SEARCH

Enter a search term... Search

ABOUT IRE CONFERENCE BLOG

Keep up with the latest news out of the IRE Conference. Students from the University of Missouri School of Journalism and universities near conference sites will be writing posts about sessions throughout the conference. IRE staff members will post insights and links to training materials as well.

IRE CONFERENCE BLOG ARCHIVES

2015
December 2015
November 2015
October 2015
September 2015
August 2015
July 2015
June 2015
May 2015
April 2015
March 2015
February 2015
January 2015
2014
2013
2012
2011
2010
2009
2008

IRE BLOGS

IRE News
Extra Extra
On the Road
Uplink
Transparency Watch

HOME » IRE CONFERENCE BLOG » DEALING WITH INACCESSIBLE DATA AND FINDING A ...

Dealing with inaccessible data and finding a needle in a million haystacks

By IRE Conference Blog | 06.20.2013

IRE Conference Blog

TAGGED
data cleaning, data, PDF, IRE Conference, crowdsourcing

Leave a comment

By Jordan Gass-Poore'

Amanda Zamora of ProPublica answers questions during a panel on how to build a thorough data-based investigation with inaccessible, incomprehensible, and indeterminate data. Photo: Travis Hartman.

Leading journalism professionals spoke about the search for finding meaning in messy data during Thursday morning's session "Finding the needles in a million haystacks: How to build a thorough data-based investigation with inaccessible, incomprehensible and indeterminate data."

Amanda Zamora, ProPublica engagement editor, used the publication's "Free the Files" initiative -- last fall's attempt to make sense of thousands of political campaign ad spending documents from various U.S. markets -- to give examples of how best to give data, and the public, a voice.

The initiative helped to reveal hidden spending, or "dark money," in political elections through various crowd-powered avenues, like a Facebook group, Leaderboard (which tracked the status of thousands of volunteers) and an app that helped translate the files into structured data.

ProPublica initially enlisted volunteers to contact their local television stations and request access to the "public inspection file." Zamora explained the reason for this was because stations had been required by the FCC to keep detailed records of political ad buys, but they were only available on paper and required an in-person visit.

Zamor said most of those documents -- an estimated 40,000 files from 200 stations -- were posted as difficult-to-search PDFs.

The result of the ProPublica initiative resulted in the cataloging of $1.12 billion worth of ad buys, Zamora said.

When working on a story that incorporates data, Zamora suggested narrowing the scope and channeling a crowd for support.

To view her entire presentation, visit bit.ly/FreeTheFilesIRE, or for tips on story ideas using Free the Files visit bit.ly/freethefiles-recipe.

Robert Benincasa, NPR producer, also provided tips on how to use data to tell stories, like those the company used to investigate whether or not the Mexican government favored a particular drug cartel.

The investigation began with Spanish-language press releases on drug arrests. NPR later examined PDFs of cartel IDs, names and charges.

After manually entering the data into a Microsoft Access database, NPR was able to analyze the information and report that the Sinaloa cartel might be favored by the Mexican government.

Benincasa said that one cartel represented 40 percent of arrests with more than 1,100 individuals arrested.

“It was pretty clear why they were arrested and who it was,” he said in response to a seminar attendee’s question.

Glen Howatt, computer-assisted reporting editor at the Star Tribune, warned that must realize when using data that something will always go wrong because of missing or complicated information that might have resulted from pressures put on those who compiled it.

Howatt said reporters should take preventative measures when using data, like requesting comprehensible data sets (and avoid those that are pre-packaged), link to other data sets and get documentation explaining how the data are compiled.

“Try to find some kind of form from which the data originated,” he said, adding that reporters can then compare it to the data set given.

Questions Howatt said reporters should ask themselves are do the numbers add up and do the names and dates of birth match.

He said ways to confirm these questions are by switching the first and middle names, checking nicknames and date-of-birth transpositions, like 9/21 versus 9/12.

“Take control of data,” Howatt said.

Jordan Gass-Poore’ is a journalism student at Texas State University

Log in or register to comment on this story.

IRE INVESTIGATIVE REPORTERS & EDITORS

141 Neff Annex
Missouri School of
Journalism
Columbia, MO 65211

573 882 2042

info@ire.org

Staff Directory

Advertise With Us

Privacy Policy

DATA AND RESOURCES

Database Library

Story Library

Tipsheets

Story packs

Listservs

NEWS AND INFORMATION

IRE News

Extra Extra

Uplink

IRE Journal

Member News

Job Postings

OUR ORGANIZATION

About Us

History

Board of Directors

Legal Documents

GET INVOLVED

Join

Donate

Contact