

Na era do Big Data, o jornalismo de dados tem profunda importância para a sociedade

Traduzido e adaptado por Natália Mazotte do artigo “In the age of big data, data journalism has profound importance for society”, de Alex Howard

A promessa do jornalismo de dados foi um tema forte durante toda a edição de 2012 da conferência do National Institute for Computer-Assisted Reporting (ou Instituto Nacional de Reportagem Assistida por Computador, o NICAR como é conhecido por sua sigla em inglês). Em 2012, dar sentido ao *big data* através de narrativa e contexto, particularmente de dados não estruturados, será um objetivo central para os cientistas de dados ao redor do mundo, quer eles trabalhem em redações, em Wall Street ou no Vale do Silício. Esse objetivo será possibilitado substancialmente por um conjunto crescente de ferramentas gratuitas, sejam elas utilizadas por técnicos do governo como feito em Chicago, técnicos de saúde ou programadores em redações.

No NICAR 2012, você podia ver literalmente o código que sustenta o futuro do jornalismo sendo escrito - ou, pelo menos, projetado - nas paredes.

"O nível de energia era incrível", disse David Herzog, professor associado de jornalismo impresso e digital da Escola de Jornalismo do Missouri, em uma entrevista por email após o NICAR. "Eu não vi participantes com as mãos suadas e preocupados com o futuro do jornalismo. Eles estavam muito ocupados o construindo."

Assim como softwares cívicos abertos fazem cada vez mais parte do governo, softwares de código aberto **estão desempenhando um papel fundamental no novo jornalismo de dados.**

"Ferramentas gratuitas e de código aberto dominaram a conversa", disse Herzog. "É evidente a partir dos painéis e das aulas práticas que ferramentas de código livre e aberto eliminaram a barreira de entrada em termos de muitos custos de software."

Embora muitos programadores sejam agnósticos com relação a quais ferramentas eles usam para realizar um trabalho, as pessoas que estão construindo e compartilhando ferramentas de jornalismo de dados estão fazendo isso muitas vezes por meio de códigos de fonte aberta. Como

Dan Sinker, o chefe da Parceria de Tecnologia Jornalística Knight-Mozilla para o Mozilla, escreveu depois, jornalistas-programadores levaram o NICAR 12 "a um nível totalmente novo".

- O Projeto PANDA foi lançado oficialmente em versão beta em St. Louis, com uma festa de lançamento. Você pode conferir o PANDA no Github.
- O Overview da organização Associated Press tornará a procura através de documentos mais fácil.
- O LA Times Datadesk compartilhou uma ferramenta, Django Bakery, para transformar aplicativos Django em arquivos HTML.
- Jonathan Soma, com a ajuda de John Keefe da WNYC, desenvolveu Tabletop.js como uma maneira de usar uma planilha pública do Google Spreadsheets como uma fonte de dados para um aplicativo Web. Tabletop está no Github também.

Enquanto uma parte do desenvolvimento de código aberto foi definitivamente impulsionada pelas exigências do Knight News Challenge, que financiou o PANDA e o projeto Overview, há também um espírito de colaboração em evidência em toda esta comunidade.

Este é um grupo de pessoas altamente dedicadas "a mostrar o seu trabalho" - e, para os programadores dentro de redações, isso significa compartilhar seu código. Para colocar de outra forma, programe, não diga. Sessões sobre Python, Django, mapeamento, Google Refine e Google Fusion estavam lotadas no NICAR 12.

"Eu acho que esta foi a melhor conferência NICAR desde a primeira, em 1993", disse Herzog. "Sempre foi difícil fazer escolhas a partir do cardápio de painéis, demonstrações e aulas práticas em conferências NICAR. Mas eu achava que havia uma abundância de sessões excelentes e muito informativas desenvolvidas pelos próprios participantes. Além disso, acho que o NICAR oferece uma boa variedade de opções para iniciantes e especialistas. Por exemplo, os participantes podem aprender a mapear usando tabelas do Google Fusion na parte para novatos, ou ainda usando PostGIS e qGIS no nível avançado. Coletar dados através da raspagem tornou-se um negócio cada vez maior para os jornalistas de dados. Ao mesmo tempo, está ficando mais fácil para as pessoas com pouca ou nenhuma formação em programação raspar usando ferramentas como planilhas, Google Refine e ScraperWiki. "

Sobre a história do NICAR

De acordo com o IRE, o NICAR foi fundado em 1989. Desde então, o instituto já treinou milhares de jornalistas em como achar, coletar e publicar informações eletrônicas.

