nossos aplicativos. Esboçamos rapidamente diferentes versões e mostramos o material para o pessoal da redação. O mais importante é que trabalhamos muito próximos aos repórteres e também acompanhamos os rascunhos de seus textos — mesmo os mais crus. Trabalhamos muito mais como repórteres do que como desenvolvedores tradicionais. Além de escrever código, conversamos com fontes, levantamos informações, nos tornamos especialistas no tema. Seria bem difícil criar um bom aplicativo jornalístico a partir de material que não compreendemos.

Por que uma redação deve se interessar por desenvolver aplicativos jornalísticos baseados em dados? Por três razões: é jornalismo de qualidade, é muito popular — os produtos do ProPublica que fizeram mais sucesso são aplicativos — e, se não fizermos, o concorrente fará. Pense em todos os furos de reportagem que perderíamos! Mais importante, as redações precisam entender que podem criar tudo isso também. É mais fácil do que parece.

— Scott Klein, ProPublica

A visualização como carro-chefe do jornalismo de dados

Antes de traçar ou mapear seus dados, reserve um minuto para pensar sobre os muitos papéis que elementos gráficos, estáticos ou interativos, têm no jornalismo.

Durante a apuração, visualizações podem:

- Ajudar a identificar temas e perguntas para o resto da reportagem
- Identificar valores atípicos: boas histórias, ou talvez erros, nos seus dados
- Ajudar a encontrar exemplos
- Mostrar falhas em sua reportagem

Visualizações também exercem múltiplos papéis na hora da publicação. Elas podem:

- Ilustrar um ponto levantado no texto de forma mais atraente
- Excluir do texto dados técnicos desnecessários
- Quando são interativas e permitem um certo grau de exploração, deixam mais transparente o processo de apuração

Esses papéis sugerem que você deve começar cedo a incluir visualizações na sua reportagem, mesmo que não inicie ao mesmo tempo o trabalho eletrônico com os dados. Não considere a visualização uma etapa separada, após a redação de grande parte da matéria. Deixe-a ajudar a guiar suas reportagens.

Às vezes, começar significa apenas colocar uma forma visual nas anotações que você já fez. Considere este gráfico abaixo, publicado pelo Washington Post em 2006.

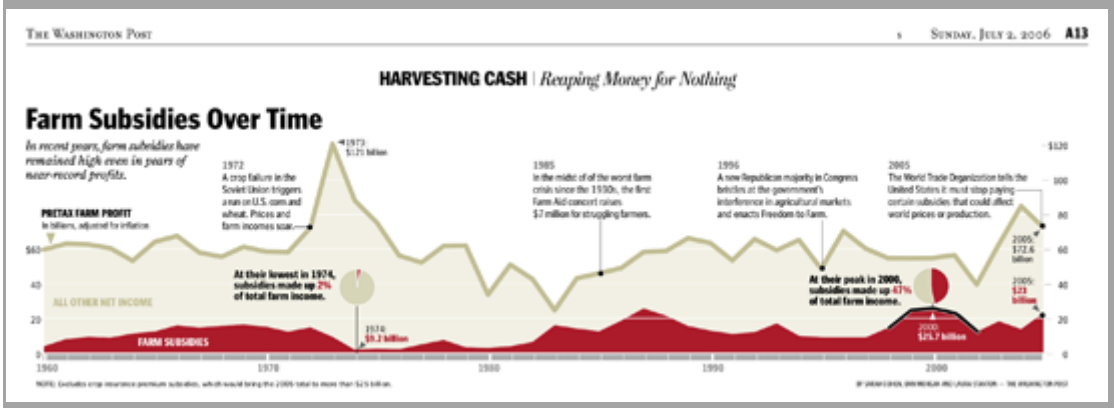


Imagem 2. Subsídios agrícolas ao longo do tempo (Washington Post)

Ele mostra a renda agrícola dos últimos 45 anos associada a subsídios e eventos importantes. Esta visualização levou meses para ficar pronta. Foi um desafio encontrar dados com definições e significados semelhantes que podiam ser comparados ao longo do tempo. Investigar todos os altos e baixos do processo nos ajudou a manter sempre em mente o contexto geral da reportagem até o fim da apuração. Também significou que uma parte importante da investigação foi concluída mesmo antes de as histórias começarem a ser escritas.

Aqui vão algumas dicas de uso da visualização para começar a explorar séries de dados.

Dica 1: Utilize múltiplos pequenos para se orientar rapidamente em meio a um grande conjunto de dados

Usei essa técnica no Washington Post quando analisamos uma sugestão de pauta ligada à administração de George W. Bush. Recebemos uma informação que dizia que o governo norte-americano estava concedendo subsídios por motivos políticos, e não de mérito. A maioria desses programas de ajuda são calculados por fórmula, e outros têm sido financiados por anos. Portanto, ficamos curiosos se poderíamos encontrar um padrão nesse processo, olhando para cerca de 1.500 diferentes fluxos de informações.

Criei um gráfico para cada programa, com os pontos vermelhos indicando um ano de eleição presidencial e os verdes, um ano de eleição legislativa. O problema: sim, houve um aumento em de financiamento em vários desses programas nos seis meses anteriores à eleição presidencial (os pontos vermelhos, com os números de pico ao lado deles). Mas isso aconteceu no ano que não procurávamos. Em vez de encontrarmos os picos durante a tentativa de reeleição de George W. Bush em 2004, os picos apareciam nas eleições de 2000, quando Bill Clinton estava na Casa Branca e seu vice, Al Gore, concorria.

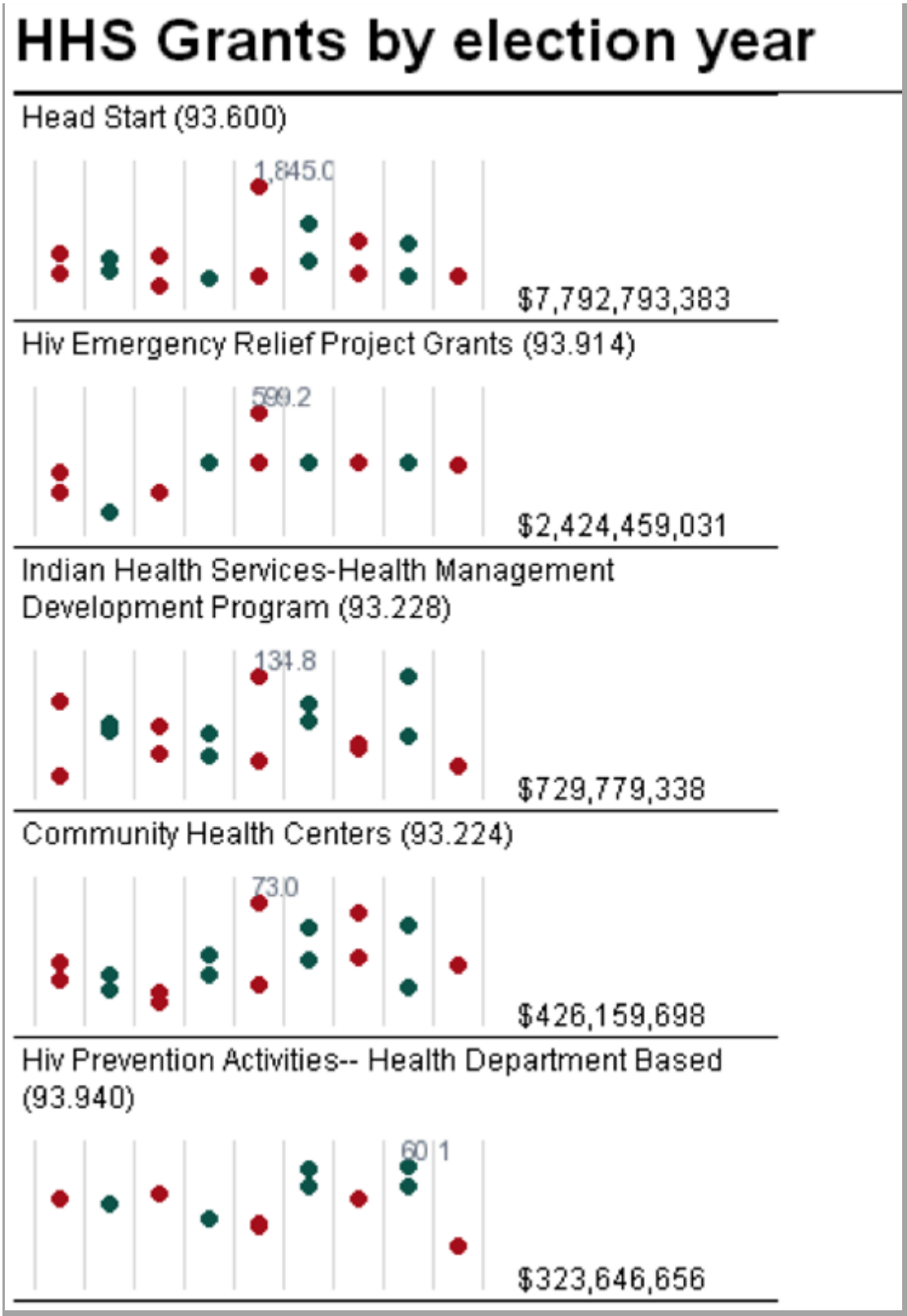


Imagem 3. Gráficos ajudam a focalizar a história (Washington Post)

Neste caso, foi realmente mais fácil ver isso em uma série de gráficos, e não em uma tabela de números. Um formulário interativo nos permitiu checar vários tipos de subsídios, regiões e agências. Mapas divididos em múltiplos pequenos

podem ser o caminho para mostrar o tempo e o local em um gráfico estático. Este tipo de visualização é bem fácil de comparar — às vezes, até mais fácil do que uma que seja interativa.

Este exemplo foi criado com um pequeno programa escrito em PHP, mas agora é muito mais fácil fazer algo do tipo por meio dos gráficos das versões 2007 ou 2010 do Excel. Edward Tufte, especialista em visualização, inventou estes "gráficos intensos e simples, que lembram o uso da palavra", para transmitir rapidamente informações de uma grande base de dados. Você agora pode vê-los em todos os lugares, dando forma a informações tão variadas como resultados esportivos e cotações do mercado de ações.

Dica 2: Olhe para os dados de cima para baixo e de um lado para o outro

Ao tentar entender uma história ou um conjunto de dados, não há formas erradas de olhar para eles. Busque os mais diversos pontos de vista para obter uma perspectiva diferente. Se estiver escrevendo sobre crime, por exemplo, você pode analisar um grupo de gráficos sobre a evolução dos crimes violentos em um ano; a variação percentual; a comparação entre as taxas de crimes de várias cidades; e o comportamento do crime ao longo do tempo. Use números brutos, percentuais e índices.

Olhe para eles em diferentes escalas. Tente seguir a regra de que o eixo X deve ser zero. Em seguida, quebre essa regra e veja se você aprende mais com isso. Experimente logaritmos e raízes quadradas para dados com distribuições ímpares.

Tenha em mente as pesquisas realizadas na área de percepção visual. Experimentos de William Cleveland mostraram que os olhos veem mudanças em uma imagem quando a inclinação média é de cerca de 45 graus. Isso sugere que é preciso ignorar as recomendações de que sempre devemos começar do zero, indicando a necessidade de trabalhar gráficos que permitam ver mais coisas. Outro estudo da área de epidemiologia sugere que níveis altos são entendidos como limites para o gráfico. Cada nova perspectiva ajuda a ver mais informações dentro dos dados. Quando eles pararem de revelar novidades, seu trabalho está concluído.

Dica 3: Não faça suposições

Agora que você já olhou os dados de várias maneiras, provavelmente encontrou registros que parecem incorretos — você não entende o que eles dizem, ou

existem valores atípicos que parecem erros de digitação, ou há tendências que parecem o oposto do que deveriam ser.

Se quiser publicar qualquer coisa com base em suas investigações iniciais ou em uma visualização, é preciso resolver essas questões e não fazer nenhum tipo de suposição. Ou elas são histórias interessantes ou são erros; ou são desafios interessantes para o senso comum ou mal-entendidos.

É comum ver governos locais oferecendo planilhas cheias de erros, e também é fácil de entender erroneamente o jargão do governo em um conjunto de dados.

Em primeiro lugar, reveja o seu próprio trabalho. Você já leu a documentação dos dados, suas advertências e viu se o problema está também na versão original dos dados? Se tudo o que fez parece estar correto, então é hora de pegar o telefone. Você vai ter de resolver essa dúvida para poder usar a base de dados.

Dito isto, nem todo erro é importante. Nos registros de financiamento de campanha, é comum ter várias centenas de códigos postais que não existem em um banco de dados de 100.000 registros. Desde que não sejam todos da mesma cidade ou estejam relacionados a um mesmo candidato, o registro ocasionalmente equivocado simplesmente não importa.

A questão a se perguntar é: se usar isso, os leitores terão uma visão fundamentalmente precisa do que os dados dizem?

Dica 4: Evite ficar obcecado com a precisão

Não fazer perguntas suficientes é ruim, mas há um outro extremo: ficar obcecado com a precisão sem que isso importe. Seus gráficos exploratórios devem ser corretos no geral, mas não se preocupe se tiver de fazer arredondamentos, se os números não somam exatamente 100 por cento ou se faltam informações de um ou dois anos em um período de 20 anos. Tudo isso é parte da apuração e não impedirá você de ver as grandes tendências, assim como saber o que pesquisar antes da publicação.

Na verdade, você pode considerar a eliminação de marcadores e indicadores de escala, como nos gráficos acima, para ter uma melhor visão do sentido geral dos dados.

Dica 5: Crie cronologias de casos e eventos

No início de toda história complexa, comece a montar cronologias de casos e eventos-chave. Você pode usar o Excel, um documento de Word ou uma

ferramenta especial como o TimeFlow para a tarefa, mas em algum momento encontrará um conjunto de dados que pode usar como referência. A releitura periódica do material vai mostrar os buracos do seu trabalho que ainda precisam ser preenchidos.

