Research on the Development and Practice of Data Journalism: A Case Study of The Paper's "Meishuke" Postprint

Authors: Guo Yan

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

Abstract

To investigate the development and practical status of domestic data journalism, this study selects the "Meishuke" data journalism column of The Paper as its research object. Using web scraping, it collected the titles, publication dates, keywords, like counts, and comment counts of all data news articles published since 2016, as well as their visual presentation forms. Detailed statistical data analysis was conducted to examine the column's practices in data journalism from perspectives including content focus, presentation forms, and communication effects. Finally, recommendations are proposed: enhancing data journalism quality through emphasis on topic selection and innovation in visual presentation forms, leveraging the distinctive features and advantages of data journalism to promote its development.

Full Text

Data Journalism Development and Practice Research: A Case Study of The Paper's "Meishu Ke"

Author: Guo Yan (Xinhua News Agency, China Economic Information Service Co., Ltd., Beijing 100052)

Abstract: To investigate the current state of data journalism development and practice in China, this study examines The Paper's "Meishu Ke" data journalism column. Using web scraping, we collected comprehensive data on all data journalism pieces published since 2016, including titles, publication dates, keywords, like counts, comment counts, and visualization formats. Through detailed statistical analysis, we analyze the column's practices in terms of content focus, presentation forms, and communication effects, and conclude with recommendations for improving data journalism quality through enhanced topic selection

and innovative visualization methods to better leverage the unique strengths of data journalism and promote its development.

Keywords: data journalism; Meishu Ke; visualization formats; communication effects

CLC Number: G210
Document Code: A
Article ID: 1671-0134(2021)04-098-03
DOI: 10.19483/j.cnki.11-4653/n.2021.04.027

Citation Format: Guo Yan. Research on the Development and Practice of Data Journalism—A Case Study of The Paper' s "Meishu Ke" [J]. China Media Technology, 2021(04): 98-100.

The concept of data journalism can be traced back to 2006, gradually gaining momentum around 2010. During this period, data journalism received increasing recognition within the industry and entered the public and governmental consciousness. Tim Berners-Lee, known as the "father of the internet," declared that data analysis would become a defining characteristic of future journalism [1].

This study selects The Paper, a news platform with a large user base, as its research subject. We scraped data from its "Meishu Ke" data journalism column over the past five years, conducting detailed statistical analysis to explore the characteristics, communication effects, development trends, and reflections on data journalism, ultimately offering recommendations for its future development and practice.

1.1 Overview of Data Journalism

Data journalism represents a new field in journalism that has emerged in the big data era. Grounded in open data, it transforms complex, abstract, and difficult-to-understand data into vivid, concrete, and engaging news reports.

In China, the rise of data journalism has largely kept pace with international developments. In 2011, Sohu launched its "Digital Path"column; the following year, NetEase, Sina, and Tencent successively established "Data Reading," "Graphic World," and "Data Control." In 2013, Caixin established a data journalism center and launched its "Digital Talk" column. In 2015, Yicai Media Group created DT Finance, a new media platform focused on financial data journalism content, while The Paper launched its "Meishu Ke" column. Concurrently, mainstream Chinese media outlets made bold attempts in data journalism: People' s Daily Online introduced its "Graphic News" channel in 2013, produced by the People' s Daily Online News Department, while Xinhua News Agency launched its data journalism project and "Data News" column in 2012 [2].

1.2 Current State of "Meishu Ke" Column Practice

The Paper's "Meishu Ke" column was established in 2015 with the motto: "Data is the skeleton, design is the soul. Related to news, yet unrelated to news." The column is categorized under "Current Affairs," focusing primarily on trending social topics and analytical reporting on breaking events. The column publishes an average of 14 data journalism pieces per month, employing visualization formats including static infographics, interactive visualizations, videos, and animations. Readers can like, comment on, share, and repost published pieces.

