Data Leading the Future of Journalism Postprint

Authors: Cao Shen-shen

Date: 2023-10-08T00:00:00+00:00

Abstract

The advent of the big data era has brought new opportunities to traditional journalism. The emergence and development of data journalism have enriched the content and modalities of news reporting. This article begins with the definition of data journalism, systematically reviews its origins, development, and reporting methodologies, and through multiple case studies from the data journalism platforms ProPublica and The Upshot, offers reflections and summaries on the stages of data journalism reporting and the characteristics it exhibits.

Full Text

Abstract and Keywords

Abstract: The advent of the big data era has brought new opportunities to traditional journalism. The emergence and development of data journalism have enriched both the content and methods of news reporting. This paper begins with the definition of data journalism, reviews its origins, development, and reporting methods, and through multiple case studies from two data journalism platforms—ProPublica and The Upshot—reflects on and summarizes the stages and characteristics of data journalism reporting.

Keywords: data journalism; data; personalization; visualization
CLC Number: G210
Document Code: A
Article ID: 1671-0134(2018)08-112-03
DOI: 10.19483/j.cnki.11-4653/n.2018.08.046
Author: Cao Shenshen (China Civil Aviation News)

Data and algorithms are transforming every industry and reshaping our lives. Today, even storytelling relies on data. Several years ago, the prestigious Columbia University launched a distinctive journalism program called the Lede,

which quickly became a model for data journalism education. The program offers two tracks for students with little or no computer science background but who are eager to learn programming: first, an intensive 12-week summer bootcamp covering computing foundations, data and databases, algorithms, and data analysis studios; second, after completing the bootcamp, students may choose to continue with fall courses that more systematically cover data analysis and algorithms, including essential data structures, data analysis for social sciences, storytelling with data, and digital activism. Evidently, the Lede curriculum is grounded in social and human sciences, preparing students for statistically-driven media careers.

1. Data Generates News

Although the term "big data" has become familiar in recent years, understanding and exploration of what constitutes data journalism remain in their infancy. Jonathan Gray and colleagues write in *The Data Journalism Handbook* (2012) that superficially, data journalism means reporting news with data, yet both "data" and "journalism" are troubling terms—especially when we live in an age where every aspect of life can be described numerically. Data can provide sources for news, serve as a tool for storytelling, or exist in a mutually causal relationship with journalism. It represents "a way of reporting news that combines journalistic skills with the ability to work with big data information." Troy Thibodeaux states in "5 Tips for Getting Started in Data Journalism" that it involves "applying data to the process of producing and disseminating information," reflecting interactions between content producers and designers, computer scientists, and statisticians. Chinese scholar Zhang Gehao defines data journalism as "a method of creating news reports through data analysis and filtering," while Fang Jie and colleagues describe it as "a new form of news reporting based on data scraping, mining, statistics, analysis, and visualization."

Simon Rogers, an experienced data journalist who has worked at *The Guardian*, Twitter, and Google and founded the Datablog, published *Facts Are Sacred* in 2013. The book reviews and organizes data journalism pieces he produced during his 15 years at *The Guardian*, revealing the stories behind data journalism, how data changes our lives, and what we can learn from it. For Rogers, the primary task of data journalism is storytelling—not creating graphics or writing code. Data is open and belongs to everyone, and through its presentation, people can better understand news. Meanwhile, as digital technologies continue to develop and evolve, an increasing number of tools are becoming available to help media organizations analyze data and process results.

2. Data Journalism Reporting Methods

In August 2010, Deutsche Welle journalist Mirko Lorenz proposed four steps for data journalism reporting: first, data mining; second, data filtering; third, data visualization; and finally, news production. Similar to this is Rogers' introduc-

tion to the "data journalism workflow" on his personal blog. Rogers argues that data journalism presents a multi-linear, comprehensive reporting process that handles data on one hand while continuously verifying and questioning its credibility and value on the other, ultimately publishing reports through multiple means and channels.

Paul Bradshaw, a professor at Birmingham City University, proposed the "Inverted Pyramid of Data Journalism" in his "Double Pyramid Model," which more vividly reveals how data transforms during the dissemination process. For the data processing phase, Bradshaw uses an inverted pyramid structure comprising four parts: Compile, Clean, Context, and Combine. The purpose of data processing is to achieve effective dissemination, while data journalism communication follows a "positive pyramid structure." Bradshaw believes that reporting should incorporate visualization, narrativization, socialization, and humanization.

[Figure 1: see original paper]

In fact, regardless of the reporting method employed, data acquisition, processing, and presentation are indispensable steps in data journalism. The impact of big data on journalism is manifested in at least two aspects: first, applying technology to collect and deeply analyze data; and second, presenting news through visualization and interactive effects.

3. Data Journalism Practice: ProPublica and The Upshot

In the United States, the first names that come to mind when discussing data journalism are ProPublica and *The New York Times* data journalism section "The Upshot."