Dica 6: Reúna-se sempre e desde o princípio com seu departamento gráfico

Troque ideias sobre a produção dos gráficos com ilustradores e designers de sua redação. Eles podem indicar boas alternativas de visualização dos dados, sugerir formas de interação e também dar ideias sobre como conectar dados e histórias. Sua tarefa será muito mais fácil se souber, desde o começo, o que tem de pesquisar ou, então, se deve alertar sua equipe de que não é possível fazer um determinado tipo de gráfico quando não se tem os dados necessários.

Dicas para publicação

Você pode ter gasto apenas alguns dias ou algumas horas na apuração, ou ter levado meses para reunir as informações necessárias para a sua história. Mas quando chega o momento de publicá-la, precisa ficar atento a dois importantes aspectos.

Lembra daquele ano sobre o qual faltavam informações e que deixou sua apuração incompleta? De repente, você se dá conta de que não pode mais avançar na investigação sem esses dados. E todas aquelas informações problemáticas que acabaram sendo ignoradas? Reaparecem para assombrá-lo. A questão é que não dá para escrever sobre dados ruins. Não há solução intermediária para um gráfico: ou se tem tudo o que é necessário para construí-lo, ou não se tem.

Combine o esforço de coleta de dados com o gráfico interativo

Não há esconderijo em um gráfico interativo. Se você realmente vai permitir que seus leitores explorem seus dados da forma como quiserem, então cada um de seus elementos tem que ser o que diz ser. Os usuários podem encontrar erros a qualquer momento no material, e isso pode assombrá-lo por meses ou anos. Se você constrói o seu próprio banco de dados, também deve revisá-lo, checá-lo e editar todo o conteúdo. Se estiver usando informações governamentais, é preciso decidir qual será o nível de apuração desses dados e o que vai fazer quando encontrar um erro inevitável.

Design para dois tipos de leitores

O gráfico — seja um elemento interativo autônomo ou uma visualização estática ao lado da reportagem — deve satisfazer dois tipos de leitores. Deve ser fácil de entender à primeira vista, mas também complexo o suficiente para oferecer algo interessante a quem queira se aprofundar nas informações. Se seu gráfico se tornar interativo, certifique-se de que seus leitores vão obter algo mais do que um único número ou nome.

Transmita uma ideia e, depois, simplifique

Certifique-se de que há uma única coisa específica que você quer que as pessoas vejam. Decida qual é a impressão geral que deseja transmitir ao leitor e faça todo o resto desaparecer. Em muitos casos, isso significa remover as informações, mesmo quando a internet permite ampliar o contexto. A menos que seu principal objetivo seja garantir a transparência do trabalho jornalístico, a maioria dos detalhes reunidos em sua linha do tempo e cronologia simplesmente não são importantes. Em um gráfico estático, são intimidantes. Em um gráfico interativo, chatos.

— Sarah Cohen, Universidade de Duke

Usando visualização para contar histórias

A visualização de dados merece ser considerada por várias razões. Não somente porque pode ser belíssima e chamar atenção — elemento valioso para ser compartilhada e atrair leitores — , mas também porque conta com uma poderosa vantagem cognitiva. Metade do cérebro humano é dedicada ao processamento de informação visual. Quando você apresenta um gráfico com informações a um usuário, consegue ser mais efetivo para chegar à mente dele. A visualização de dados, quando bem projetada, pode dar uma impressão imediata e profunda aos espectadores, acabando com a desorganização de uma história complexa e indo direto ao ponto.

Mas ao contrário de outros recursos visuais (como a fotografia e o vídeo), a visualização de dados está profundamente enraizada em fatos mensuráveis. Embora seja esteticamente envolvente, tem menos carga emocional e se preocupa mais com o esclarecimento do que com a parte emocional do tema. Em uma época de meios de comunicação muitas vezes focados em públicos específicos, a visualização de dados (e o jornalismo de dados em geral) oferece a oportunidade tentadora de narrar histórias orientadas principalmente pelos fatos, não pelo fanatismo.

Além disso, a visualização pode ser eficaz tanto para apresentar notícias — transmitindo rapidamente informações pontuais como a localização de um acidente e o número de vítimas — como para reportagens, nas quais pode aprofundar um tema e oferecer uma nova perspectiva sobre algo familiar.

Enxergando o que é familiar de uma nova maneira

A capacidade da visualização de dados para testar algo que seja consenso é exemplificada por um gráfico interativo publicado pelo The New York Times no final de 2009, um ano após o início da crise econômica mundial. Com a taxa de desemprego dos Estados Unidos pairando os 9%, os usuários podiam filtrar a população do país com critérios demográficos e educacionais para ver quão grande era a variação do desemprego. O resultado é que a taxa ia de menos de 4%, entre mulheres de meia-idade com alto grau de instrução, a 50%, quando o grupo era de jovens negros que não concluíram o Ensino Médio. Essa disparidade não era novidade — um fato sublinhado por valores históricos de cada um desses grupos.

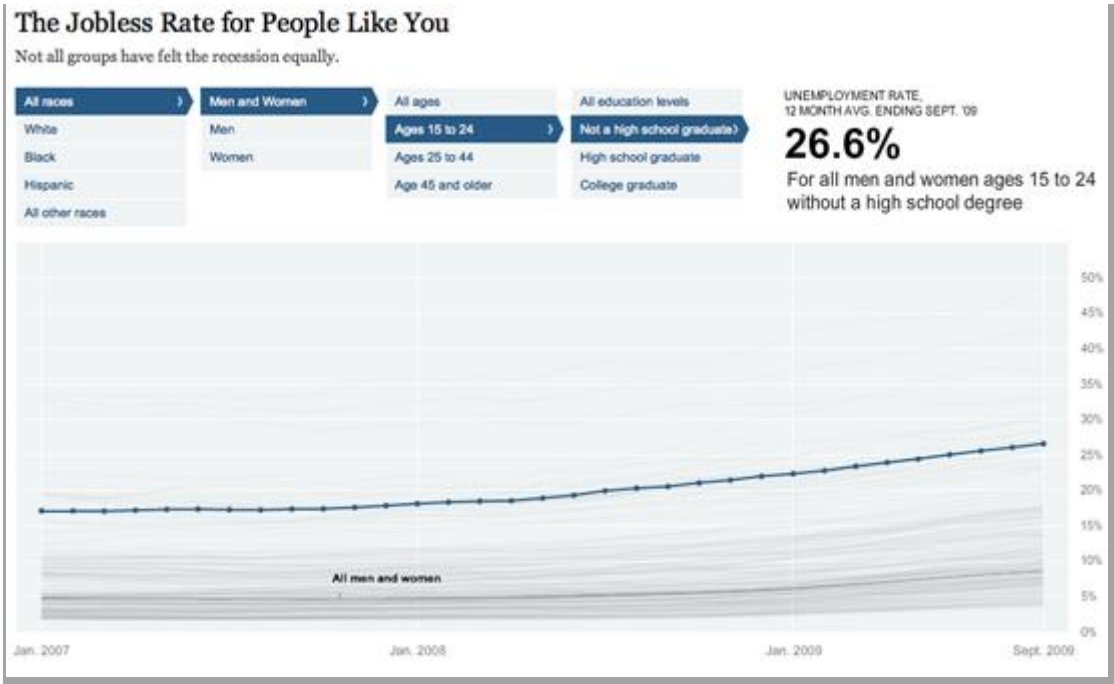


Imagem 4. A taxa de desemprego para pessoas como você (The New York Times)

Mesmo depois que você para de observá-la, uma boa visualização de dados fica na memória e deixa um modelo mental duradouro de um fato, tendência ou processo. Quantas pessoas viram as animações de tsunamis apresentadas por pesquisadores em dezembro de 2004, que mostravam ondas em cascata sendo irradiadas pelo Oceano Índico por conta de um terremoto indonésio, ameaçando milhões de moradores de áreas costeiras no sul da Ásia e leste da África?

A visualização de dados — e as associações estéticas que ela engendra — pode até se tornar uma referência cultural, tal como a representação de profundas divisões políticas nos Estados Unidos após as eleições de 2000 e 2004. Naquele momento, estados republicanos "vermelhos" encheram a área central, e os democratas "azuis" agruparam-se no nordeste e no extremo oeste. Não importa que, na mídia dos EUA anterior ao ano 2000, as principais redes de televisão alternassem livremente o vermelho e o azul para representar cada partido, algumas fazendo isso de quatro em quatro anos. Daí a lembrança de alguns norte-americanos da vitória épica de Ronald Reagan em 49 estados "azuis" para os republicanos em 1984.

Mas para cada gráfico que gera um clichê visual, aparece outro com um poderoso testemunho factual, como no mapa de 2006 do The New York Times. O material usou círculos de uma forma diferente para mostrar onde centenas de milhares de pessoas retiradas de Nova Orleans após o furacão Katrina estavam

vivendo, espalhadas por todo o país devido a uma mistura de conexões pessoais e programas sociais. Será que essas pessoas farão o caminho de volta para casa?

Portanto, agora que já discutimos o poder da visualização de dados, é justo perguntar: quando devemos usá-la, e quando *não* devemos usá-la? Primeiro, olhemos para alguns exemplos em que ela pode ser útil para ajudar a contar uma história.

Mostrando a mudança através do tempo

Talvez o uso mais comum da visualização de dados — personificada por um simplório gráfico de linha — é mostrar como os valores mudaram ao longo do tempo. O crescimento da população da China desde 1960 ou o aumento da taxa de desemprego desde a crise econômica de 2008 são bons exemplos. Mas a visualização de dados também pode mostrar, com outras formas gráficas, a mudança ao longo do tempo. O pesquisador português Pedro M. Cruz utiliza gráficos de círculo animados para mostrar o declínio radical dos impérios europeus ocidentais. Dimensionado pfaela população total, Grã-Bretanha, França, Espanha e Portugal estouram como bolhas quando seus territórios ultramarinos alcançam a independência. Lá se vão México, Brasil, Austrália, Índia...

Um gráfico do Wall Street Journal mostra o número de meses que uma centena de empresários levou para alcançar o número mágico de US$ 50 milhões em receita. Criado com mapas abertos e a ferramenta de análise de dados Tableau Public, a comparação assemelha-se aos rastros deixados por vários aviões ao decolar, alguns mais rápidos, outros lentos.

Falando de aviões, outro gráfico interessante que mostra mudanças ao longo do tempo traz a participação de mercado das principais companhias aéreas norte-americanas durante décadas. Após a administração Carter desregulamentar a aviação de passageiros nos EUA, ocorreu uma série de aquisições financiadas com empréstimos que criou companhias as aéreas nacionais, como mostra este gráfico do The New York Times.

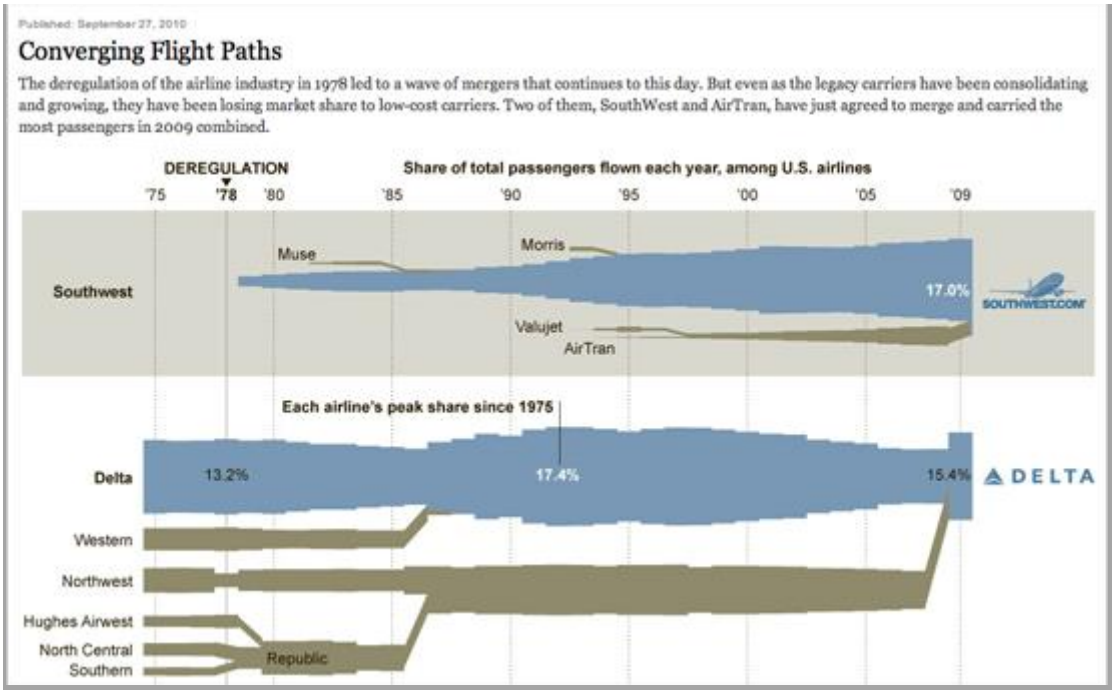


Imagem 5. Rotas de voos convergentes (The New York Times)

Tendo em conta que quase todos os leitores eventuais enxergam o eixo horizontal X' de um gráfico como o que representa o tempo, às vezes, é fácil pensar que *todas* as visualizações deveriam mostrar mudanças ao longo do tempo.

