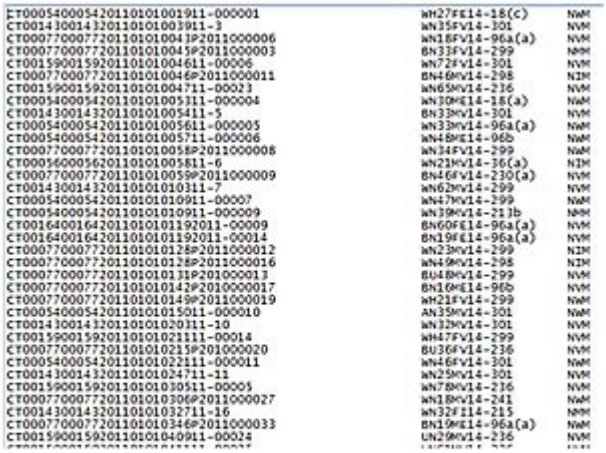
New England Newspaper Press Association (NENPA) E-Bulletin
Published April 24, 2014

Available online with more visuals at: http://www.seapubs.com/eBulletin4-24-14/

Using tools, techniques that help data miners to dig deep

By Aneri Pattani
Bulletin Staff

When Matthew Kauffman, investigative reporter at The Hartford (Conn.) Courant, wanted to do a story on mental health in the military, he didn't limit his reporting to anecdotes and interviews. Instead, he and co-author Lisa Chedekel – also then at the Courant – gathered information on mental health evaluations before deployment. Statistical analysis of that data became the foundation of their story.



What the raw data for traffic stops looks like on a computer screen.

"We really wanted to have a statistical element to the story, and that was what made it compelling," Kauffman said. "Combining a systemic big picture with anecdotal smaller pictures made a better story."

Kauffman is one of many journalists turning to data more frequently for their stories. The surge in data-driven reporting has even led to the creation of a new subfield: data journalism.

Data journalism – often also called computer-assisted reporting – is the convergence of several fields, including programming, design, statistics and investigative research, according to an article published by Poynter Institute.

Todd Wallack, staff reporter for the Spotlight team at The Boston Globe, sees data journalism as reporting that involves more intensive data and tools.

"It is about trying to gather data in a more intensive way than just writing down a statistic you hear at a press conference," he said.

That could mean conducting a poll or survey, downloading readily available data and doing statistical analysis, filing for public records and going through them, or visually showing that data in a different way, Wallack said.

Although the title of data journalism covers a vast array of tasks, most can be placed into one of three categories:
gathering data, analyzing data or visualizing data. Countless tools – the majority of which are free – are available to assist reporters with each aspect.

Gathering Data

The first step to any data story is to compile all the necessary information. That often involves accessing various public and private databases.

Many journalists are finding that to be an easier task than in previous years because so much data is now available for free online.

'It's useful to know a little bit about how programming works and to have some people in the newsroom know how to code, but not everyone needs to know that particular skill.'

-- Todd Wallack,
Staff reporter, Spotlight team, Boston Globe

Paul Parker, a staff writer at The Providence (R.I.) Journal, noted that trend during the past 10 years.

“A lot of data you used to have to request is now just routinely put online. There are still occasional fights to get more sensitive or detailed data, but they are fewer and further between,” Parker said.

The increasing availability of information often simplifies a reporter’s job, but sometimes it is still necessary to hunt down the desired data. That might involve using the federal Freedom of Information Act, asking multiple sources, or simple persistence, which Kauffman explained is a vital component of data journalism.

“Making assumptions is usually a dangerous thing in journalism, but one assumption I want you to confidently make is that if you want data and you have any reason to believe someone is collecting it, then it is out there and you will get it,” he said.

Although traditional reporting tactics are important in acquiring data, sometimes the process requires new tools as well. Those often include coding languages, which can be used to scrape data off a website.

Scraping data, as Kauffman describes it, is building a robot to run the same type of queries over and over again to get all the information you want out of a database that is designed to only dole out one piece at a time. Using a robot -- that is, a computer program -- turns what could be a tedious manual task into an efficient algorithm.

Scraping can be useful, Wallack said. He mentioned an instance where the state of Massachusetts published a report with the percentage of people who smoked in each community. The information for each community was stored in a separate spreadsheet. To aggregate all the percentages meant accessing 351 separate spreadsheets. Rather than attempting that task by hand, Wallack wrote a program that searched each file to pull out the number he wanted and listed that in a spreadsheet.

While Wallack appreciates the ability that comes from knowledge of a coding language, he does not think that that is a necessary tool for all journalists.

“It’s useful to know a little bit about how programming works and to have some people in the newsroom know how to code, but not everyone needs to know that particular skill,” he said.

According to Wallack, it is more important to understand data conceptually.

“The main thing reporters need to know is how data can be used for stories,” he said. “Learn a little bit about what you can do about getting data, what you can get, what you can do with it, and start asking for data more and more frequently.”

Parker seconded the idea that not all reporters need to know how to code, but he recommended understanding programming concepts such as loops and object-oriented programming.

The coding languages used most in the field of journalism include Python, Ruby and R, according to an article published by Poynter Institute. For those interested in enhancing their skill set, those are good coding languages to explore.

