Data-Driven or Data Illusion: A Critical Inquiry of Data Journalism Practitioners, Practices and Problems in China

Zhang, Shuling

*Zhang, Shuling, School of International Studies, University of International Business and Economics (UIBE).
ORCID:https://orcid.org/0000-0002-3553-7276*

***Corresponding Author: *Zhang, Shuling,* *Zhang, Shuling, School of International Studies, University of International Business and Economics (UIBE)*

Abstract: *This study investigates occurrences of data journalism in Chinese media in order to build a profile of data journalism practitioners, examine their practices and understand the main problems that pose constraints to the development of this evolving form of journalism in China. A combined study of in-depth interviews of Chinese practitioners and a content analysis of 268 samples show that while data journalism is being enthusiastically practiced as an innovation in Chinese newsrooms, it is only at a cursory level with limited scope of coverage and depth of analysis. Despite the hope of using data to tell more compelling stories and avoid news censorship of the government, the absence of an open data culture in China results in limited access to public data and the scarcity of quality data. This, along with a low level of data literacy among practitioners and doubts about its sustainability, all raise questions about the further development of data journalism in China. This new form of quantitatively oriented journalism is still in its infancy and is not comprised of the mainstream journalism practices in China. Pioneering interest coexists with skepticism and uncertainties about its development path.*

Keywords: *big data; Chinese practitioner; content analysis; semi-structured interview; data-driven journalism; journalistic practice*

1. INTRODUCTION

In a big data era that is awash with the exponential increase and availability of data (Manyika, et al. 2011), there have been major advances in information gathering, sense-making, storytelling and the dissemination of news. News consumers are demanding more interactive, engaging and effective ways for media outlets to provide information. Data journalism, or often referred to as data-driven journalism, has gained momentum with the rise of the global open data movement, increased availability of open-source software, cheaper data storage costs, and a need for news organizations to adapt to the changing market landscape. Following early experimentation in U.S. and European newsrooms and bolstered by fellowship programs, boot camps and cross-border data-intensive investigations, developing countries have started taking baby steps toward data journalism (Constantaras 2015). An awareness of using data to improve accountability and transparency is driving the rapid adoption of data journalism practices, although these newsrooms still lag behind much of the world (Oberai 2015).

With the rise of the "Big Data" concept, Chinese journalism practitioners have started their own experiments with this new strand of quantitatively oriented journalism. China's internet portal giant Sohu.com took the lead and kick-started its first data journalism blog Matrix in May 2011. The following year was widely regarded as a landmark for the development of big data projects with an increasing prevalence for data journalism. Many of these outlets publish data stories on a regular basis and in ways that are "at least as compelling as their prestigious western counterparts", which puts the country "at the brink of something that looks genuinely exciting," as the former data editor at the Guardian's Datablog, Simon Rogers, once put it.

2. THE DATA TURN IN THE JOURNALISTIC FIELD

This emerging journalistic practice has aroused research interest in various disciplines and a large body of literature has been published in recent years (Ausserhofer, et al. 2017). Regarded by many as

anything but new, data journalism evolved from computer-assisted reporting (CAR) (e.g., Appelgren and Nygren 2014; Coddington 2015;Fink and Anderson 2016; Splendore 2016; Uskali and Kuutti 2015), which started making significant inroads into western newsrooms in the 1990s, and from social science-based methods descended from precision journalism (Meyer 1973). This data-driven strain of journalism is rooted in the deeply democratic traditions of open-government advocacy (Coddington 2015; Parasie and Dagiral 2013), an open-source culture and the watchdog role of investigative journalism (Coddington 2015; Felle 2015). This has given rise to epistemological challenges (Lewis and Westlund 2015; Parasie and Dagiral 2013) to traditional journalism, which is reshaping fundamental aspects of news production and distribution (Lewis 2015).

