Strata

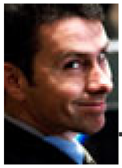
O'REILLY®

Strata

Making Data Work

Join the Data Revolution.

- Twitter
- Facebook
- Google
- LinkedIn
- YouTube
- Subscribe to Strata



In the age of big data, data journalism has profound importance for society

The 2012 NICAR conference highlighted how new tools and platforms are energizing the practice of journalism.

by Alex Howard | @digiphile | +Alex Howard | Comment | March 1, 2012

The promise of data journalism was a strong theme throughout the National Institute for Computer-Assisted Reporting's (NICAR) 2012 conference. In 2012, making sense of big data through narrative and context, particularly unstructured data, will be a central goal for data scientists around the world, whether they work in newsrooms, Wall Street or Silicon Valley. Notably, that goal will be substantially enabled by a growing set of common tools, whether they're employed by government technologists opening Chicago, healthcare technologists or newsroom developers.

At NICAR 2012, you could literally see the code underpinning the future of journalism written – or at least projected – on the walls.

"The energy level was incredible," said David Herzog, associate

Profiles

- The API architect
- The Daily Visualizer
- The Data Editor
- The Long Form Developer
- The Elections Developer
- The Human Algorithm
- The Visualizer

professor for print and digital news at the Missouri School of Journalism, in an email interview after NICAR. “I didn’t see participants wringing their hands and worrying about the future of journalism. They’re too busy building it.”

- The Storyteller and The Teacher
- The Hacks Hacker

Just as open civic software is increasingly baked into government, open source is playing a pivotal role in the new data journalism.

“Free and open-source tools dominated,” said Herzog. “It’s clear from the panels and hands-on classes that free and open source tools have eliminated the barrier to entry in terms of many software costs.”

While many developers are agnostic with respect to which tools they use to get a job done, the people who are building and sharing tools for data journalism are often doing it with open source code. As Dan Sinker, the head of the Knight-Mozilla News Technology Partnership for Mozilla, wrote afterwards, journo-coders took NICAR 12 “to a whole new level.”

- The PANDA Project officially launched in beta in St. Louis, including a provisioning party. You can check out PANDA on Github.
- The Associated Press’ Overview Project is going to make digging through documents easier.
- The LA Times Datadesk shared a tool, Django Bakery, to turn Django applications into flat HTML files.
- Jonathan Soma, with help from John Keefe of WNYC, built Tabletop.js as a way to use a public Google Spreadsheet as a source of data for a Web app. Tabletop is on Github too.

While some of that open source development was definitely driven by the requirements of the Knight News Challenge, which funded the PANDA and Overview projects, there’s also a collaborative spirit in evidence throughout this community.

This is a group of people who are fiercely committed to “showing your work” — and for newsroom developers, that means sharing your code. To put it another way, code, don’t tell. Sessions on Python, Django, mapping, Google Refine and Google Fusion tables were packed at NICAR 12.

> Journalists want to learn python: #nicar12 twitter.com/MeghanHoyer/st…
>
> — Meghan Hoyer (@MeghanHoyer) February 24, 2012

No, this is not your father’s computer-assisted reporting.

“I thought this stacked up as the best NICAR conference since the first in 1993,” said Herzog. “It’s always been tough to choose from the menu of panels, demos and hands-on classes at NICAR conferences. But I thought there was an abundance of great, informative, sessions put on by the participants. Also, I think NICAR offered a good range of options for newbies and experts alike. For instance, attendees could learn how to map using Google Fusion tables on the beginner’s end, or PostGIS and qGIS at the advanced level. Harvesting data through web scraping has become an ever bigger deal for data journalists. At the same time, it’s getting easier for folks with no or little programming chops to scrape using tools like spreadsheets, Google Refine and ScraperWiki. ”

> We’ve scraped a webpage using a simple chrome extension with @michelleminkoff at #nicar12 twitter.com/ChrisLKeller/s…
>
> — Chris Keller (@ChrisLKeller) February 25, 2012

On the history of NICAR

According to IRE, NICAR was founded in 1989. Since its founding, the Institute has trained thousands of journalists how to find, collect and public electronic information.

Today, “the NICAR conference helps journalists, hackers, and developers figure out best practices, best methods,and best digital tools for doing journalism that involves data analysis and classic reporting in the field,” said Brant Houston, former executive director of Investigative Reporters and Editors, in an email interview. “The NICAR conference also obviously includes investigative journalism and the standards for data integrity and credibility.”