Atualmente, "as conferências NICAR ajudam jornalistas, hackers e programadores a entender as melhores práticas, métodos e ferramentas digitais para fazer um jornalismo que envolva análise

de dados e reportagem clássica de campo”, disse Brant Houston, ex-diretor executivo da Investigative Reporters and Editors, em uma entrevista por e-mail. “A conferência NICAR obviamente também abrange jornalismo investigativo e padrões para integridade e credibilidade de dados.”

“St. Louis foi minha décima conferência NICAR”, disse Anthony DeBarros, o editor sênior de banco de dados do jornal USA Today, em uma entrevista por e-mail. “Minha primeira foi em 1999, em Boston. A conferência é um lugar onde jornalistas nerds podem se reunir e se lembrar de que não estão sozinhos em seu amor por números, análise de dados, programação e pela busca de grandes histórias ao se debruçarem sobre colunas em uma planilha. Ela serve como um importante espaço de treinamento para jornalistas começando a mexer com dados em redações, e tem sempre mantido jornalistas informados sobre os desenvolvimentos tecnológicos que oferecem novas maneiras de encontrar e contar histórias. Ao mesmo tempo, sua conexão com o IRE a mantém firmemente enraizada nos melhores aspectos do jornalismo investigativo - desenterrando histórias que servem ao bem público.”

Baby, you can drive my CAR

Muito antes de nós começarmos a falar sobre "jornalismo de dados," a prática da reportagem assistida por computador (CAR, na sigla em inglês) foi crescendo ao redor do mundo.

"A prática de CAR mudou ao longo do tempo assim como as ferramentas e o ambiente no mundo digital mudaram", disse Houston. "Ela começou no tempo dos *mainframes* no final dos anos 60, passou para PCs (que aumentaram a velocidade e a flexibilidade de análise e apresentação) e seguiu para a Web, o que acelerou a capacidade de coletar, analisar e apresentar dados. Os objetivos básicos permaneceram os mesmos. Vasculhar dados e dar sentido a eles, muitas vezes com métodos das ciências sociais. A prática CAR tende a ser um termo "guarda-chuva" - um que inclui jornalismo de precisão, jornalismo de dados e qualquer metodologia que dê sentido a dados, tais como visualizações e apresentações de dados eficazes".

De uma certa forma, o termo CAR ainda existe porque o mundo do jornalismo não cunhou uma expressão melhor para usar em seu lugar.

"Reportagem assistida por computador" é uma expressão antiquada, mas a maioria das pessoas que a pratica reconhece isso há anos ", disse DeBarros. "Ela se mantém porque ninguém ainda inventou um substituto *matador*." Phil Meyer, o padrinho do movimento, escreveu um livro seminal chamado Precision Journalism, e essa expressão é ótima para descrever esse segmento de CAR que lida com estatísticas e com o uso de métodos de ciências sociais na busca por notícias. Como um termo guarda-chuva, o jornalismo de dados parece ser a melhor descrição no momento,

provavelmente porque abrange adequadamente a maioria das áreas que CAR tornou-se - desde a tradicional reportagem baseada em dados até as novas categorias de aplicativos de notícia.

A mudança mais significativa na CAR pode muito bem ter ocorrido quando todos os computadores que estavam sendo utilizados para reportagens foram conectados em rede (com a Internet) na década de 1990.

"Pode parecer óbvio, mas é claro que a Internet mudou tudo, e por um tempo CAR foi esmagada enquanto se tentava aprender a navegar na Internet por histórias, e a como fazer o download de dados", disse Cohen. "Em seguida, houve uma fase em que todo mundo estava desenvolvendo intranets para entregar registros públicos dentro de redações, para ajudar a encontrar pessoas no prazo de entrega, etc. Por isso, na maior parte do tempo, a CAR estava focada em comunicação, não na publicação ou apresentação. Agora o pessoal de jornalismo de dados surgiu com outra direção: pessoas que estão usando dados obtidos por meio de APIs e que muitas vezes ignoram o lado de reportagem e usam as mesmas técnicas para entregar a informação não filtrada para os seus leitores em um formato mais fácil do que o que o governo está nos dando. Mas eu acho que tudo está começando a se consolidar - os chamados jornalistas de dados estão cada vez mais interessados em reportagens, e os repórteres que utilizam as práticas de reportagem assistida por computador de forma mais tradicional estão interessados na obtenção de suas histórias na web de maneiras mais interessantes.