Comparando valores

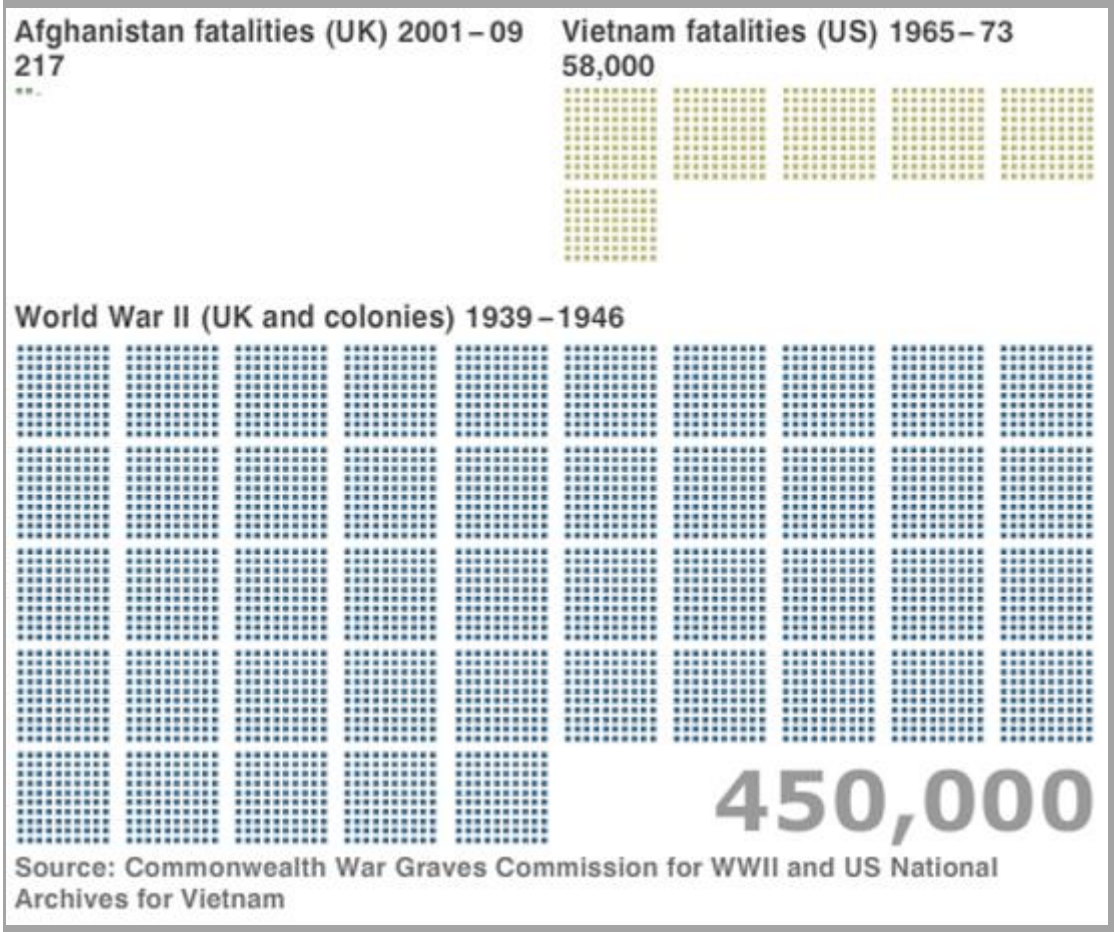


Imagem 6. Contando o custo humano da guerra (BBC)

A visualização de dados também é útil ao ajudar os leitores a comparar dois ou mais valores. Pode, por exemplo, contextualizar a perda trágica de militares no Iraque e no Afeganistão (comparando-a com os mortos no Vietnã na Segunda Guerra Mundial, como fez a BBC em um slideshow animado). Também pode, neste gráfico minimalista da National Geographic, mostrar como é mais provável morrer de doença cardíaca (1 em 5 mortes) ou acidente vascular cerebral (1 em 24) do que, digamos, acidentes de avião (1 em 5.051), ou picada de abelha (1 em 56.789) — todas ofuscadas por um grande arco que representa a probabilidade de morte em geral: 1 em 1!

A BBC, em colaboração com a agência de design Berg, também desenvolveu o site"Dimensões", que mostra o contorno do impacto de grandes eventos mundiais — o derramamento de petróleo da plataforma marinha Deepwater Horizon, ou as inundações no Paquistão, por exemplo, a um mapa do Google de seu país.

Mostrando fluxos e conexões

A introdução do trem de alta velocidade na França, em 1981, não fez o país ficar menor, mas uma representação visual inteligente mostra quanto tempo se economiza para chegar a diferentes destinos em comparação com o trem convencional. Um gride sobre o país aparece sobreposto à imagem de "antes", mas é esmagado em direção ao centro, Paris na de "depois", mostrando não apenas que os destinos estão mais próximos, mas também que a maior economia de tempo se dá na primeira parte da viagem, antes de os trens desacelerarem para passar por vias precárias.

Para comparações entre duas variáveis distintas, veja o gráfico de Ben Fry que avalia o desempenho dos principais times de beisebol da Liga segundo o salário de seus jogadores. Uma linha traçada em vermelho (baixo desempenho) ou azul (performance acima da média) conecta os dois valores, transmitindo de forma rápida a sensação de que os proprietários de times estão se arrependendo de seus jogadores caros. Além disso, uma linha do tempo oferece um retrato fiel da concorrência presente no campeonato.

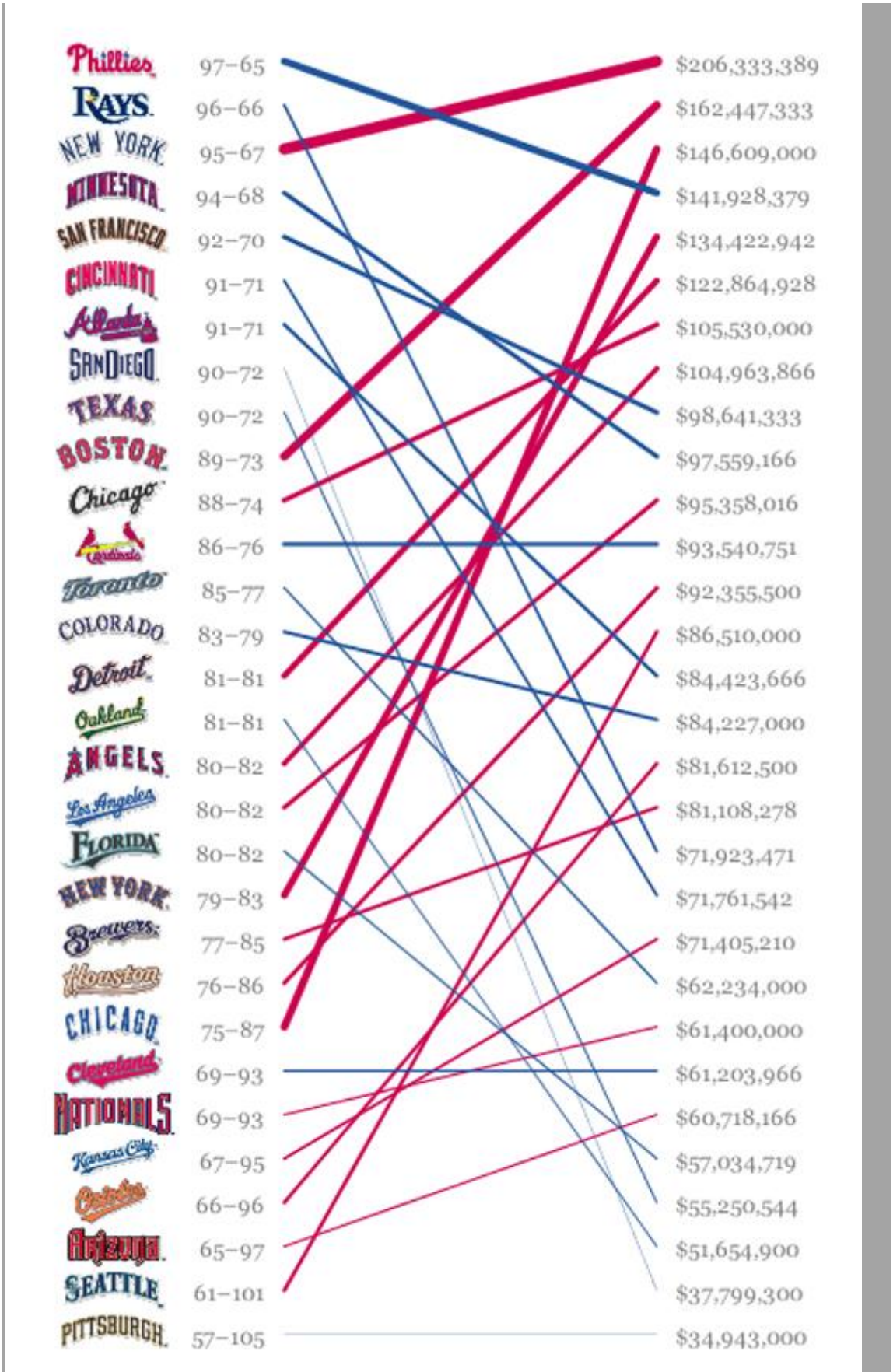


Imagem 7. Salário vs. desempenho (Ben Fry)

Desenhando com dados

Similares às conexões gráficas, os diagramas de fluxo também codificam informações nas linhas de ligação, geralmente pela espessura e/ou a cor. Por exemplo, com a crise da Zona do Euro e vários membros incapazes de quitar suas dívidas, o The New York Timesprocurou desvendar a teia de empréstimos que prendia os integrantes da UE aos seus parceiros comerciais além do Atlântico e na Ásia. Em um dos "estados" da visualização, a largura da linha reflete o montante de crédito que passa de um país para o outro, e tons que vão do amarelo ao laranja indicam o quanto a situação é "preocupante" — ou seja, pouco provável de que o dinheiro seja devolvido.

Numa temática mais alegre, a revista National Geographic produziu uma tabela enganosamente simples, mostrando as conexões de três cidades dos EUA — Nova York, Chicago e Los Angeles — às principais regiões produtoras de vinho. A peça revela como os meios de transporte que trazem o produto de cada uma das fontes podem resultar em grandes pegadas de carbono, fazendo com que seja mais ecológico para os nova-iorquinos comprar o vinho de Bordeaux em vez do californiano.

"Mapa de origem", um projeto iniciado na escola de negócios do MIT, usa diagramas de fluxo para analisar com rigor a cadeia global de produtos manufaturados, seus componentes e matérias-primas. Graças a uma série de pesquisas, o usuário pode procurar produtos que vão desde sapatos da marca Ecco até suco de laranja, e descobrir em que lugar do planeta foi produzido e qual é a sua pegada de carbono.

Mostrar hierarquia

Em 1991, o pesquisador Ben Shneiderman inventou uma nova forma de visualização chamada "treemap", que consiste em várias caixas concêntricas, umas aninhadas dentro das outras. A área de cada caixa indica a quantidade que representa. O treemap é uma interface compacta e intuitiva para o mapeamento de uma entidade e suas partes constituintes. Seja para visualizar o orçamento nacional dividido entre instituições oficiais e contratadas; o mercado de ações por setor e empresa; ou mesmo uma linguagem de programação por classes e subclasses. Outro recurso eficaz é o dendrogram — formato que se parece a um organograma típico, no qual as subcategorias saem de um único tronco central.

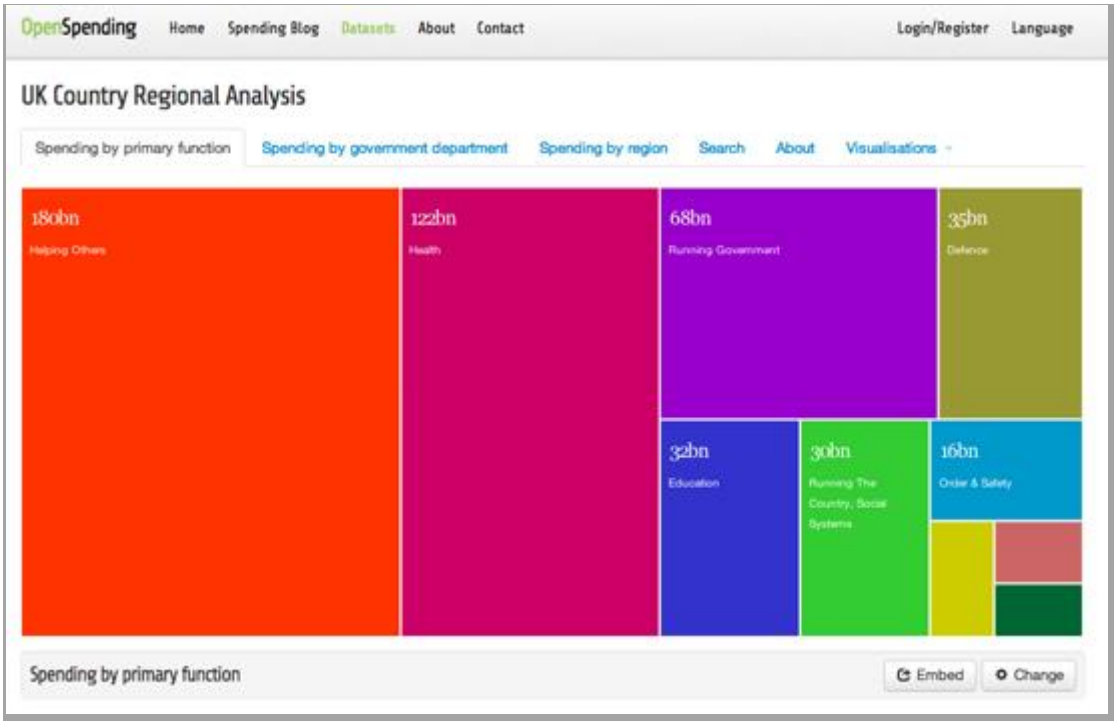


Imagem 8. OpenSpending.org (Open Knowledge Foundation)

Navegando por grandes bancos de dados

Às vezes, a visualização de dados é muito eficaz ao exibir informações familiares a partir de ângulos novos. Mas o que acontece quando você tem informações inéditas? A era dos dados traz descobertas surpreendentes quase todos os dias, da brilhante análise geográfica das fotos do Flickr de Eric Fischer até a divulgação de milhares de avaliações confidenciais de professores de Nova York.

Essas bases de dados são mais poderosas quando os usuários podem se aprofundar no conteúdo e navegar pelas informações mais relevantes para eles.

No início de 2010, o The New York Times teve acesso aos registros do Netflix, normalmente privados, sobre quais gêneros de filmes são mais alugados. Mesmo com a negação do Netflix de publicar números brutos, o Times criou um banco de dados interativo que permite que os usuários explorem os 100 filmes mais alugados em 12 áreas metropolitanas dos Estados Unidos, esmiuçadas até o nível do código postal. Um ou mapa de calor sobreposto a cada comunidade deixava conferir rapidamente onde um filme em particular era mais popular.