For those who cannot scrape data through coding, there is a simpler tool called Import.io. It is a visual scraper, which means it can go to a website with formatted data -- for example, the Academy Awards website -- and use a visual interface to locate the title of each movie nominee. It will then collect that data in a table. In that way, even non-coders can scrape data.

> *'Making assumptions is usually a dangerous thing in journalism, but one assumption I want you to confidently make is that if you want data and you have any reason to believe someone is collecting it, then it is out there and you will get it.'*
>
> -- **Paul Parker,**
> *Staff writer,*
> *Providence (R.I.) Journal*

Analyzing Data

Once reporters compile a set of data, they have to analyze it to see whether it contains a story worth telling.

Many times before the data can be viewed critically for trends and patterns, it has to be cleaned. This means correcting misspellings and formatting everything consistently.

Google Refine is a great, free tool for data cleaning, said Patty Reaves, user experience and audience manager at the Bangor (Maine) Daily News. Reaves uses Google Refine to manipulate “very large, messy data sets.”

The Courant’s Kauffman also endorses it as “very intuitive and easy to use.” He recalled an instance when he was looking through death certificate data from a medical examiner’s office and encountered several different ways of referring to heart attacks. He used Google Refine to standardize the data and make it easier for him to analyze.

After data cleaning comes time for critical analysis. One of the most powerful tools for that task, is Microsoft Excel, Kauffman said. If you learn the basics of arithmetic and pivot tables – interactive tables that extract, organize and summarize data through comparisons and pattern detection – in Excel, there is a lot you can do, he said.

Kauffman said knowing something as simple as how to calculate the change between two numbers, possibly the town budget from one year to the next, can add a new dimension to a story.

When it comes to exceptionally large data sets, Excel might not always be an option because it is limited to 1 million records. In that case, Wallack suggests using Microsoft Access, which can hold up to two gigabytes of data, or MySQL, which is another database manager. It is a free, open-source tool.

MySQL is essentially a way to tell a database program how to pull certain types of data out and how to match separate information and mash it together, Wallack said.

For the most intricate data analysis, it can be useful to have a statistical package like SPSS or its free, open-source counterpart, PSPP. Those programs can run tests that indicate the significance of certain trends in data.

Many reporters think that that level of analysis is not particularly useful in journalism, however.

“It’s rare that I’m ever doing analysis that requires calculating p-values or confidence intervals,” Kauffman said. “Typical journalism doesn’t get into causation as much, which is the main reason you’d need such statistical testing. Usually raw numbers are just as meaningful.”

Wallack seemed to agree that most reporters only need to understand the basics of statistics – percentages, rates, median vs. mean, and margins of error.

For anyone interested in gaining statistical knowledge for the purposes of journalism, Wallack recommends reading Numbers in the Newsroom: Using Math and Statistics in the News by Sarah Cohen.

Visualizing Data

Typically the final component of a data story involves visualizing data. That includes the creation of charts, graphs, maps and other forms of graphic design.

It is the element of data journalism that often receives the most attention. It is what draws readers to the story, as shown by the famous “Snowfall” feature in The New York Times.

Kauffman said the power of data visualization comes from giving readers a lot more content than traditional print.

“Instead of that classic ugly newspaper, you get projects that take up the entire screen. As you scroll, charts and photos and videos fade into the screen. This is journalists taking advantage of the multimedia tools of the Web,” he said.

One highly recommended tool for creating charts, graphs and other interactive visuals is Tableau Public. Kauffman refers to it as a “robust data visualizer.”

“It gives you lots of options for what data looks like and is shockingly easy for making graphs that are more interactive than the static ones created in Excel,” he said.

Parker uses Tableau for mapping as well, although other tools such as Google Fusion and QGIS are more popular.

Google Fusion is optimal for simple mapping projects, which typically refer to geographical maps although the term can also include diagrams. Fusion works with basic x and y coordinates to create a straightforward map layout. Its intuitive design allows for a quick learning curve, Kauffman said.

In comparison, QGIS is a more advanced open-source geographic information system. It allows users to edit, visualize, analyze and publish geospatial information on almost any operating system, including Windows, Mac, Linux, and BDS.


Patty Reaves

Although those tools provide convenient platforms for journalists to create visuals, most advanced graphic design requires some knowledge of HTML or JavaScript, Wallack said. Those coding languages provide the flexibility to create interactive programs that can bring the story to life for readers, he said.

For many aspiring data journalists, the abundance of tools can seem overwhelming. The flipside of overwhelming, however, is exciting – at least that is how the Bangor Daily News' Reaves sees it.

"It's a really exciting time to be a journalist because the number of tools and the abilities we have to define stories is amazing," she said

Side Bar Featured with the Article

Free tools for data journalists

Gathering Data

Coding languages -- Useful for scraping data from websites. Recommended: Python, Ruby, R

Import.io – Simple visual scraper, no coding knowledge required

Analyzing Data

Google Refine – Data cleaner that removes misspellings and standardizes format

Excel – Useful for arithmetic calculations and some basic statistical analysis

MySQL – Database manager for larger data sets than Excel can process

Visualizing Data

Tableau Public – Create charts, graphs and some maps; more interactive than Excel-produced graphics

Google Fusion – Basic mapping tool; uses x and y coordinates

QGIS – More advanced mapping capabilities; functions on almost any operating system

HTML/JavaScript – Coding languages used for more free-form graphic design