Não importa o nome, os objetivos ainda são os mesmos.

"CAR sempre foi relacionada ao uso de dados para encontrar e contar histórias", disse DeBarros. "E ainda é. O que mudou nos últimos anos é a maior ênfase dada a apresentações on-line (mapas e aplicativos interativos) e as habilidades de programação necessárias para produzi-las (JavaScript, HTML / CSS, Django, Ruby on Rails). As conferências NICAR anteriores giravam muito mais em torno das melhores histórias do ano e de como usar técnicas de dados para cobrir determinados tópicos. Isso ainda existe. Mas, mais recentemente, a conferência e a prática cresceram para incluir mais tópicos de programação e apresentação de dados. Isso reflete o estado da mídia - todas as redações estão fazendo hora extra para fazerem o seu conteúdo funcionar bem na web, no celular e com aplicativos, e os jornalistas de dados tendem a ser vanguarda nesse sentido, de maneira que não é surpreendente que a conferência se expanda para incluir esses temas."

O que foi destaque no NICAR 2012?

As ferramentas e táticas em exposição no NICAR foram suficientes para convencer Tyler Dukes da Universidade de Duke a escrever que "o NICAR me mostrou que eu não sei nada". Navegue através das ferramentas, slides e links da NICAR 2012 com curadoria de Chrys Wu para ter uma noção de quanta novidade existe lá fora. O grande tema, no entanto, sem dúvida, eram os dados.

"Os dados realmente são o filé mignon da conferência, e uma verificação rápida da programação mostra as toneladas de sessões realizadas sobre todos os tipos de tópicos de dados, desde o censo populacional à saúde, do crime à educação", disse DeBarros.

O que eu vi em toda parte no NICAR foi o interesse não apenas em dados, no entanto, mas em como obtê-los e utilizá-los, desde encontrar histórias e fontes que forneçam evidências empíricas para sustentar outras reportagens até contar histórias com mapas e visualizações.

"Um dos temas principais foi a análise de dados (usando planilhas, softwares de gerenciamento de dados, GIS), que dá mais credibilidade ao jornalismo ao identificar padrões, tendências e pontos fora da curva", disse Houston. "Outros temas envolveram coleta e análise de mídias sociais, visualização de dados, planejamento e organização de histórias baseadas na análise de dados, programação para web scraping (coleta de dados na Web) e a combinação de vários programas da Web."

"Coleta de dados através de raspagem web vem se tornando um negócio cada vez maior para os jornalistas de dados", disse Herzog. "Ao mesmo tempo, está ficando mais fácil para as pessoas com pouca ou nenhuma formação em programação raspar usando ferramentas como planilhas, Google Refine e ScraperWiki. Dito isto, uma outra mensagem para mim foi como a programação tem se tornado importante. Não, nem todos os jornalistas ou até mesmo jornalistas de dados precisam aprender programação. Mas como Rich Gordon da Medill disse, todos os jornalistas devem compreender o que ela pode fazer. "

Cohen destacou os dados de forma similar, especificamente a sua forma. "O tema que eu vi este ano foi um foco em dados não estruturados, em vez de estruturados", disse ela. "Durante muito tempo, martelamos governos para nos darem 'dados' em colunas e linhas. Acho que estamos vendo cada vez mais que as histórias vêm da informação não estruturada na mesma intensidade (senão em intensidade maior) em que vêm de documentos, áudio e vídeo, tweets e outras mídias sociais - a partir de fontes governamentais e não-governamentais. O outro tema é que há muito mais colaboração, abertura e compartilhamento entre as empresas de notícias concorrentes. (Veja, por exemplo, PANDA, census.ire.org e a [API do New York Times sobre financiamento de campanha](#)). Mas isso também não vai tão longe - você não vê a ProPublica compartilhando com outros os dados médicos de mais de 40 estados que Dan conseguiu através de raspagem (eu tenho que admitir, no entanto, que eu não pedi a ele para compartilhar.) O IRE sempre foi relacionado a técnicas e ferramentas de compartilhamento - agora estamos realmente compartilhando fontes"

Embora os dados tenham dominado o NICAR 12, outras tendências também foram importantes, das ferramentas abertas de mapeamento às tendências macroeconômicas na indústria de mídia. "Diversas redações estão sofrendo com a rápida mudança na tecnologia de mapeamento", disse DeBarros. "Muitos de nós, durante anos, utilizamos bem o Flash, mas a falta de suporte para Flash no iPad tem alimentado a exploração em mapas baseados em tecnologias de código aberto que

trabalham em uma variedade de ambientes online. Muitas redações estão sofrendo com isso, e o número de sessões sobre mapas na conferência refletiu isso."