No final do mesmo ano, o jornal publicou os resultados do Censo dos Estados Unidos logo depois de ser divulgado. A interface, construída em Flash, ofereceu uma série de opções de visualização e permitiu aos usuários explorar dados de

cada bloco do estudo. Eles podiam ver, por exemplo, a distribuição da população por raça, renda e educação. Tal era a resolução do mapa que era fácil se perguntar se você era a primeira pessoa a consultar determinado dado em tamanho banco de dados poucas horas depois de sua publicação.

Igualmente louvável é a investigação da BBC sobre mortes na estrada e muitas das tentativas de indexar rapidamente grandes quantidades de dados como o WikiLeaks War Logs, sobre as guerras do Iraque e do Afeganistão.

A regra do 65 mil

Ao receber a primeira pilha de dados do WikiLeaks sobre a guerra no Afeganistão, a equipe encarregada de processá-la começou a demonstrar entusiasmo por poder ter acesso a 65 mil registros militares.

Isto imediatamente fez soar os alarmes entre os que já tinham experiência com Excel. Graças a uma limitação histórica ligada às linhas do programa, a ferramenta de importação do software não processa mais de 65.536 registros. Neste caso, verificou-se que faltavam "apenas" 25.000 linhas.

A moral da história (além de evitar o uso do Excel para tais tarefas) é desconfiar sempre de alguém que se vanglorie de ter 65.000 linhas de dados.

— Alastair Dant, Guardian

Imagem 9. Todas as mortes em estradas na Grã-Bretanha 1999-2010 (BBC)

Imaginando explicações alternativas

No The New York Times, o "gráfico porco-espinho" de Amanda Cox, com projeções otimistas do deficit dos EUA ao longo dos anos, mostra como, às vezes, o que aconteceu é menos interessante do que o que não ocorreu. O gráfico de linha de Cox, com o crescente déficit orçamentário após uma década de guerra e de incentivos fiscais mostra como as previsões podem ser irreais.

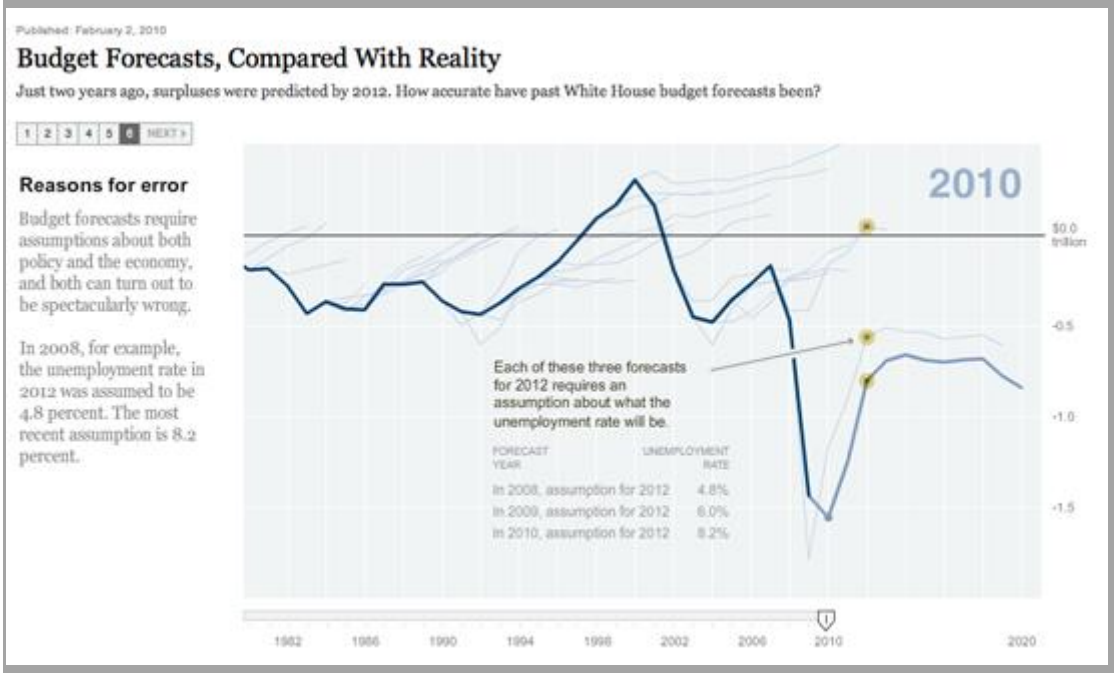


Imagem 10. Previsões orçamentárias comparadas com a realidade (New York Times)

Bret Victor, há muito tempo designer de interface da Apple (e criador da teoria "mate a matemática", de visualização para comunicar informação quantitativa), criou um protótipode uma espécie de documento que se atualiza em conjunto cada vez que um dado é modificado. Em seu exemplo, as dicas para economizar energia incluem premissas editáveis, em que um passo simples como desligar as luzes de quartos vazios pode ajudar os norte-americanos a economizar energia equivalente à produzida por entre 2 a 40 usinas de carvão. Alterando a porcentagem que aparece no meio de um parágrafo do texto, você modifica, ao mesmo tempo e de forma coerente, todo o resto da página!

Para mais exemplos e sugestões, aqui está uma lista de links de diferentes usos de visualizações, mapas e gráficos interativos organizada por Matthew Ericson, do The New York Times.

Quando não usar a visualização de dados

A visualização de dados, para ser eficaz, depende de informação boa, limpa, precisa e significativa. Assim como boas aspas, fatos e descrições alimentam o bom jornalismo narrativo, a visualização de dados é tão boa como as informações por trás dela.

Quando a sua história pode ser melhor contada com um texto ou recurso multimídia

Às vezes, os dados por si só não contam a história da maneira mais convincente. Um simples gráfico de uma tendência ou estatística pode ser útil, mas uma narrativa relacionando as consequências reais de um problema tem mais chances de causar um impacto maior no leitor.

Quando você tem pouquíssimos dados

Há um ditado que diz que "um número sozinho não diz nada". Uma frase comum dos editores de notícias em resposta a uma estatística citada é: "em comparação com o quê?" A tendência é subir ou baixar? O que é normal?

Quando seus dados variam pouco e não revelam uma tendência clara

Às vezes, você coloca os seus dados no Excel e acaba descobrindo que a informação é apenas um ruído, flutua demais ou tem tendências muito sutis. Você ergue a base do gráfico do zero até o o ponto mais baixo para ressaltar as diferenças? Não! Parece que você tem dados ambíguos e precisa fazer mais pesquisas e análises.

Quando um mapa não é um mapa

Às vezes, o elemento espacial não é significativo ou convincente. Também pode ser que distraia a atenção de tendências numéricas pertinentes, como a mudança ao longo do tempo ou a exibição de semelhanças entre áreas não adjacentes.

Quando uma tabela resolve

Se você tem relativamente poucos dados, mas conta com informações que podem ser úteis para os leitores, considere apenas apresentar os dados em uma tabela. É um recurso limpo, fácil de ler e não cria expectativas irreais com relação à matéria. Na verdade, as tabelas podem ser um formato muito elegante e eficiente para transmitir informação básica. — *Geoff McGhee, Universidade de Stanford*

Gráficos diferentes contam histórias diferentes

Neste mundo digital com promessa de experiências 3D imersivas tendemos a esquecer que, por muito tempo, só tivemos tinta e papel. Nós agora relegamos a esse meio estático um tratamento de segunda classe, mas, durante centenas de anos em que temos escrito e feito impressões, conquistamos uma incrível riqueza de conhecimento e de práticas para representar dados em uma página. Ao mesmo tempo que gráficos interativos, visualizações de dados e infográficos são a última tendência, eles abandonam muitas das melhores práticas que aprendemos. Somente quando você olha para trás, através da história de mapas e gráficos, pode entender aquele banco de conhecimento e trazê-lo para os novos meios.

Alguns dos mais famosos mapas e gráficos surgiram da necessidade de explicar melhor tabelas densas de dados. William Playfair foi um poliglota escocês que viveu entre o final dos anos 1700 até o início dos anos 1800. Sozinho, ele apresentou ao mundo muitos dos mesmos mapas e gráficos que nós ainda usamos atualmente. No seu livro de 1786,*Commercial and Political Atlas (Atlas Comercial e Político)*, Playfair apresentou o gráfico de barra para mostrar, de uma forma nova e visual, as importações e exportações da Escócia.

Ele popularizou o temido gráfico de pizza no seu livro de 1801 *Statistical Breviary (Breviário Estatístico)*. A demanda por essas novas formas de mapas e gráficos surgiu do comércio, mas, na medida em que o tempo foi passando, outros formatos apareceram e foram usados para salvar vidas. Em 1854, John Snow criou o seu, agora famoso, "Mapa da Cólera de Londres", adicionando uma pequena barra preta sobre cada endereço onde um incidente havia sido registrado. Com o tempo, uma densidade óbvia do surto podia ser vista e providências podiam ser tomadas para conter o problema.

Com o passar do tempo, praticantes destes novos mapas e gráficos tornaram-se mais ousados e experimentaram mais, forçando o setor a assumir a direção que conhecemos hoje. André-Michel Guerry foi o primeiro a publicar a ideia de um mapa em que regiões individuais eram representadas por cores diferentes com base em alguma variável. Em 1829, ele criou o primeiro mapa coroplético ao escurecer regiões da França para representar os níveis de criminalidade. Hoje em dia, vemos mapas assim sendo usados para mostrar quem votou em quem, distribuição de riqueza e outras muitas variáveis relacionadas geograficamente.

A ideia parece simples, mas, mesmo nos dias atuais, é difícil dominá-la e compreendê-la caso não seja usada com sabedoria.

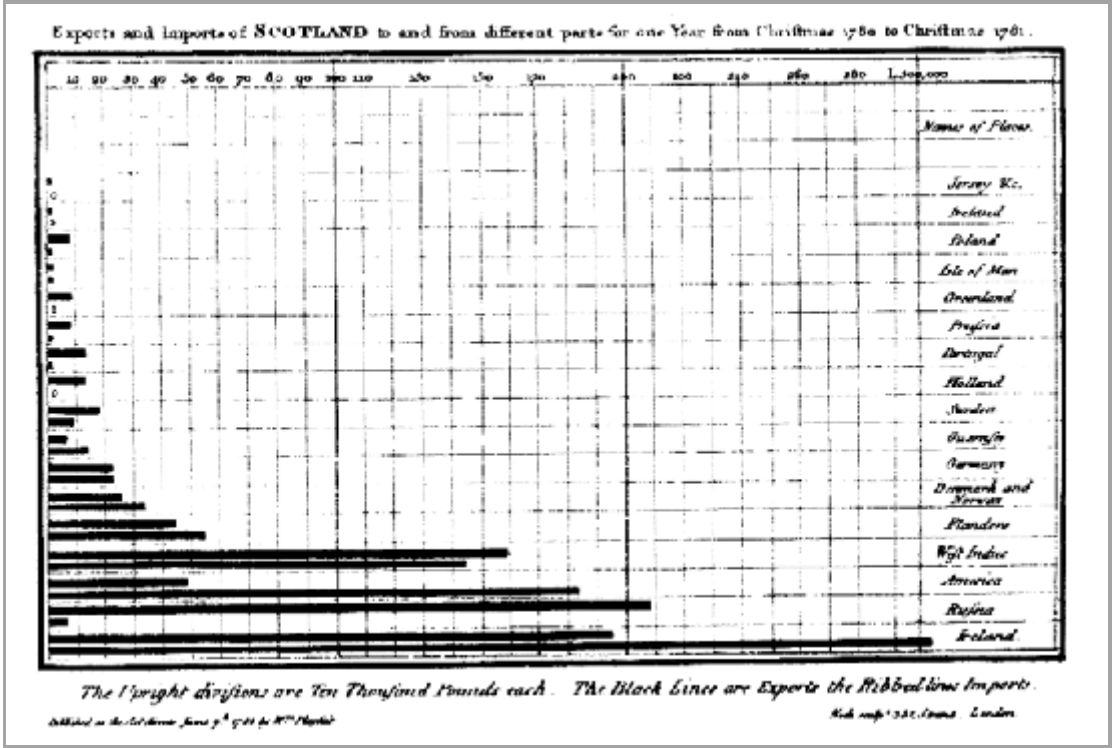


Imagem 11. Um dos primeiros gráficos de barra (William Playfair)

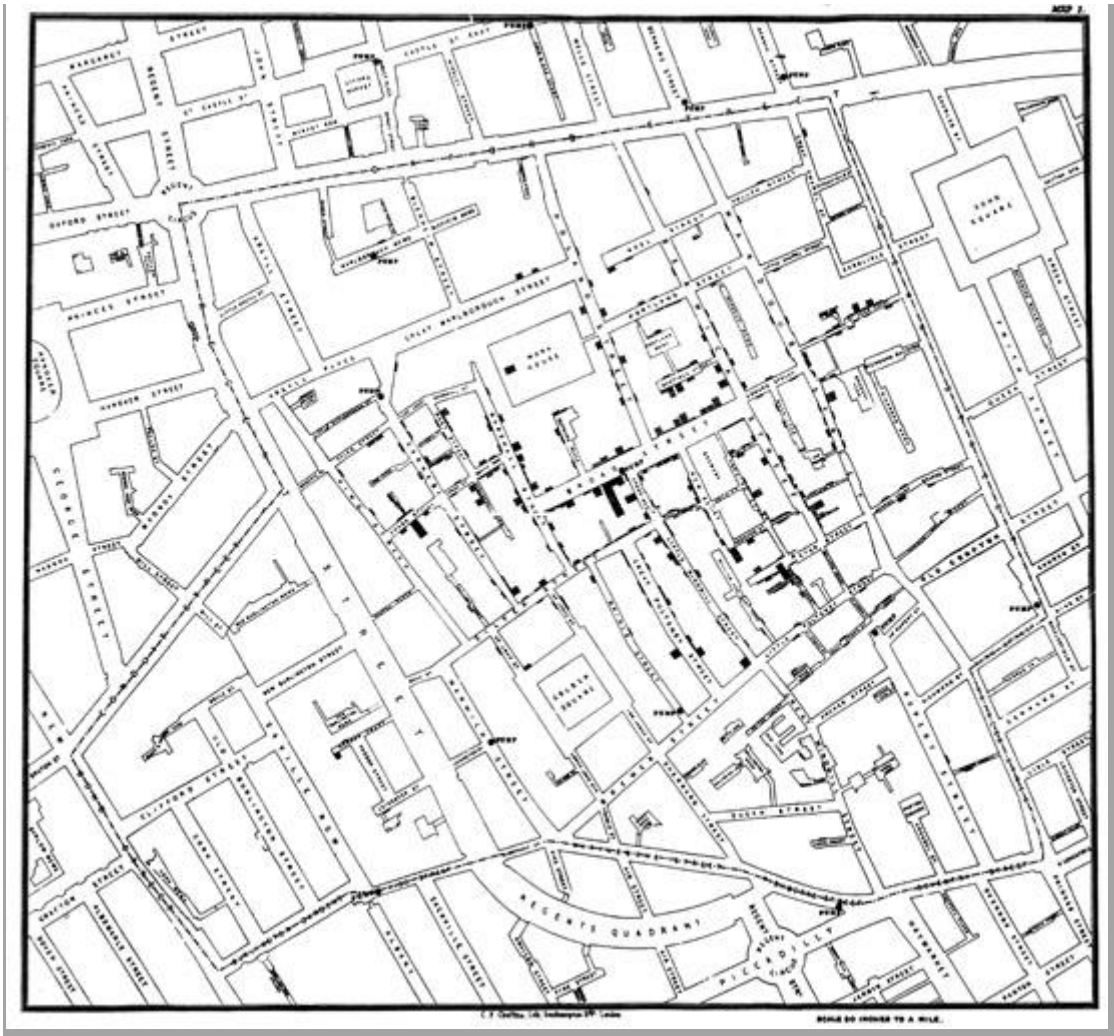


Imagem 12. Mapa da cólera em Londres (John Snow)