“I believe the first IRE-sponsored [conference] was in 1993 in Raleigh, when a few reporters were trying to acquire and learn to use spreadsheets, database managers, etc. on newly open electronic records,” said Sarah Cohen, the Knight professor of the practice of journalism and public policy at Duke University, in an email interview. “Elliott Jaspin was going around the country teaching reporters how to get data off of 9-track tapes. There really was no public Internet. At the time, it was really, really hard to use the new PC’s, and a few reporters were trying to find new stories. The famous ones had been Elliott’s school bus drivers who had drunk driving records and the Atlanta Color of Money series on redlining.”

“St. Louis was my 10th NICAR conference,” said Anthony DeBarros, the senior database editor at USA Today, in an email interview. “My first was in 1999 in Boston. The conference is a place where news nerds can gather and remind themselves that they’re not alone in their love of numbers, data analysis, writing code and finding great stories by poring over columns in a spreadsheet. It serves as an important training vehicle for journalists getting started with data in the newsroom, and it’s always kept journalists apprised of technological developments that offer new ways of finding and telling stories. At the same time, its connection to IRE keeps it firmly rooted in the best aspects of investigative reporting — digging up stories that serve the public good.

Baby, you can drive my CAR

Long before we started talking about "data journalism," the practice of computer-assisted reporting (CAR) was growing around the world.

"The practice of CAR has changed over time as the tools and environment in the digital world has changed," said Houston. "So it began in the time of mainframes in the late 60s and then moved onto PCs (which increased speed and flexibility of analysis and presentation) and then moved onto the Web, which accelerated the ability to gather, analyze and present data. The basic goals have remained the same. To sift through data and make sense of it, often with social science methods. CAR tends to be an "umbrella" term – one that includes precision journalism and data driven journalism and any methodology that makes sense of date such as visualization and effective presentations of data."

On one level, CAR is still around because the journalism world hasn't coined a good term to use instead.

"Computer-assisted reporting" is an antiquated term, but most people who practice it have recognized that for years," said DeBarros. "It sticks around because no one has yet to come up with a dynamite replacement. Phil Meyer, the godfather of the movement, wrote a seminal book called "Precision Journalism, and that term is a good one to describe that segment of CAR that deals with statistics and the use of social science methods in newsgathering. As an umbrella term, data journalism seems to be the best description at the moment, probably because it adequately covers most of the areas that CAR has become — from traditional data-driven reporting to the newer category of news applications."

The most significant shift in CAR may well be when all of those computers being used for reporting were connected through the network of networks in the 1990s.

"It may seem obvious, but of course the Internet changed it all, and for a while it got smushed in with trying to learn how to navigate the Internet for stories, and how to download data," said Cohen. "Then there was a stage when everyone was building internal intranets to deliver public records inside newsrooms to help find people on deadline, etc. So for much of the time, it was focused on reporting, not publishing or presentation. Now the data journalism folks have emerged from the other direction: People who are using data obtained through APIs who often skip the reporting side, and use the same techniques to deliver unfiltered information to their readers in an easier format the the government is giving us. But I think it's starting to come back together — the so-called data journalists are getting more interested in reporting, and the more traditional CAR reporters are interested in getting their stories on the web in more interesting ways.

Whatever you call it, the goals are still the same.

"CAR has always been about using data to find and tell stories," said DeBarros. "And it still is. What has changed in recent years is more emphasis toward online presentations (interactive maps and applications) and the coding skills required to produce them (JavaScript, HTML/CSS, Django, Ruby on Rails). Earlier NICAR conferences revolved much more around the best stories of the year and how to use data techniques to cover particular topics and beats. That's still in place. But more recently, the conference and the practice has widened to include much more coding and presentation topics. That reflects the state of media — every newsroom is working overtime to make its content work well on the web, on mobile, and on apps, and data journalists tend to be forward thinkers so it's not surprising that the conference would expand to include those topics."

What stood out at NICAR 2012?

The tools and tactics on display at NICAR were enough to convince Tyler Dukes at Duke to write that NICAR taught me I know nothing. Browse through the tools, slides and links from NICAR 2012 curated by Chrys Wu to get a sense of just how much is out there. The big theme, however, without a doubt, was data.

"Data really is the meat of the conference, and a quick scan of the schedule shows there were tons of sessions on all kinds of data topics, from the Census to healthcare to crime to education," said DeBarros.

What I saw everywhere at NICAR was interest not simply in what data was out there, however, but how to get it and put it to use, from finding stories and source to providing empirical evidence to back up other reporting to telling stories with maps and visualizations.

"A major theme was the analysis of data (using spreadsheets, data managers, GIS) that gives journalism more credibility by seeing patterns, trends and outliers," said Houston. "Other themes included collection and analysis of social media, visualization of data, planning and organizing stories based on data analysis, programming for web scraping (data collection from the Web) and mashing up various Web programs."