ProPublica, a non-profit organization headquartered in Manhattan, New York, was founded by former *Wall Street Journal* executive editor Paul Steiger. It is an independent newsroom dedicated to investigative reporting for the public interest. From its inception, ProPublica assembled a group of top journalism talent. In addition to editor-in-chief Steiger, former *New York Times* investigative editor Stephen Engelberg joined as executive editor, with all reporters and editors hailing from traditional media outlets.

On April 18, 2011, the 95th Pulitzer Prizes—the highest honor in journalism—were announced. ProPublica won the Pulitzer Prize for National Reporting for its work "The Wall Street Money Machine." The year-long project used visualization to comprehensively disclose how Wall Street financial giants secured massive profits for themselves. This was not ProPublica's first Pulitzer; in 2010, just two years after its founding, it shared the Investigative Reporting award with *The New York Times* for Sheri Fink's investigation into a hospital in Louisiana that administered lethal injections to patients during Hurricane Katrina. While that work appeared both on ProPublica's website and in print in *The Times*, the 2011 award marked the first time a purely digital-native investi-

gation won a Pulitzer, establishing ProPublica as one of America's most active data journalism sites. Today, ProPublica has 104 partners, including *The New York Times*, *The Washington Post*, *Los Angeles Times*, *USA Today*, MSNBC, and CNN.

ProPublica journalists view a news application as an interactive database that can tell stories—essentially a news work where software replaces text and images. "Dollars for Docs" is one such application, tracking hundreds of millions of dollars in payments from pharmaceutical companies to doctors in the form of consulting fees, speaking engagements, and travel. By creating this application, readers can search for their own doctors to see what compensation they have received, and journalists from partner news organizations can also use this data. Since its launch, more than 125 local news organizations have used the data to investigate local doctors, though only a handful were formal ProPublica partners—the majority conducted independent reporting using the application and data. The ProPublica team believes that if their data can help reporters who understand local contexts tell impactful stories, they have fulfilled their mission, which aligns perfectly with ProPublica's consistent goal of producing journalism that is "truly important" and "morally powerful."

If ProPublica is a product of the big data era, then The Upshot represents a traditional media outlet's foray into data journalism. In spring 2014, *The New York Times* launched The Upshot, a section focused on data journalism covering politics and economics. Led by former Washington bureau chief David Leonhardt, the launch team consisted of 15 members, including three full-time graphic designers. The Upshot aims to help audiences understand the significance behind news through data analysis and presentation. Beyond news coverage, The Upshot has designed many interesting interactive products that allow audiences to find information tailored to their needs and receive relevant recommendations through data analysis and modeling.

[Figure 3: see original paper]

Through the cases of The Upshot and ProPublica, three characteristics of data journalism become apparent. First, personalization: through a data journalism piece or an interactive product, audiences can develop resonance and find answers that match their own circumstances. Second, visualization: unlike traditional news that relies on "text + images," data journalism more often tells stories through charts, achieving effects that text alone cannot through graphical presentation and data analysis. For example, The Upshot's "Where is the hardest place to live in America?" uses maps and a series of indicators (overall ranking, average income, education level, unemployment rate, life expectancy, etc.) to show living conditions in each region. These political and economic themes may seem macro-level, but they are closely related to people's daily lives and thus have won audience appreciation. Third, teamwork: in the traditional news era, a reporter could complete a story independently, with editors, graphic designers, and proofreaders forming a production chain. In the data journalism era, news products are the result of team collaboration, requiring

reporters, designers, and programmers to jointly develop plans during production. Reporters provide material, programmers analyze and process data, and designers present visual effects—the three parties maintain a cooperative and interactive relationship.

References

[1] Huang Zhimin, Wang Min, Li Wei. Survey Report on Data Journalism Education. *News and Writing*, 2017(9).

[2] Jonathan Gray, Liliana Bounegru, Lucy Chambers. *The Data Journalism Handbook*. http://datajournalismhandbook.org/1.0/en/, 2012.

[3] Troy Thibodeaux. 5 Tips for Getting Started in Data Journalism. http://www.poynter.org/how-tos/digital-strategies/147734/5-tips-for-getting-started-in-data-journalism/, 2011.

[4] Zhang Gehao. Data Journalism for Open News—Data Journalism Practice at *The Guardian* in the UK. *Shanghai Journalism Review*, 2013(6).

[5] Fang Jie, Yan Dong. "Data Journalism" from a Global Perspective: Concepts and Practice. *Journal of International Communication*, 2013(6).

[6] Simon Rogers. Open Data Journalism. http://simonrogers.net/2013/01/24/open-data-journalism/.

[7] Simon Rogers. A Data Journalism Workflow. http://simonrogers.net/2013/01/27/a-data-journalism-workflow/.

[8] Image source: Bradshaw, P. The Inverted Pyramid of Data Journalism. http://onlinejournalismblog.com/2011/07/07/the-inverted-pyramid-of-data-journalism/.

[9] Image source: http://www.nytimes.com/interactive/2014/upshot/buy-rent-calculator.html.

Note: Figure translations are in progress. See original paper for figures.