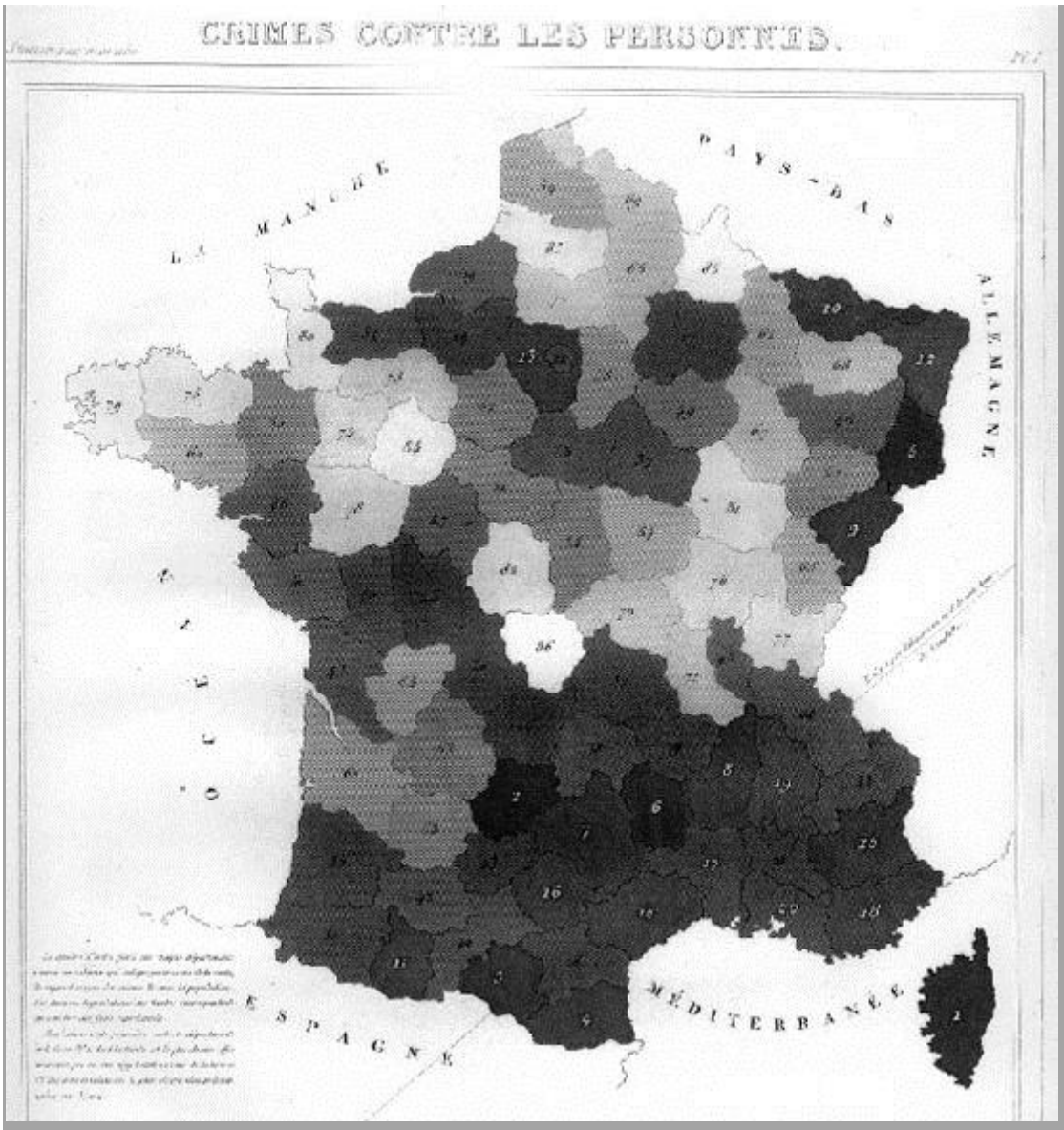


Imagem 13. Mapa coroplético da França indicando os níveis de criminalidade (André-Michel Guerry)

Existem muitos recursos que um bom jornalista precisa compreender para construir visualizações. Ter excelentes conhecimentos de base sobre mapas e gráficos é importante. Tudo o que você cria precisa ser originado de uma série de mapas e gráficos atômicos. Se você pode dominar o básico, então pode melhorar a construção de visualizações mais complexas, feitas a partir destas unidades iniciais.

Dois dos mais básicos tipos de gráfico são os de barra e o de linha. Ao mesmo tempo em que têm usos muito similares, diferenciam-se imensamente por seus significados. Tomemos, por exemplo, as vendas mensais de uma empresa por um ano. Teríamos 12 barras representando a quantidade de dinheiro recebida cada mês.

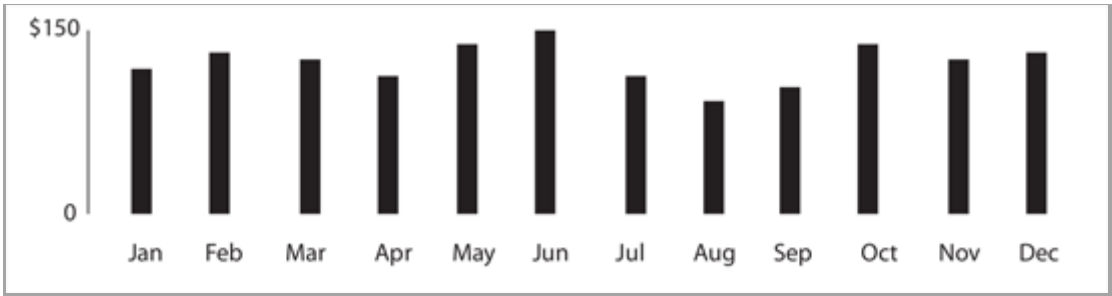


Imagem 14. Um gráfico de barras simples: útil para representar informações descontínuas

Vamos ver porque deveriam ser barras, em vez de um gráfico de linha. Gráficos de linha são ideais para dados contínuos. Com nossos números de vendas, tem-se o somatório do mês, não contínuo. Como uma barra, sabemos que, em janeiro, a empresa gerou US$ 100, e, em fevereiro, US$ 120. Se tratássemos essas informações como um gráfico de linha, ele continuaria a representar US$ 100 e US$ 120 no começo de cada mês, mas o gráfico de linha faria com que parecesse que a empresa gerou apenas US$ 110 no dia 15. O que não é verdade. Barras são usadas para unidades descontínuas de medida, enquanto linhas são usadas quando se tem um valor contínuo, tal como a temperatura.

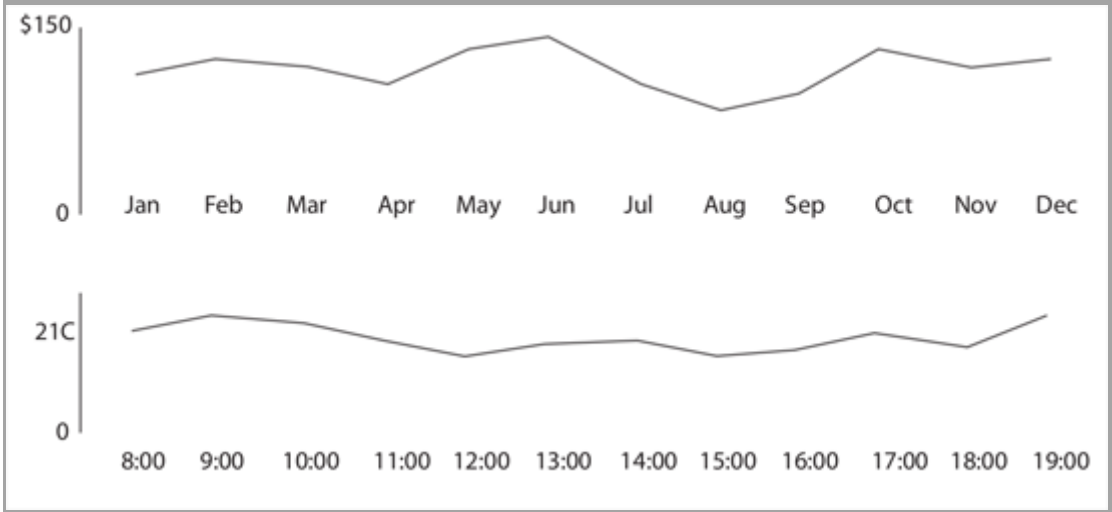


Imagem 15. Gráficos de linha simples: úteis para representar informações contínuas

Podemos ver no gráfico acima que, às 8h, a temperatura era de 20°C e, às 9h, de 22°C. Se olharmos para a linha para adivinhar a temperatura às 8h30, diríamos que era de 21°C, o que é uma estimativa correta, uma vez que a temperatura é contínua e cada ponto não é a soma de outros valores. Ela representa o valor exato naquele momento ou uma estimativa entre duas medições exatas.

Tanto os gráficos de barra quanto a linha possuem uma variação onde se empilham as variáveis. Essa é uma excelente ferramenta para contar histórias e pode funcionar de diferentes formas. Tomemos, por exemplo, uma empresa com 3 locais.

Para cada mês, temos 3 barras, uma para cada uma das lojas — um total de 36 para o ano. Quando as colocamos próximas umas às outras, podemos ver rapidamente qual loja estava faturando mais. Essa é uma história interessante e válida, mas existe outra escondida dentro dos mesmos dados. Se empilharmos as barras, para que tenhamos apenas a cada mês, perdemos a habilidade de ver qual loja é a mais lucrativa, mas veremos em quais meses a empresa faz o melhor negócio como um todo.

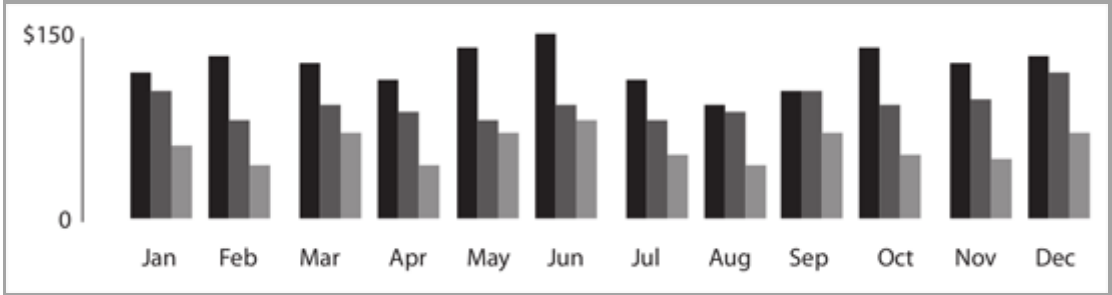


Imagem 16. Gráfico de barras agrupadas mostra diferença de vendas entre lojas

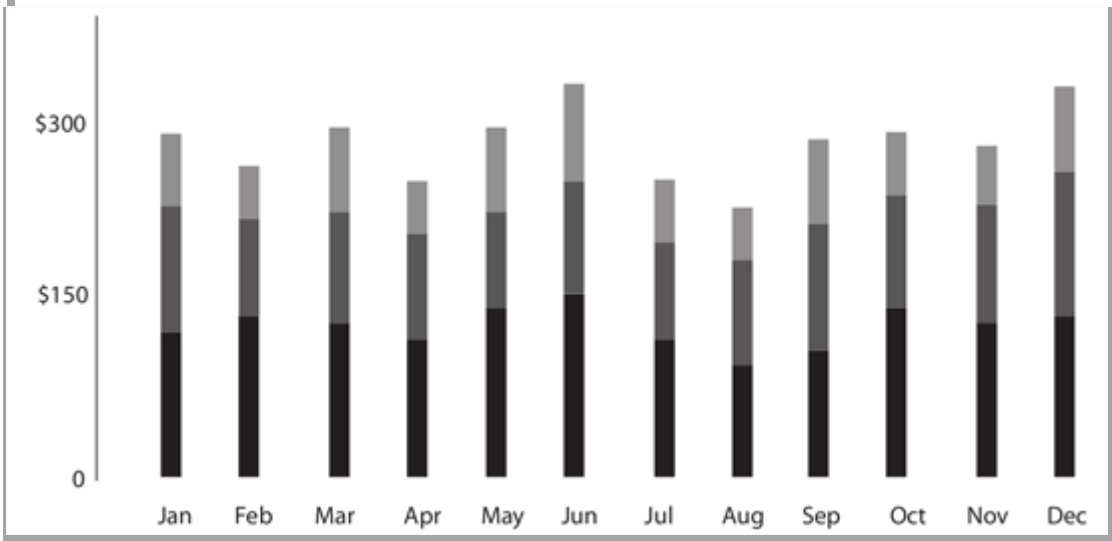


Imagem 17. Gráfico de barras empilhadas mostra melhor o faturamento total

Ambos são exibições válidas da mesma informação, mas contam duas histórias diferentes. Como jornalista, o aspecto mais importante é que você escolha a história que deseja contar. É sobre o melhor mês para os negócios, ou sobre qual loja é a principal? Este é um simples exemplo, mas mostra, na verdade, todo o foco do jornalismo de dados — fazer a pergunta correta antes de ir longe demais. A história vai guiar a escolha de visualização.