This study selects The Paper's "Meishu Ke" column for several reasons: its high publication frequency, large and active user base, and comprehensive production and operational capabilities that generate significant communication effects, providing rich analytical material [3].

1.3 Data Selection Methodology

Data was collected on March 1, 2021, capturing all data journalism pieces from the "Meishu Ke" column, including titles, publication dates, keywords, like counts, comment counts, and visualization formats. After excluding several reader-voting posts, the dataset comprised 795 data journalism pieces, with the earliest piece dating from June 2016.

2.1 Content Focus on Hot Topics

Data journalism topics can be categorized into event-based and topic-based selections. Event-based data journalism focuses on a specific news event, interpreting and presenting newsworthy data emerging from that event, commonly seen in coverage of major events such as conferences, sports competitions, or disasters. Topic-based data journalism concentrates on a particular news topic, collecting and analyzing data around that theme.

Analysis of the 795 scraped pieces reveals that the vast majority are topic-based selections, with event-based pieces accounting for only 7%. This is because major news events occur unpredictably and infrequently. Consequently, data journalism production typically selects popular topics and thought-provoking themes to generate reader interest.

Word frequency analysis of titles and keywords from the collected pieces identified the top ten high-frequency terms: "COVID-19," "epidemic," "United States," "Olympics," "election," "income," "film," "accident," "games," and "college entrance exam." This demonstrates that "Meishu Ke" focuses on social hot topics. Compared to traditional text-based news, data journalism collects and mines relevant data to provide deeper analytical reporting on trending social issues from a data perspective, offering more objective positioning that stimulates further reader reflection and enhances understanding of relevant issues.

2.2 Visualization Formats

Data journalism visualization formats can include static infographics, interactive visualizations, video/animation, VR/AR news, and others. The collected sample primarily employs three formats: static infographics, video/animation, and H5 interactive pages. H5 interactive pages and video/animation formats are relatively rare, accounting for 6% and 9% respectively, while static infographics dominate at 85%.

Interactive visualizations and video formats are more dynamic, novel, and interactive, better engaging readers and sparking interest, but are relatively more difficult and time-consuming to produce. Therefore, "Meishu Ke" predominantly uses static infographics, supplemented by H5 interactive and video/animation visualizations. This approach allows for bold experimentation with interactive and video/animation formats while ensuring timeliness and high update frequency through simpler, faster static infographic production.

3. Communication Effects Analysis

Communication effects can be measured through like and comment counts. This study collected like and comment data for each piece. Since H5 interactive pages lack like and comment functionality, 48 H5 interactive pieces were excluded from the 795-piece sample, leaving 747 pieces for analysis.

The progression from liking to commenting represents deepening reader engagement. After reading, satisfied readers may choose to like a piece. Those who feel a deeper resonance or develop their own insights may comment on the article or engage with other readers' comments, representing a natural deepening of interaction with the news content.

3.1 Like Statistics

In the sample data, the average number of likes per piece is 190, with a median of 89. The most-liked piece received 3,587 likes, while the least-liked received only 1. The top ten pieces by like count are shown in Table 1 .

Table 1: Top Ten Data Journalism Pieces by Like Count

Publication Date	Title	Likes
2020/2/5	763 Confirmed Patient Stories: Tracing the National Spread Path of COVID-19	3,587

Publication Date	Title	Likes
2020/5/27	A Mind Map to Understand the Civil Code: What Rights Does It Protect for You and Me?	2,456
2019/10/13	Data on Marathon “Breaking 2” : His 42-Kilometer Speed May Be Faster Than Your Cycling	1,890
2019/10/6	AR: China’ s Urbanization Process Over 70 Years	1,654
2020/3/1	National Discharge Rate Exceeds 50%: This Recovery Map Has Been Turning Greener Over the Past Month	1,432
2020/4/7	Wuhan Lockdown Countdown: To Walk Out of the Psychological Haze	1,287
2020/3/18	“Global Chinese Shops Closed, Can’ t Go Home” ? How Fake News Is Mass-Produced	1,156
2020/2/18	COVID-19 Infection Rate on “Diamond Princess” Reaches 14.6%	1,089

