

A journalist who sought to democratize data

BY: LOUISE LIEF

Posted: March 9, 2017 | Tags: data

Earlier this year, some of the nation's top data journalists flew to Washington to pay their respects to one of their own. My Investigative Reporting Workshop colleague David Donald, former data editor at the watchdog journalism nonprofit Center for Public Integrity, died in December after battling cancer. His colleagues gathered to celebrate his life and remarkable professional achievements.



IRW photo

Data Editor Jennifer LaFleur designed “D-squared” buttons to honor her friend and co-trainer David Donald.

and remarkable professional achievements.

Before he became ill, Donald and his colleagues had been developing a new, more democratic vision for data journalism, one that would invite the public into the process. His efforts came at a critical moment, as the integrity of data is being tested in a post-truth world.

I met Donald at the Workshop. He, Chuck Lewis, the workshop's executive editor, and I shared an interest in making data journalism's tools and techniques more available to the public and other disciplines.

“What fascinated him and still fascinates me,” says Lewis, “is that this is a little too esoteric of a field, intended for the ‘cognoscenti.’ He and I felt that wasn't necessary. For the future of truth, we need to knock those walls down... Journalists are wonderful, but they're not the only people into accessing primary records, nor should they be. The world is too big, and there are not enough journalists.”

One step Donald took to lower barriers was designing a workshop on storytelling with his friend and colleague Maggie Mulvihill at Boston University. “There are not a lot of opportunities for a regular person to get this training,” says Mulvihill, co-founder of the New England Center for Investigative Reporting and a professor of computational journalism at Boston University. “We're all drowning in data. It's overwhelming.”

Rather than have journalists just train each other, she and Donald thought that by offering a workshop to the outside world they could educate people about data journalism while creating a new revenue stream to support nonprofit journalism.

“All occupations use data,” says Mulvihill. “And everyone has to tell a story to his or her boss.”

Their inaugural workshop in early 2016 sold out in three weeks. They were surprised by who signed up. Housewives re-entering the workforce, social workers, real-estate brokers, designers, business managers, pediatricians, even zoologists all wanted to learn how to tell stories with data. Participants were eager to learn how investigative journalists think and work. Donald suggested using the open source statistical program “R” to make data comprehensible and useful to non-statisticians, a move she says was instrumental in making the workshop a success.

In late 2016, more incentives have emerged to lower barriers. As worries increased that the new Trump administration might delete environmental and other scientific data that did not support its political agenda, a [team at the University of Pennsylvania](#) launched the [Data Refuge Project](#) to harvest and secure vulnerable government scientific and environmental data. Soon after, a more comprehensive [Libraries Network](#) project launched to rescue data at 15 federal agencies.

The projects copy federal data sets and secure them in multiple locations in the event the originals are altered or disappear. A virtual “chain of custody” tags metadata to track changes and prevent unauthorized alterations. Data Refuge events have spread to colleges and universities in 17 states and the District of Columbia.

Data journalism picks up where these data rescue efforts leaves off. To understand the role it can play, it helps to know how this small, specialized branch of journalism came to be.

The field's guru is Philip Meyer, a former journalist and emeritus Knight Chair in Journalism at the University of North Carolina at Chapel Hill. Data journalism's version of the Pulitzer Prize, the Philip Meyer Journalism Award, is named after him. His seminal book, “[Precision Journalism](#),” was first published in 1973 and updated through 2002. It was the first book Donald gave me.

Meyer's thesis, revolutionary at the time, was that journalists should apply social-science research methods to newsgathering and conduct a “disciplined search for verifiable truth” by incorporating powerful data and analytic tools. He argued that journalism's “objectivity model” was intended for a simpler, bygone era, and couldn't adequately manage the growing complexity of information.

“The scientific method,” he wrote, “is still the one good way invented by humankind to cope with its prejudices, wishful thinking and perceptual blinders.”

Jennifer LaFleur, senior editor for data journalism at the Center for Investigative Reporting and former director of computer-assisted reporting at ProPublica, co-taught data workshops with David at various conferences. She is also the former training director for the National Institutes of Computer Assisted Reporting. The field has not

spread as rapidly in journalism as she thought it would. Many journalists find the math intimidating. “Telling [journalists] to do a regression model makes their heads explode,” she says.

One of data journalism's strengths is that it gets beyond personalities and identifies broader trends and systemic issues. One of its drawbacks is you can really screw up.

“If your methodology is flawed or your data is incomplete,” says Lewis, “you look like an idiot and get skewered.”

Even under normal circumstances, government data is not pristine. It often contains discrepancies and omissions. It needs to be “cleaned” and, as data journalists like to say, “interviewed.” Such issues surfaced in a 2009 award-winning project Donald did at the Center for Public Integrity examining how colleges respond to campus sexual assault. That series was done in collaboration with five regional nonprofit news organizations in New England, Wisconsin, Texas and Western states, all members of the Institute for Nonprofit News: <https://inn.org/>

Mulvihill was the lead reporter in New England. She says they discovered huge gaps in Department of Education data. Schools were required to report disciplinary actions for guns and armed robbery but not for rape or sexual assault. Donald and his colleagues found other government databases containing some of the information they needed and built new ones based on their own questionnaires and information culled from legal documents.

As computing power increased, Donald became one of the early pioneers to use big data for journalism. A complex 2012 project at CPI on [Medicare billing abuses](#) analyzed more than 700 million medical claims. That and another big data Medicare story netted him two Philip Meyer awards.

A central Meyer tenet is that journalists should show their work, describe their methods and explain how they reached their conclusions. Arizona State University Knight Chair in Journalism Steve Doig, who teaches data journalism, describes it as “open-source peer review.” Journalists publish their results and the public can pick it apart. One of Donald's big strengths, says Doig, was his rigorous approach to methodology. A scholar of applied statistics, “He had learned the power and limitations of different statistical techniques.”

Some researchers view data journalism as a watered-down version of academic social science. But though they use social-science research methods, data journalists see what they do as a different animal, built for a different purpose. They focus on storytelling for the public.

“Journalists are cross-disciplinary. They take methods from a lot of different places,” says Brant Houston, the former executive director of Investigative Reporters and Editors (IRE) and current Knight Chair of Investigative and Enterprise Reporting at the University of Illinois, who had hired Donald as a data trainer. “Social scientists find us very creative in the ways we think about this.”

Journalists also place a premium on timeliness. Any editor, says Lewis, will ask journalists two questions: “What have you got? How long will it take?” And there is always the omnipresent, “Why do we care?”

In a period when many academics, journalists, civil-society groups and citizens fear that federal data may be altered to suit political agendas, the tools and techniques data journalists use will help keep it honest. They are the watchdogs of the data world. They scrutinize methodologies, compare data from different sources and interview the people who did the analyses. They distill their findings into a story and present it to the public.

Donald, with his persistent, systematic approach to what seem like impossible projects, would have welcomed the coming challenge. Gifted teacher that he was, he would have taught others how to do it, too.

But the scale of this is formidable. Donald's Medicare projects involved around 5 terabytes of data and took two years. With federal data we are talking about many hundreds of terabytes of data and dozens of databases.

Now would be a good time to knock down some walls.

Louise Lief is a scholar in residence at the Investigative Reporting Workshop.