O gráfico de barras e o gráfico de linhas são realmente o pão com manteiga de qualquer jornalista de dados. A partir daí, você pode trabalhar com histogramas, gráficos de horizonte, "sparklines", gráficos de fluxo, entre outros. Todos eles

têm características comuns e são apropriados para situações um pouco diferentes entre si — que mudam de acordo com a quantidade de dados, fontes de dados ou a localização do gráfico no que se refere ao texto.

No jornalismo, um dos recursos gráficos mais usados é o mapa. Tempo, quantidade e geografia são itens comuns aos mapas. Sempre queremos saber quanto existe em uma área em detrimento de outra ou como os dados fluem de uma área para outra. Diagramas de fluxo e mapas coropléticos são ferramentas muito úteis e merecem ser incluídas em seu arsenal. Saber codificar corretamente um mapa por meio de cores, sem causar distorções, é fundamental. Mapas políticos são geralmente codificados com cores sólidas que preenchem por completo certas regiões ou as deixam vazias, mesmo se um candidato ganhou somente uma parte do país com uma diferença de 1%. Colorir não precisa ser uma escolha binária: tonalidades podem ser usadas com cautela. Entender mapas é uma grande parte do jornalismo. Mapas facilmente respondem à pergunta ONDE das 5 questões básicas do jornalismo ("quem?", "o que?", "quando?", "onde?" e "por que?")

Uma vez dominado o tipo básico de mapas e gráficos, você pode começar a construir visualizações de dados mais elaboradas. Se não entender estará construindo uma fundação instável. Da mesma forma que você aprende a ser um bom escritor — mantendo as frases curtas, tendo a audiência em mente e não complicando demais as coisas para parecer mais esperto, mas, sim, transmitindo significado ao leitor--, você não deveria pesar demais a mão com os dados. Começar pequeno é a forma mais efetiva de contar uma história, construindo lentamente somente quando necessário.

Uma escrita vigorosa é concisa. Uma sentença não deve conter palavras desnecessárias, um parágrafo não deve apresentar frases desnecessária, pela mesma razão que um desenho não deve ter linhas desnecessárias e uma máquina não deve ter partes desnecessárias. Isso requer não que o escritor faça sentenças curtas ou que trate todos os assuntos de forma superficial, mas sim que cada palavra faça um relato.

Elementos de Estilo (1918)— William Strunk Jr.

É aceitável não usar todos os dados encontrados na sua história. Você não deveria pedir permissão para ser conciso — essa deveria ser a regra.

— Brian Suda, (optional.is)

O faça-você-mesmo da visualização de dados: nossas ferramentas favoritas

Que ferramentas de visualização de dados estão disponíveis na web, são grátis e fáceis de usar? Aqui no Datablog e Datastore, tentamos ao máximo possível usar as opções sem custo, mas poderosas, que estão na internet.

Pode soar falso, pois obviamente temos acesso às maravilhosas equipes de gráficos e interatividade do Guardian para aqueles projetos que exigem mais tempo — como essemapa de gastos públicos (criado com Adobe Illustrator) ou essa ferramenta interativa sobre rebeliões no Twitter.

Para o trabalho cotidiano, no entanto, usamos ferramentas acessíveis para todos — e criamos gráficos que qualquer um também pode criar.

Afinal, o que usamos?

Google Fusion Tables

Essa ferramenta de bases de dados e mapeamento online tornou-se padrão para produzirmos mapas rápidos e detalhados, especialmente aqueles em que você precisa dar zoom. Oferece a alta resolução do Google Maps e aguenta um grande volume de dados — 100 MB no formato CSV, por exemplo. O Fusion parece meio complicado na primeira vez que você usa — mas insista. Nós o utilizamos para fazer mapas como esse do Iraque abaixo e também para produzir mapas com fronteiras delineadas como o sobre a falta de moradia.

Imagem 18. Os diários de guerra do WikiLeaks (Guardian)

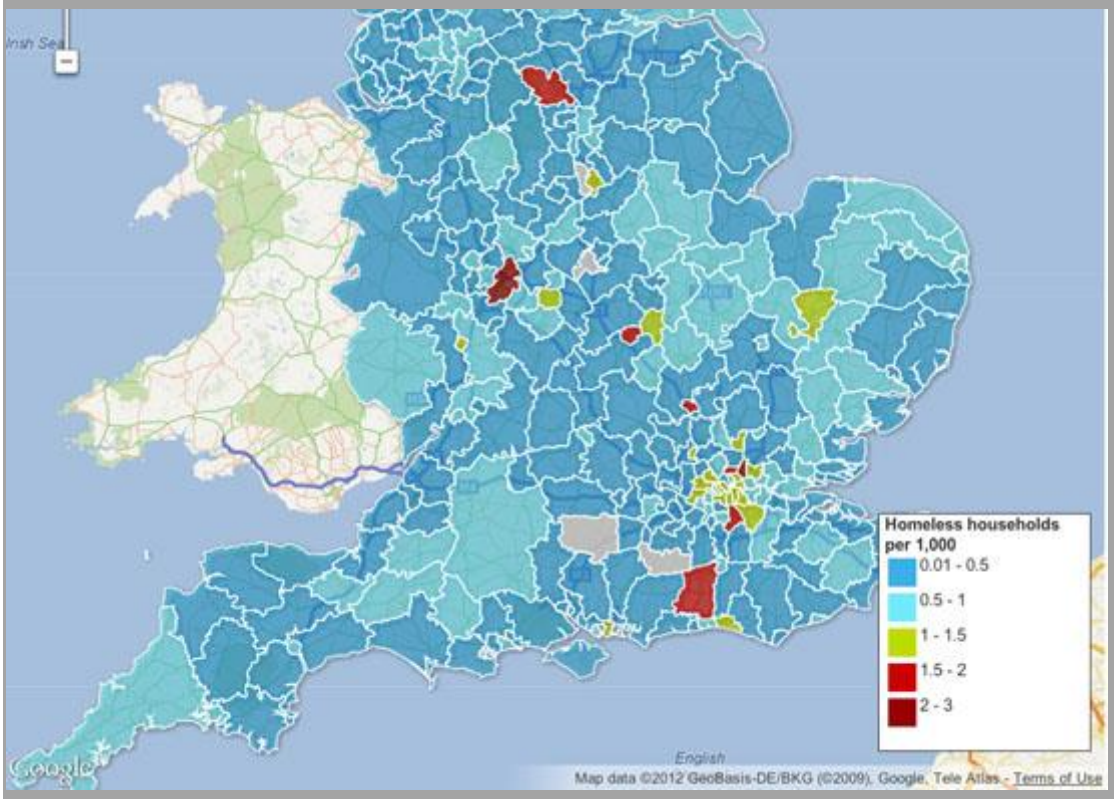


Imagem 19. Mapa interativo dos sem-teto (Guardian)

A maior vantagem é a flexibilidade — você pode fazer o upload de um arquivo KML com as divisas regionais, digamos — e então combiná-lo com uma planilha. Além disso, sua interface está sendo remodelada, o que deve torná-lo mais fácil de usar.

Você não precisa ser um programador para criar um mapa — e essa ferramenta de camadas do Fusion permite que combine diferentes mapas ou crie opções de busca e filtro, e o resultado pode ser incorporado depois a um blog ou site.

Esse excelente tutorial feito por Kathryn Hurley, do Google, é um bom ponto de partida.

Dica

Use o Shape Escape para converter arquivos .shp em tabelas Fusion. Além disso, tome cuidado com mapas muito complicados — a ferramenta não suporta mais de um milhão de pontos em cada célula.

Tableau Public

Se você não precisa do espaço ilimitado da versão profissional, Tableau Public é de graça. Com ele, é possível fazer visualizações complexas de até 100.000 colunas de forma simples e fácil. Nós usamos quando é necessário apresentar diferentes tipos de gráficos ao mesmo tempo, como nesse caso do mapa das alíquotas de imposto mais altas do mundo (que também inclui um gráfico de barras).

Ou pode ser usado para explorar os dados. Foi o que fizemos com os dados de gastos nas eleições presidenciais dos Estados Unidos (se bem que ultrapassamos a cota de espaço da versão gratuita... algo com o que se deve tomar cuidado). Também é preciso que os dados estejam formatados de maneira muito específica para se tirar o máximo do Tableau. Mas, uma vez superada esta etapa, você terá uma ferramenta intuitiva e que funciona bem. O La Nación, da Argentina, montou toda sua área de jornalismo de dados em torno ao Tableau, por exemplo.

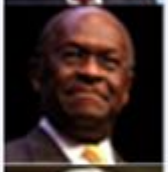
2012 Presidential Campaign Finance

Republican Candidates
(click name to view by state)

Mitt Romney $50.37M

Rick Perry $18.53M

Ron Paul $11.84M



Herman Cain $7.01M

Newt Gingrich $6.31M

Jon Huntsman $2.96M

Michelle Bachmann $2.61M

Rick Santorum $1.41M

Incumbent President
(click name to view by state)

Barack Obama $72.45M

Imagem 20. Financiamento da campanha presidencial de 2012 (Guardian)

Para quem quiser começar a usar o Tableau, há bons tutoriais emhttp://www.tableausoftware.com/learn/training.

Dica

O Tableau é feito para PCs, mas está sendo elaborada uma versão para Mac. Enquanto isso, use um mirror como o parallels para fazê-lo funcionar.

Gráficos do Google Spreadsheet

Você pode acessar essa ferramenta em http://www.google.com/google-d-s/spreadsheets/.

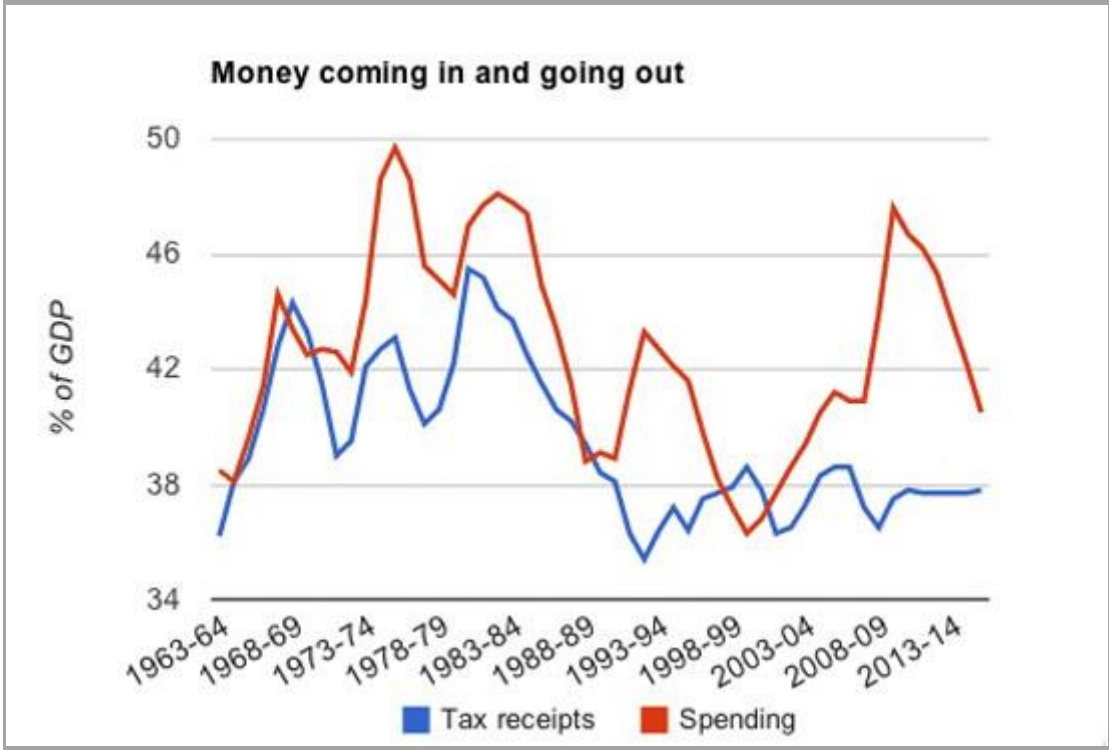


Imagem 21. Gastos públicos e impostos no Reino Unido (Guardian)

Procurando por algo simples, como um gráfico de barras, linhas ou pizza? As planilhas do Google (que podem ser criadas a partir do menu Documentos em sua conta) também podem se tornar gráficos bem legais — incluindo as bolhas animadas usadas por Hans Rosling no Gapminder. Ao contrário dos gráficos API, não é necessário saber códigos de programação; é bem similar à criação de um gráfico no Excel, pois basta selecionar os dados e clicar na janela de gráficos. Também vale a pena explorar as opções de personalização; você pode mudar cores, títulos e proporções. Eles são

bastante neutros no que se refere ao design, o que é bem útil no caso de gráficos pequenos. Há ainda opções interessantes para os gráficos de linha, incluindo anotações.

Dica

Explore as opções de personalização de gráficos; você pode criar sua própria paleta de cores.

Datamarket

Mais conhecido como fornecedor de dados, o Datamarket é também uma ótima ferramenta de visualização de números. Você pode fazer o upload dos seus ou usar uma das várias bases de dados que eles oferecem, e as opções ficam melhores com a conta Pro.

Dica

O Datamarket funciona melhor com séries históricas de dados, mas confira a extensa gama de dados que eles oferecem.

Many Eyes

Um site que precisa ser tratado com um pouco mais de carinho é o Many Eyes, da IBM. Quando foi lançado, criação de Fernanda B. Viégas e Martin Wattenberg, era uma ferramenta singular ao permitir o upload e a visualização de bases de dados. Agora que seus criadores trabalham para o Google, o site parece meio abandonado com suas paletas de cores sem graça; não apresenta nada novo em termos de visualização há algum tempo.