Há também um contexto significativo de interesse no desenvolvimento de habilidades de jornalismo de dados. Mais de 166 jornais dos EUA pararam de publicar edições impressas ou fecharam completamente desde 2008, resultando em mais de 35.000 postos de trabalho perdidos ou em aquisições na indústria de jornal desde 2007.

"A crise econômica e a mudança fundamental no negócio de publicações impressas significam que os jornalistas estão mais conscientes sobre a parte comercial do que nunca", disse DeBarros, "e eu acho que a conferência refletiu isso mais do que no passado. Houve uma grande sessão sobre como transformar um bom trabalho em dinheiro por Chase Davis e Matt Wynn, por exemplo. Eu estive em um painel sobre as razões de negócio para lançar APIs. Existe um mal estar geral que a maioria dos jornalistas sente sabendo que a nossa indústria ainda enfrenta tempos econômicos difíceis. Assistir a uma nova geração de jornalistas abraçar a causa é emocionante."

Um aspecto notável dessa nova geração de jornalistas de dados é que ela não parece ou ecoa as redações do século 20.

"Esta foi a conferência com maior diversidade de que eu posso me lembrar", disse Herzog. "Eu vi mais mulheres e negros do que nunca. Tivemos jornalistas de dados de vários países: Coréia, Reino Unido, Sérvia, Alemanha, Canadá, América Latina, Dinamarca, Suécia e mais. Além disso, a conferência é também muito mais diversificada em termos de habilidades e interesses profissionais. Empresários da Web 2.0, programadores, defensores de dados abertos, especialistas em visualização de dados, educadores e desenvolvedores de aplicativos misturados com jornalistas tradicionais de CAR. Eu também vi um público mais jovem, uma nova geração de jornalistas de dados que está abraçando a causa. Para muitos dos participantes, esta foi a sua primeira conferência."

Que problemas o jornalismo de dados enfrenta?

Embora as ferramentas estejam evoluindo, ainda há imensos desafios pela frente, desde o lado da tecnologia em si até o treinamento de redações. "Um grande desafio não solucionado é tornar a análise de dados não estruturados mais fácil e rápida de fazer. Além de mim, trabalham com isso a Sarah Cohen, a equipe do DocumentCloud, equipes da AP e Chicago Tribune e muitos outros", disse Houston.

Há também a questão de melhorar o nível de matemática básica entre os profissionais da mídia. "Isso vai soar básico, mas ainda há jornalistas de todo o mundo que não podem abrir uma planilha do Excel, classificar os valores ou escrever uma função para determinar variação percentual", disse DeBarros. "E isso inclui um grande número de estagiários universitários que

vejo ano após ano, o que realmente me assusta. Programas de jornalismo precisam se atualizar e entender que vivemos em uma sociedade rica em dados, e conhecimentos de matemática e competências básicas de análise de dados são altamente relevantes para o jornalismo. Os mais de 400 jornalistas no NICAR ainda representam uma espécie de *outliers* na indústria, e isso tem que mudar se o jornalismo quiser continuar a ser relevante em uma cultura baseada na informação".

Qual o papel do jornalismo de dados em governos abertos?

No terceiro dia do NICAR 2012, eu falei sobre "jornalismo de dados abertos", o que, parafraseando Jonathan Stray, eu defino como a obtenção, curadoria e produção de reportagens com dados abertos de interesse público. Como alguém que seguiu o movimento de abertura de governos de perto por alguns anos, os paralelos com o que os hackers cívicos estão fazendo e o que esta comunidade de jornalistas de dados faz é inevitável. Eles estão focados em colocar dados para trabalhar pelo bem público, seja com ou sem fins lucrativos, para utilidade cívica ou responsabilidade governamental.