Publication Date	Title	Likes
2020/4/2	Asymptomatic Cases Included in Epidemic Reports: How Far Has Our Understanding of This Group Progressed?	987
2019/9/29	Women' s Volleyball Team Wins 10 Consecutive World Cup Victories: Data on 70 Years of Glorious Achievements	876

Among these top ten pieces, six (numbers 1, 5, 6, 7, 8, and 9) focus on COVID-19 topics, indicating that during the pandemic, readers showed heightened interest in epidemic-related data journalism, demonstrating that social hot topics generate better communication effects. Additionally, all top ten pieces were published in 2019 and 2020, with seven from 2020 alone, suggesting that The Paper' s data journalism column has gradually gained reader attention and recognition in recent years, with an increasing user base.

Plotting like counts by publication date reveals a year-over-year upward trend, indicating growing readership. Details are shown in Figure 1 [Figure 1: see original paper].

3.2 Comment Statistics

In the sample data, the average number of comments per piece is 99, with a median of 31. The most-commented piece received 4,590 comments, while 26 pieces received no comments. The top ten pieces by comment count are shown in Table 2 .

Table 2: Top Ten Data Journalism Pieces by Comment Count

Publication Date	Title	Comments
2019/3/5	Southern China Rainy Weather Competition: 60 Years of Data Shows This Is the Longest Sun “Wandering”	4,590
2018/8/23	One Graphic to Understand Shouguang Flood: Upstream Discharge Causes Downstream Disaster	3,876
2020/1/24	One Graphic to Understand: What Is the Level-One Response Activated by Hubei and 31 Other Provinces?	3,234
2019/1/19	Data Investigation on Infinitus: How Did It Become China’s Largest Direct Sales Company?	2,890
2019/9/10	Behind 53,027 Messages: The Self-Rescue and Mutual Aid of the Desperate in Internet “Tree Holes”	2,567

Publication Date	Title	Comments
2016/10/11	Knowledge Boost: After Years of CET-4/6 Exams, Do You Know How Much More You' ve Paid Than Others?	2,345
2020/7/22	Data Talk: Xu Mingchao Refuses to Apologize to Yamy, How Many Have Experienced Workplace PUA?	2,123
2017/12/11	Graphic Analysis of Jiang Ge Case "Rashomon" : What Are the Differences Between Prosecution and Defense Accounts?	1,987
2021/2/23	Calculation: The Value of Housework That Should No Longer Be Ignored	1,765
2017/7/21	Data on High Temperature : 65 Years of Data Shows How the New "Top Ten Furnaces" Were Forged	1,654

Among these top ten pieces, three (numbers 1, 2, and 10) address disaster weather such as floods and high temperatures, while number 3 covers COVID-19. Most others focus on negative news topics, indicating that socially negative topics more effectively stimulate discussion and are more likely to become hot or viral topics. Attention generates power, and observation changes the world. Reader attention and discussion of negative news often help events develop in positive directions and even drive social progress, enabling negative news to exert positive influence—demonstrating the value of data journalism. The comment top ten includes pieces from 2021 (1), 2020 (2), 2019 (3), 2018 (1), 2017 (2), and 2016 (1), showing no clear temporal pattern.

Plotting comment counts by publication date reveals no obvious temporal trend compared to like counts. Details are shown in Figure 2 [Figure 2: see original paper].

3.3 Analysis of Likes and Comments Correlation

To objectively analyze the correlation between likes and comments, this study calculated their Pearson correlation coefficient. In statistics, the Pearson correlation coefficient [4] measures linear correlation between two variables X and Y, with values ranging from -1 to 1. It is defined as the covariance of the two variables divided by the product of their standard deviations, with the formula:

The larger the absolute value of the correlation coefficient, the stronger the correlation. Coefficients approaching 1 or -1 indicate strong correlation, while those approaching 0 indicate weak correlation. Typical ranges for interpreting correlation strength are: 0.8-1.0 (very strong), 0.6-0.8 (strong), 0.4-0.6 (moderate), 0.2-0.4 (weak), and 0.0-0.2 (very weak or no correlation).