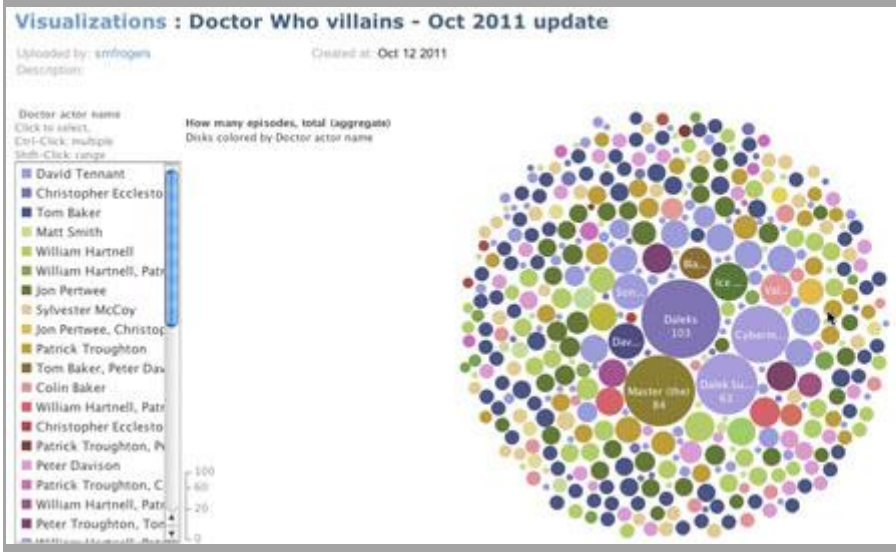


Imagem 22. Vilões do Doctor Who; Guardian

Dica

Você não pode mais editar os dados depois de fazer o upload, então tenha certeza de que estão corretos antes de enviar.

Color Brewer

Não é exatamente uma ferramenta de visualização. O Color Brewer serve para escolher cores de mapas. Você escolhe a cor básica e ele sugere os códigos para o resto da paleta.

E mais alguns

Se nenhuma dessas dicas é o que procurava, vale a pena conferir essa lista do DailyTekk, que tem ainda mais opções. As ferramentas acima não são as únicas, mas apenas aquelas que usamos com mais frequência. Há muitas outras opções, incluindo:

- Chartsbin, uma ferramenta para criar mapas-múndi interativos
- iCharts, que é especializada em widgets de gráficos simples
- GeoCommons, que compartilha dados geográficos para criar mapas locais e globais
- Ah, tem também o piktochart.com, que oferece templates para as visualizações mais populares do momento.

— Simon Rogers, the Guardian

Como mostramos os dados no Verdens Gang

Jornalismo é levar novas informações ao leitor o mais rápido possível. A forma mais rápida pode ser um vídeo, uma fotografia, um texto, um gráfico, uma tabela ou uma combinação de tudo isso. A respeito de visualizações, o objetivo deve ser o mesmo: informação rápida. Novas ferramentas de dados permitem aos jornalistas encontrar histórias com as quais eles não teriam contato de outra forma, assim como apresentá-las de novas maneiras. Aqui estão alguns exemplos de como nós apresentamos dados no jornal mais lido na Noruega, o Verdens Gang (VG).

Números

Esta história é baseada em dados do Instituto de Estatísticas Norueguês, dados de contribuintes e dados do monopólio nacional de loterias. No gráfico interativo abaixo, o leitor podia encontrar diferentes tipos de informação de cada municipalidade ou condado norueguês. A tabela mostra a porcentagem da renda gasta em jogos e foi construída usando-se o Access, Excel, MySql e Flash.

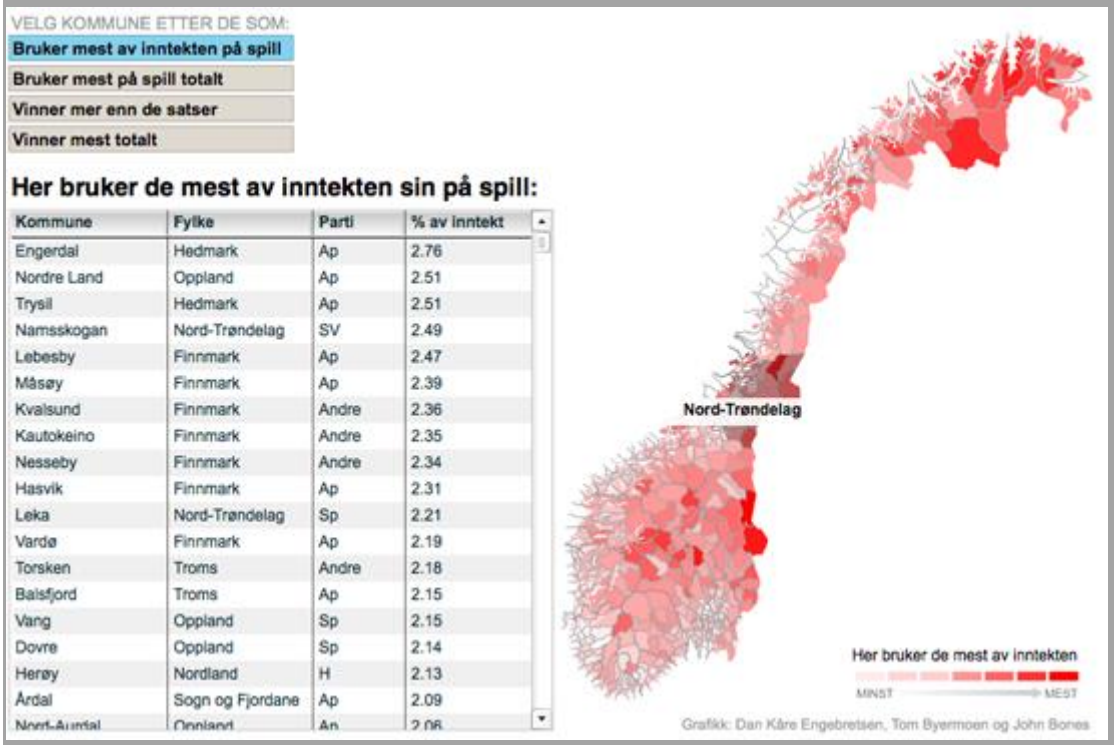


Imagem 23. Mapeando dados dos contribuintes e da Lotto (Verdens Gang)

Redes

Nós utilizamos análises de redes sociais para estudar as relações entre os 157 filhos e filhas das pessoas mais ricas da Noruega. Nossa investigação mostrou

que os herdeiros dos mais ricos da Noruega também herdaram as redes sociais dos seus pais. Ao todo, foram mais de 26.000 conexões, e os gráficos foram todos finalizados manualmente com o Photoshop. Usamos Access, Excel, Bloco de Notas e a ferramenta de análise de redes sociais Ucinet.

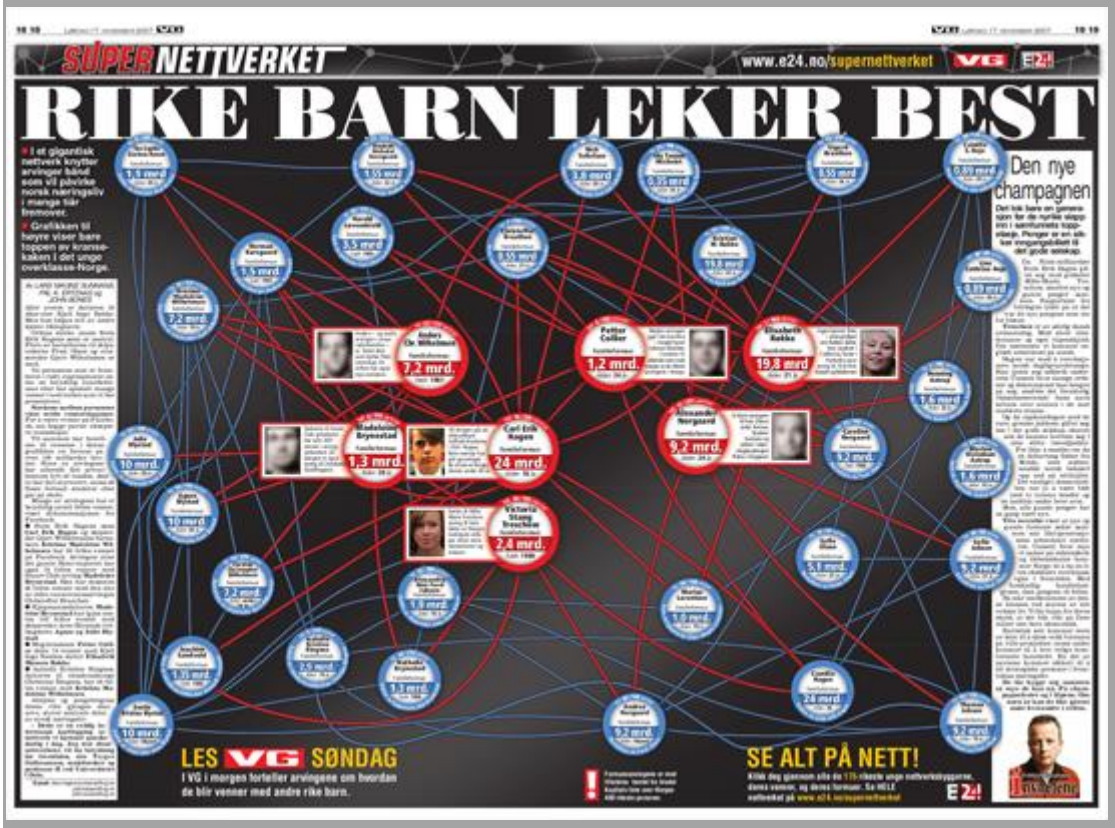


Imagem 24. Aves ricas de mesma plumagem voam juntas (Verdens Gang)

Mapas

Neste mapa de calor animado combinado com um gráfico simples de barras, você pode ver a incidência de crimes no centro de Oslo, hora a hora, no fim de semana, por vários meses. No mesmo mapa, é possível conferir o número de oficiais da polícia trabalhando ao mesmo tempo. Quando o crime está realmente acontecendo, a quantidade de policiais está no nível mais baixo. O mapa foi feito usando ArcView com Spatial Analyst.

Imagem 25. Mapa de calor animado (Verdens Gang)

Mineração de texto

Para esta visualização, fizemos mineração de dados (extração de padrões ocultos em bases de dados) nos discursos feitos por sete líderes de partidos noruegueses durante suas convenções partidárias. Todos os discursos foram analisados, e esses estudos forneceram ângulos para algumas reportagens. Cada reportagem foi relacionada a um gráfico e os leitores puderam explorar e conhecer melhor a linguagem dos políticos. Essa visualização foi feita usando Excel, Access, Flash e Illustrator. Se tivesse sido feito em 2012, teríamos feito o gráfico interativo em JavaScript.

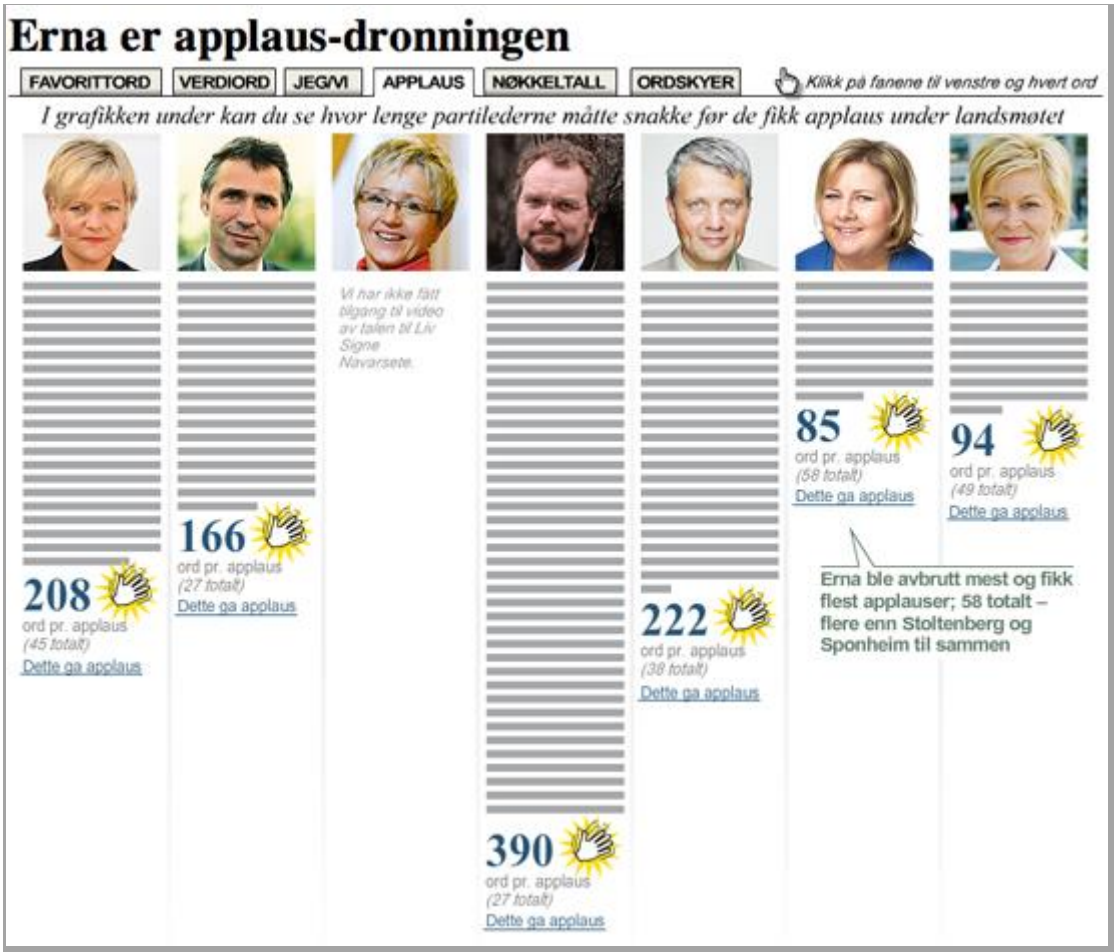


Imagem 26. Mineração de texto dos discursos de líderes partidários (Verdens Gang)

Conclusão

Quando precisamos visualizar uma reportagem? Na maioria das vezes não precisamos, mas há momentos em que queremos fazer isso para ajudar nossos leitores. Reportagens que contêm uma grande quantidade de dados geralmente precisam de visualização. No entanto, temos de ser críticos ao escolher que tipo de dados vamos apresentar. Conhecemos todos os detalhes quando informamos sobre algo, mas o que o leitor realmente precisa saber na reportagem? Talvez uma tabela seja suficiente, ou um gráfico simples mostrando uma evolução do ano A para o ano C. Ao trabalhar com jornalismo de dados, a questão não é necessariamente apresentar grandes quantidades de dados. É sobre jornalismo!