"Harvesting data through web scraping has become an ever bigger deal for data journalists," said Herzog. "At the same time, it's getting easier for folks with no or little programming chops to scrape using tools like spreadsheets, Google Refine and ScraperWiki. That said, another message for me was how important programming has become. No, not all journalists or even data journalists need to learn programming. But as Rich Gordon at Medill has said, all journalists should have an appreciation and understanding of what it can do."

Cohen similarly pointed to data, specifically its form. "The theme that I saw this year was a focus on unstructured rather than structured data," she said. "For a long time, we've been hammering governments to give us 'data' in columns and rows. I think we're increasingly seeing that stories just as likely (if not more likely) come from the unstructured information that comes from documents, audio and video, tweets, other social media — from government and non-government sources. The other theme is that there is a lot more collaboration, openness and sharing among competing news organizations. (Witness PANDA and census.ire.org and the New York Times campaign finance API). But it only goes so far — you don't see ProPublica sharing the 40+ states' medical licensure data that Dan scraped with everyone else. (I have to admit, though, I haven't asked him to share.) IRE has always been about sharing techniques and tools — now we're actually sharing source material."

While data dominated NICAR 12, other trends mattered as well, from open mapping tools to macroeconomic trends in the media industry. "A lot of newsrooms are grappling with rapid change in mapping technology," said DeBarros. "Many of us for years did quite well with Flash, but the lack of support for Flash on iPad has fueled exploration into maps built on open source technologies that work across a range of online environments. Many newsrooms are grappling with this, and the number of mapping sessions at the conference reflected this."

There's also serious context to the interest in developing data journalism skills. More than 166 U.S. newspapers have stopped putting out a print edition or closed down altogether since 2008, resulting in more than 35,000 job losses or buyouts in the newspaper industry since 2007.

"The economic slump and the fundamental change in the print publishing business means that journalists are more aware of the business side than ever," said DeBarros, "and I think the conference reflected that more than in the past. There was a great session on turning your good work into money by Chase Davis and Matt Wynn, for example. I was on a panel talking about the business reasons for starting APIs. The general unease most journalists feel knowing that our industry still faces difficult economic times. Watching a new generation of journalists come into the fold has been exciting."

One notable aspect of that next generation of data journalists is that it does not appear likely to look or sound the same as the newsrooms of the 20th century.

"This was the most diverse conference that I can remember," said Herzog. "I saw more women and people of color than ever before. We had data journalists from many countries: Korea, the U.K., Serbia, Germany, Canada, Latin America, Denmark, Sweden and more. Also, the conference is much more diverse in terms of professional skills and interests. Web 2.0 entrepreneurs, programmers, open data advocates, data visualization specialists, educators, and app builders mixed with traditional CAR jockeys. I also saw a younger crowd, a new generation of data journalists who are moving into the fold. For many of the participants, this was their first conference."

What problems does data journalism face?

While the tools are improving, there are still immense challenges ahead, from the technology itself to education to resources in newsroom. "A major unsolved challenge is making the analysis of unstructured data easier and faster to do. Those working on this include myself, Sarah Cohen, the DocumentCloud team, teams at AP and Chicago Tribune and many others," said Houston.

There's also the matter of improving the level of fundamental numeracy in the media. "This is going to sound basic, but there are still far too many journalists around the world who cannot open an Excel spreadsheet, sort the values or write an equation to determine percentage change," said DeBarros, "and that includes a large number of the college interns I see year after year, which really scares me. Journalism programs need to step up and understand that we live in a data-rich society, and math skills and basic data analysis skills are highly relevant to journalism. The 400+ journalists at NICAR still represent something of an outlier in the industry, and that has to change if journalism is going to remain relevant in an information-based culture."

In that context, Cohen has high hopes for a new project, the Reporters Lab. "The big unsolved problem to me is that it's still just too hard to use "data" writ large," she said. " You might have seen 4 or 5 panels on how to scrape data [at NICAR]. People have to write one-off computer programs using Python or Ruby or something to scrape a site, rather than use a tool like Kapow, because newsrooms can't
(and never have) invest that kind of money into something that really isn't mission-critical. I think Kapow and its cousins cost $20,000-$40,000 a year. Our project to find those kinds of holes and create, commission or adapt free, open source tools for regular reporters to use, not the data journalist skilled in programming. We're building communities of people who want to work on these problems."

What role does data journalism play in open government?

On the third day of NICAR 2012, I presented upon "open data journalism, which, to paraphrase Jonathan Stray, I'd define as obtaining, reporting upon, curating and publishing open data in the public interest. As someone who's been following the open government movement closely for a few years now, the parallels to what civic hackers are doing and what this community of data journalists are working on is unescapable. They're focused on putting data to work for the public good, whether it's in the public interest, for profit, in the service of civic utility or, in the biggest crossover, government accountability.