The calculated Pearson coefficient between likes and comments is 0.2, indicating weak correlation. Both metrics reflect reader recognition and interest, hence the positive correlation—more likes generally correspond to more comments. However, the correlation is not strong, and their temporal trends do not align perfectly. Each reader can only like once but can comment multiple times and engage with other comments. The increasing readership in recent years has driven the upward trend in likes. Comment counts more directly reflect article popularity. When a piece becomes a hot topic, its comment count can rapidly exceed its like count, multiplying communication effects. Consequently, the comment count timeline shows dispersed peaks rather than a clear trend.

4. Reflections on Data Journalism Practice and Development

Analysis of The Paper' s "Meishu Ke" column reveals that even during the initial user acquisition phase, high-quality data journalism can generate comment peaks and significant communication effects. This demonstrates that beyond focusing on publication volume, quality should be prioritized. Producing more hot-topic data journalism can increase likes, comments, shares, and reposts,

triggering exponential growth in communication effects and rapidly building column recognition.

4.1 Topic Selection Catering to Reader Preferences Effective topic selection is the first step in data journalism production. Topics should align with reader preferences to generate interest. Data journalism' s distinctive feature is its foundation in data mining and analysis, offering objectivity and logical rigor that traditional journalism cannot replace. It can analyze and elaborate on news events and topics from novel data perspectives. Therefore, topic selection should incorporate trending social issues or news events, potentially using like and comment data to analyze reader preferences and identify topics suitable for data-driven interpretation.

4.2 Innovating Visualization Formats The Paper' s "Meishu Ke" column predominantly uses simple, fast static infographics, with relatively few video or interactive formats. Novel, rich, and multi-dimensional visualization formats can effectively enhance user experience and represent a key advantage of data journalism over traditional text reporting. Given that non-static formats require more production time, we recommend balancing timeliness with format diversity while strengthening technical capabilities to improve production efficiency and more fully demonstrate data journalism' s unique strengths.

Data journalism represents the convergence of journalism and technology. In the big data era, massive datasets increase the difficulty of extracting valuable information. Data journalism helps readers uncover hidden information, filter and organize valuable components, and present them through diverse formats. Data' s objectivity enables deeper reader understanding of news events—something traditional text reporting cannot convey.

Many Chinese media outlets have accumulated years of practical experience in data journalism production. Through data scraping and analysis of The Paper' s "Meishu Ke" column, this study examines data journalism practices. Based on our findings, we recommend that data journalism production should emphasize topic selection, enhance technical capabilities, and experiment with innovative multi-dimensional visualization formats to further develop data journalism and better serve readers [5].

References: [1] Fang Jie. Introduction to Data Journalism [M]. Beijing: China Renmin University Press, 2019, 2.
[2] Wang Qiong, Xu Yuan, et al. China Data Journalism Development Report (2018-2019) [M]. Beijing: Social Sciences Academic Press, 2020, 1.
[3] Zhan Guihua. Practical Analysis of Data Journalism Reporting on Epidemics—A Case Study of Caixin' s "Digital Talk" [J]. China Media Technology, 2020(9):99-101.
[4] Peng Hai. Application of Pearson Correlation Coefficient in Medical Signal Correlation Measurement [J]. Electronics World, 2017(7):163.

[5] Zhang Chao. Unleashing the Power of Data—Research on Data Journalism Production and Ethics [M]. Beijing: China Renmin University Press, 2020, 1.

Author Bio: Guo Yan (1988-), female, from Shandong, engineer. Research interests: data analysis and management.

(Editor: Zhang Xiaojing)

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