> Tem havido uma tendência clara nos últimos três anos para criar gráficos interativos e tabelas que permitem ao leitor se aprofundar em temas diferentes. Uma boa visualização é como uma boa fotografia. Você entende do que se trata só de olhar para ela por um momento ou dois. Quanto mais você olhar para a visualização, mais você a vê. A visualização é ruim quando o leitor não sabe por onde começar ou terminar, e quando a visualização está sobrecarregada de detalhes. Neste cenário, talvez um texto seja melhor, não?
>
> *— John Bones, Verdens Gang*

Dados públicos viram sociais

Os dados têm valor inestimável. O acesso a eles tem o potencial de jogar luz sobre diversos assuntos de uma forma que impulsiona resultados. No entanto, um mau tratamento dos dados pode colocar os fatos em uma estrutura que não comunica nada. Se não promover discussão ou proporcionar um entendimento contextualizado, os dados podem ter um valor limitado para o público.

A Nigéria voltou para a democracia em 1999, depois de longos anos de ditadura militar. Sondar os fatos por trás dos dados era uma afronta à autoridade e visto como uma tentativa de questionar a reputação da junta. A Lei de Segredos Oficiais levou os funcionários públicos a não compartilhar informações do governo. Mesmo 13 anos depois da volta da democracia, acessar dados públicos pode ser uma tarefa difícil. Quando se trata de informações sobre gastos públicos, por exemplo, é difícil passá-las de uma maneira clara para a maioria da audiência, que não conhece bem contabilidade financeira.

Com o aumento do número de celulares e de nigerianos online, vimos uma imensa oportunidade de usar tecnologias de visualização de dados para explicar e engajar as pessoas em torno às despesas públicas. Para isso, tínhamos que envolver os usuários em todas as plataformas, assim como chegar aos cidadãos por meio de ONGs. Lançamos o projeto BudgIT, que visa fazer dos dados públicos um objeto social, e construir um extensa rede que demande mudanças.

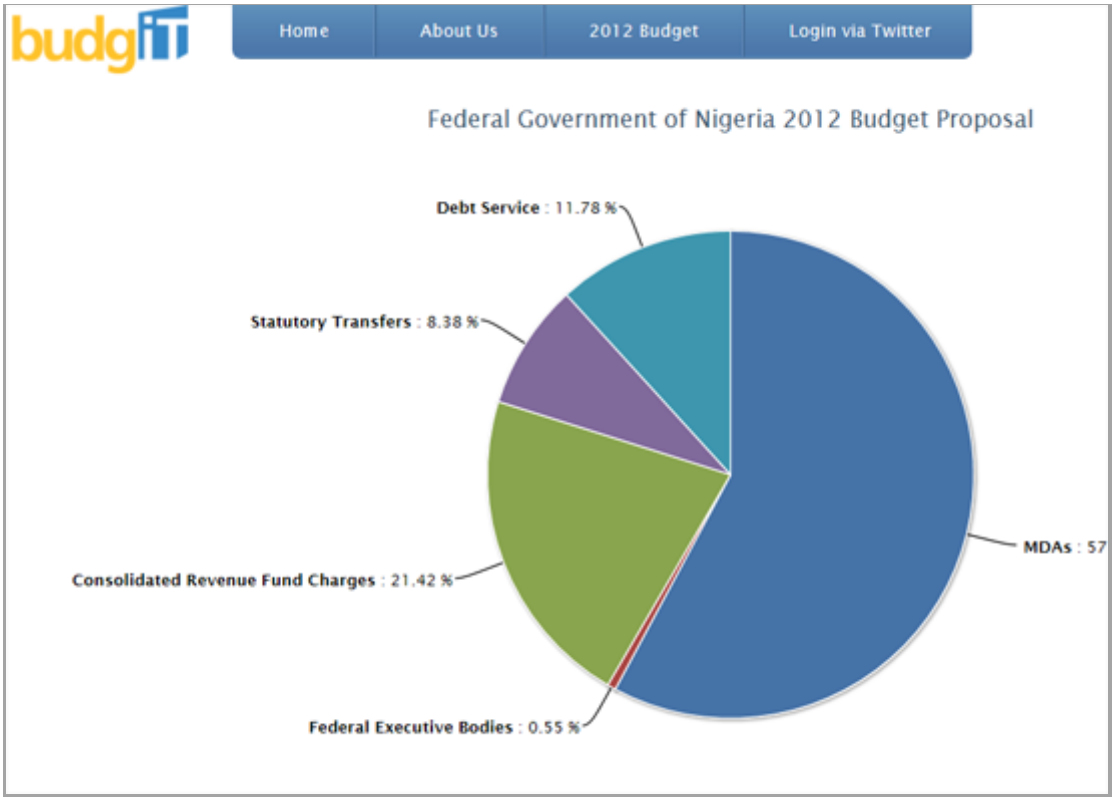


Imagem 27. Aplicativo The BudgIT (BudgIT Nigeria)

Para conseguir engajar os usuários, temos que entender o que eles querem. Com o que o cidadão nigeriano se preocupa? Onde eles veem uma lacuna de informação? Como podemos tornar os dados relevantes para suas vidas? O alvo imediato do BudgIT é o nigeriano de educação média, conectado a fóruns online e mídias sociais. Para competir pela atenção de usuários, temos que apresentar os dados de maneira breve e concisa. Após transmitir uma prévia dos dados na forma de um tweet ou infográfico, há oportunidade para um envolvimento sustentado. Isso pode ser feito por meio de uma experiência mais interativa, a fim de entregar aos usuários um contexto mais amplo.

Na visualização de dados, é importante entender o nível de compreensão que os usuários têm desse tipo de informação. Por mais bonitos e sofisticados que sejam, vimos que diagramas complexos e aplicativos interativos podem não ser ideais para uma comunicação efetiva com os nossos leitores. Uma boa visualização vai falar com o usuário por meio de um uma linguagem que ele entenda, assim como contar uma história com a qual ele sinta uma conexão imediata.

Conseguimos engajar mais de 10 mil nigerianos na questão do orçamento, e os dividimos em três categorias para dar a eles informações de maior valor. As categorias estão explicadas resumidamente abaixo:

Usuários ocasionais

Querem informação de forma simples e rápida. Estão interessados em ter uma ideia geral dos dados, não em análises detalhadas. Podemos atrai-los com tweets ou gráficos interativos.

Usuários ativos

Estimulam a discussão e usam os dados para melhorar seus conhecimentos sobre determinada área ou desafiar as suposições ligadas a eles. Para esses usuários, queremos oferecer mecanismos de feedback e a possibilidade de que compartilhem ideias com seus pares pelas redes sociais.

Consumidores massivos de dados

Querem dados brutos para visualização ou análise. Nós simplesmente damos a eles as informações que desejam.

Com o BudgIT, o engajamento do nosso usuário é baseado em:

Estimular discussões sobre tendências atuais

O BudgIT acompanha discussões online e offline e procura fornecer dados sobre os assuntos atuais. Por exemplo, durante as greves do setor de combustíveis de janeiro de 2012, houve agitação constante entre os manifestantes com relação à necessidade de reinstituir os subsídios ao combustível e reduzir gastos públicos exagerados e desnecessários. O BudgIT acompanhou o debate pelas mídias sociais e, em 36 trabalhosas horas, construiu um aplicativo que permite aos cidadãos reorganizar o orçamento nigeriano.

Bons mecanismos de feedback

Tentamos engajar os usuários por meio de canais de discussão e das redes sociais. Muitos querem conhecer as histórias ligadas aos dados, enquanto outros perguntam nossa opinião. Garantimos que nossas respostas expliquem apenas os fatos por trás dos dados, sem vínculos com visões pessoais ou políticas. Precisamos manter abertos os canais de feedback, responder ativamente a comentários e envolver a audiência criativamente para garantir que a comunidade construída ao redor dos dados se mantenha.

Tornar local

Para uma base de dados voltada a um grupo específico de usuários, o BudgIT tenta localizar ou adaptar seu conteúdo e promover um canal de discussão que se conecte às suas necessidades. Em particular, estamos interessados em engajar o público por meio de mensagens SMS.

Depois de publicar dados de gastos no site yourbudgit.com, chegamos aos cidadãos com a ajuda de várias ONGs. Também planejamos desenvolver uma rede participativa em que os cidadãos e instituições governamentais se encontrem em prefeituras para definir itens fundamentais do orçamento a serem priorizados.

O projeto teve cobertura de mídia local e estrangeira, da CP-Africa à BBC. Fizemos uma análise dos orçamentos de 2002-2011 para o setor de segurança para uma jornalista da AP, Yinka Ibukun. A maioria das organizações de mídia é composta por "usuários pesados de dados" e nos pede informações para usar em reportagens. Estamos planejando mais colaborações com jornalistas e organizações de notícias ao longo dos próximos meses.

— Oluseun Onigbinde, BudgIT Nigeria

Engajando pessoas nos seus dados

Tão importante quanto publicar dados é obter uma reação da audiência. Você é humano; vai cometer erros, perder coisas e ter ideias erradas de tempos em tempos. A sua audiência é um dos bens mais úteis que você tem. Ela pode verificar fatos e apontar outras coisas que não foram consideradas.

Engajar o público, no entanto, é complicado. Você está lidando com um grupo de pessoas condicionadas por anos de uso da internet, de navegação de site em site, e que deixam apenas um comentário sarcástico ao longo de suas caminhadas. Construir uma relação de confiança com seus usuários é crucial; eles precisam saber o que vão obter, como reagir e dar feedback ao que será ouvido.

Mas primeiro é preciso pensar no público que você tem, ou que deseja ter. O público que vai ser informado e informar por meio do tipo de dados com os quais você trabalha. Se a audiência está ligada a um setor particular, será necessário explorar formas de comunicação personalizadas. Existem organizações que você pode contatar para que ajudem na divulgação do material a um público mais amplo? Existem sites comunitários ou fóruns com os quais conversar? Há publicações comerciais especializadas que gostariam de ajudar na confecção de sua reportagem?

As redes sociais também são uma ferramenta importante. No entanto, mais uma vez, dependem do tipo de dados sobre a mesa. Se estiver trabalhando com estatísticas globais de transportes, por exemplo, vai ser complicado encontrar um grupo no Facebook ou no Twitter especialmente interessado nas suas atividades. Por outro lado, se estiver peneirando índices mundiais de corrupção ou de crimes locais, será mais fácil achar pessoas preocupadas com esses assuntos.

Quando se trata do Twitter, a melhor abordagem é entrar em contato com perfis de personalidades públicas, explicando brevemente a importância de seu trabalho e incluindo um link. Com sorte, eles retuitarão a mensagem aos seus leitores. Esta é uma ótima forma de aumentar a exposição do seu trabalho com um esforço mínimo — e sem atormentar as pessoas!

Depois de obter leitores para a sua página, pense em como eles vão interagir com seu trabalho. Claro, podem ler a história que você escreveu e ver mapas e infográficos. Mas é imensamente valioso oferecer também canais de resposta.

Mais que tudo, eles podem contribuir com ideias sobre o tema tratado, ajudando a definir as próximas tarefas do projeto de cobertura.

Primeiro, não precisa nem dizer que o ideal é publicar os dados brutos em suas reportagens. Você pode apresentar os dados em uma planilha CSV ou hospedá-los em outros serviços, como o Google Docs. Assim, você terá apenas uma versão dos dados e poderá atualizá-la a qualquer momento, por exemplo para corrigir possíveis erros. Se puder, a melhor alternativa é fazer as duas coisas. Permita que as pessoas acessem as informações brutas da sua reportagem da forma mais fácil possível.

Então, pense em outras formas de interagir com o público. Acompanhe as métricas que revelam quais partes de suas bases de dados estão conseguindo mais atenção — é provável que as áreas de maior tráfego digam algo sobre detalhes que você tenha perdido. Por exemplo, você pode não ter dado destaque para as estatísticas de pobreza da Islândia, mas se esses blocos recebem muitas visitas, é porque pode valer a pena estudá-los melhor.

Pense além da caixa de comentários. Você pode anexar comentários a células particulares de uma planilha? Ou a uma região específica de um infográfico? Enquanto a maioria dos sistemas de edição não permitem esse tipo de incorporação de informações, vale a pena avaliar essa possibilidade se estiver criando um material mais elaborado. Os benefícios que esse recurso pode trazer aos seus dados não podem ser subestimados.

Certifique-se de que os demais usuários também vejam esses comentários — em muitos casos, eles têm quase tanta importância quanto os dados originais, e se você mantiver essa informação somente para si, vai privar o público desse valor.

Finalmente, outras pessoas podem querer publicar seus próprios infográficos e histórias baseados nas mesmas fontes de dados. Por isso, pense em qual é a melhor forma de vinculá-los e alinhar o trabalho deles. Você também pode usar uma hashtag específica para o conjunto de dados. Ou, se ele for muito pictórico, compartilhe em um grupo do Flickr.

Também pode ser útil contar com uma via confidencial de compartilhamento de informações. Em alguns casos, algumas pessoas podem não se sentir seguras de fazer suas contribuições publicamente, ou mesmo não se sentir confortáveis nesse contexto. Elas podem preferir submeter informações por meio de um endereço de e-mail, ou até mesmo usar uma caixa de comentários anônimos.

A coisa mais importante que você pode fazer com seus dados é divulgá-los da forma mais ampla e aberta possível. Permitir que os leitores verifiquem seu trabalho, encontrem erros e apontem detalhes perdidos que tornarão melhores tanto o seu jornalismo como a experiência do público.

— Duncan Geere, Wired.co.uk