Para isso, é preciso que jornalistas de dados e programadores cívicos usem as poderosas ferramentas emergentes nas redações para a explosão de informações digitais de governos, empresas e nossos concidadãos.

A necessidade de jornalismo de dados, no contexto de enormes quantidades de dados do governo que estão sendo liberados, não poderia ser mais oportuna, sobretudo tendo em conta os problemas de qualidade persistentes.

"Eu não consigo encontrar quaisquer desvantagens em se ter mais dados em vez de menos", disse Cohen, "mas eu me preocupo com algumas coisas."

Em primeiro lugar, enfatizou Cohen, há a questão de saber se os dados são abertos desde o início - e as consequências da 'higienização' feita antes do lançamento. "A demanda por dados estruturados, limpos com a finalidade de criar aplicativos, pode resultar em registros falsos em vez de registros reais liberados. USASpending.gov é um bom exemplo de que - nós não temos acesso aos registros reais de gastos, como faturas e ordens de compra usadas pelas agências, ou aos sistemas que eles realmente utilizam para fechar os negócios. Em vez disso, temos um sistema parcial cujo único objetivo é ser público, por isso não tem alta prioridade dentro de agências e não há nenhuma pista de auditoria natural nele. Ele não é usado para o registro do gasto público, por isso, os erros não são suscetíveis de serem descobertos. "

Para lidar com esse tipo de problema, Cohen recomenda obter mais fontes documentais, como um historiador faria.

"Eu acho que o que podemos fazer é pressionar ainda mais para obter registros reais, e não se contentar com o que o governo quer nos dar", disse ela. "Nós também temos que melhorar no uso de registros que não estejam em formatos agradáveis e fáceis de usar - eles não nasceram assim, devemos aperfeiçoar a capacidade de usar os registros sob qualquer forma que eles existem."

Por que o jornalismo de dados e os aplicativos de notícia são relevantes?

Dado o contexto econômico e tecnológico, pode parecer que o jornalismo de dados responde por si mesmo. "CAR, jornalismo de dados, jornalismo de precisão, e aplicativos de notícias são todos cruciais para o jornalismo - e o seu futuro - porque eles extraem sentido das enormes quantidades de dados", disse Houston, "para que as pessoas possam entender o mundo e tomar decisões e criar políticas sensatas."

Considerando que o universo dos que praticam jornalismo de dados continua a ser uma pequena parcela da mídia no mundo, há claramente ainda uma necessidade de seus praticantes mais importantes mostrarem o porquê dele ser relevante, em termos de impacto.

"Nós estamos vivendo em uma cultura guiada por dados", disse DeBarros. "Um jornalista experiente no uso de dados pode aproveitar a API do Twitter ou uma planilha para encontrar notícias de um modo tão simples quanto usar o telefone para ligar para uma fonte. Não só isso, falamos a muitos leitores que estão acostumados a lidar com dados todos os dias - contadores, educadores, pesquisadores, profissionais de marketing. Se queremos captar sua atenção, precisamos falar a língua dos dados com autoridade. E eles são inteligentes o suficiente para saber se fizemos nossa pesquisa corretamente ou não. Quanto aos aplicativos de notícia, eles são importantes porque - quando bem executados- podem fazer grandes quantidades de dados ser facilmente compreensíveis e relevantes para cada pessoa que for usá-los".

Novas ferramentas, mesmas regras

Embora as plataformas e ferramentas para o jornalismo estejam evoluindo e as fontes de dados estejam crescendo, muitas outras coisas não mudaram. Por exemplo, a ética que norteia as escolhas da profissão permanece central para o jornalismo do século 21, como o novο guia de ética da NPR deixa claro.

Se os desenvolvedores de notícias estão processando dados em tempo real, validando dados do mundo real ou melhorando a cobertura de notícias com os dados, o bom jornalismo de dados ainda deve contar uma história.

Se a ferramenta é um smartphone, notebook ou conjunto de dados, ela também deve abranger a reportagem investigativa, como enfatizou Doug Smith do Los Angeles Times na conferência NICAR.

Se o texto é a próxima fronteira em jornalismo de dados, aproveitando o poder do *big data*, isso acontecerá a serviço de contar histórias de forma mais eficaz. Jornalismo digital e ciências humanas digitais estão se fundindo a serviço de uma sociedade mais informada.