> Cool app shown at #NICAR12 today. Find possible serial killings in your area using FBI data. #COMO stats: twitter.com/ElizHagedorn/s…
>
> — Elizabeth Hagedorn (@ElizHagedorn) February 24, 2012

To do so will require that data journalists and civic coders alike apply the powerful emerging tools in the newsroom stack to the explosion of digital bits and bytes from government, business and our fellow citizens.

The need for data journalism, in the context of massive amounts of government data being released, could not any more timely, particularly given persistent quality issues.

"I can't find any downsides of more data rather than less," said Cohen, "but I worry about a few things."

First, emphasized Cohen, there's an issue of whether data is created open from the beginning — and the consequences of 'sanitizing' it before release. "The demand for structured, nicely scrubbed data for the purpose of building apps can result in fake records rather than real records being released. USASpending.gov is a good example of that — we don't get access to the actual spending records like invoices and purchase orders that agencies use, or the systems they use to actually do their business. Instead we have a side system whose only purpose is to make it public, so it's not a high priority inside agencies and there's no natural audit trail on it. It's not used to spend money, so mistakes aren't likely to be caught."

Second, there's the question of whether information relevant to an investigation has been scrubbed for release. "We get the lowest common denominator of information," she said. "There are a lot of records used for accountability that depend on our ability to see personally identifiable information (as opposed to private or personal information, which isn't the same thing). For instance, if you want to do stories on how farm subsidies are paid, you kind of have to know who gets them. If you want to do something on fraud in FEMA claims, you have to be able to find the people and businesses who get the aid. But when it gets pushed out as open government data, it often gets scrubbed of important details and then we have a harder time getting them under FOIA because the agencies say the records are already public."

To address those two issues, Cohen recommends getting more source documents, as a historian would. "I think what we can do is to push harder for actual records, and to not settle for what the White House wants to give us," she said. "We also have to get better at using records that aren't held in nice, neat forms — they're not born that way, and we should get better at using records in whatever form they exist."

Why do data journalism and news apps matter?

Given the economic and technological context, it might seem like the case for data journalism should make itself. "CAR, data journalism, precision journalism, and news apps all are crucial to journalism — and the future of journalism — because they make sense of the tremendous amounts of data," said Houston, "so that people can understand the world and make sensible decisions and policies."

Given the reality that those practicing data journalism remain a tiny percentage of the world's media, however, there's clearly still a need for its foremost practitioners to show why it matters, in terms of impact.

"We're living in a data-driven culture," said DeBarros. "A data-savvy journalist can use the Twitter API or a spreadsheet to find news as readily as he or she can use the telephone to call a source. Not only that, we serve many readers who are accustomed to dealing with data every day — accountants, educators, researchers, marketers. If we're going to capture their attention, we need to speak the language of data with authority. And they are smart enough to know whether we've done our research correctly or not. As for news apps, they're important because — when done right — they can make large amounts of data easily understood and relevant to each person using them."

New tools, same rules

While the platforms and toolkits for journalism are evolving and the sources of data are exploding, many things haven't changed. For one, the ethics that guide the choices of the profession remain central to the journalism of the 21st century, as the new NPR's new ethics guide makes clear.

Whether news developers are rendering data in real-time, validating data in the real world, or improving news coverage with data, good data journalism still must tell a story. And as Erika Owens reflected in her own blog after NICAR, looking back upon a group field trip to the marvelous City Museum in St. Louis, journalism is also joyous, whether one is "crafting the perfect lede or slaying an infuriating bug."

> This is how we roll at these data journalism conferences #pandaproject #nicar12yfrog.com/nu2drnoj
>
> — jonathanstray (@jonathanstray) February 24, 2012

Whether the tool is a smartphone, notebook or dataset, these tools must also extend investigative reporting, as the Los Angeles Times Doug Smith emphasized to me at the NICAR conference.

If text is the next frontier in data journalism, harnessing the power of big data, it will be in the service of telling stories more effectively. Digital journalism and digital humanities are merging in the service of more informed society.

Profiles of the data journalist

To learn more about the people who are redefining the practice computer-assisted reporting, in some cases, building the newsroom stack for the 21st century, Radar conducted a series of email interviews with data journalists during the 2012 NICAR Conference. The first two of the series are linked below:

- ProPublica developer Dan Nguyen is redefining how longform journalism is told through data.
- Derek Willis is putting data journalism to work on campaign finance and elections at the New York Times
- Ben Welsh is augmenting the investigative work of the Los Angeles Times with open source coding.
- Michelle Minkoff is adding context and clarity to data with interactive features at the Associated Press.
- Jacob Harris is building APIs and data into elections coverage at the New York Times
- Matt Stiles oversees data journalism on NPR's State Impact project
- Meghan Hoyer is a data editor at The Virginan Pilot in Virginia.