Due to the confusion over what big data means (e.g., Boyd, 2012; Crawford, Miltner and Gray 2014; Lewis and Westlund 2015), there have been divergent views among professionals and scholars regarding the defined boundaries of data journalism (Coddington 2015;Fang and Gao 2015; Knight 2015;Splendore 2016; Wang, et al. 2016). Professional definitions have tended to be succinct but generalized, characterizing it as journalism done with data or "based on large data sets" (Rogers 2011), or any activity "that deals with data in conjunction with journalistic reporting and editing or toward journalistic ends" (Coddington 2015). However, researchers argue that this has been an oversimplification (Uskali and Kuutti 2015) and it remains paramount to clarify terms as a foundation for coherent scholarship (Coddington 2015; Lewis 2015; Fang and Gao, 2015).

Considerable efforts, therefore, have been taken into developing a typology, and differentiating data journalism from other quantitative areas of journalism (e.g., CAR, precision, computational and programmer journalism) while examining their points of overlap and divergence(Parasie and Dagiral 2013; Stavelin 2013; Gynnild 2014;Coddington 2015; Splendore 2016). Although the scholarly definitional approaches are often contradictory (Loosen, Reimer and Schmidt 2015), the hybrid and cross-disciplinary nature of data journalism are widely recognized as "encompassing statistical analysis, computer science, visualization, web design and reporting"(see Coddington 2015; Shen and Luo 2016; Guo 2013), with the use of open-source tools and open data as defining elements (Gynnild 2014).

Although the practice of data journalism varies across countries and regions, the working methods and tools are alike. It is claimed to be based upon "the application of data science to journalism from which knowledge is distracted" (see Shen and Luo 2016) and relies ona growing variety of computer skills, tools, and techniques ranging from traditional CAR techniques to the most advanced data visualization and news applications(Gray, Bounegru and Chambers 2011; Tabary, Provost and Trottier 2015). As "both a process and a product" (Ausserhofer, et al. 2017), itis operated in such a way that data forms the foundation for meaningful analysis, interpretation and storytelling, as can be seen from core practices of 1) gathering and validating data, 2) applying statistics to tease meaning from it, 3) presenting findings in interactive and engaging ways, and 4) providing transparent access to the raw data, the methodology and code used in the analysis (Howard 2014).The value tothe public accumulates as it distills the essence of complex issues into understandable and impactful news stories mostly based upon quantitative data analysis and visualizations(Lorenz 2010)

Early-stage academic research turned out to be descriptive examinations with particular focus on professional culture of journalism, newsroom structures, the epistemologies of data journalists and key features of data-intensive outputs (Ausserhofer, et al. 2017). Many sought to emphasize the relative advantages of data journalism over other forms of journalism, claiming that this quantitative turn (Petre 2013) makes it possible to put together "the traditional 'nose for news' and ability to tell a compelling story with the sheer scale and range of digital information" (Gray, Bounegru and Chambers 2011) . What distinguishes data journalism is the pivotal role of data in "obtaining, reporting on, curating and publishing data in the public interest" (Stray 2011). It is believed to have reinforced the norms and rituals of pioneering legacy media groups while exhibiting a distinctive character (Borges-Rey, 2017). While scholars have examined well-recognized professional norms (Coddington 2015; Fang and Gao 2015), they haven't found agreed-upon standards that contribute to excellence in this domain (Young, Hermida and Fulda 2017).In general, conventional news values, routines and journalistic norms are carried on in newsroom practices(Tandoc Jr. and Oh 2017). The preoccupation of data journalism, therefore, centers on data collection, interpretation and presentation for journalistic purposes without compromising fundamental journalistic norms and values.

Assessing the role and value of data within journalism can be a challenge given the numerous and varied name tags for data-intensive journalistic activities and the absence of a clearly defined boundary (Lesage and Hackett 2014). Lewis and Westlund (2015) put forward four conceptual frameworks (e.g., epistemology, expertise, economics and ethics) to "make sense of, act upon and derive value from data" that pertains to journalistic purposes. Varied theoretial frameworks have been explored, including role theory (Weinacht and Spiller 2014), trading zones (Lewis and Usher 2014), the sociology of knowledge (Shen and Luo 2016) and academic perspectives range from data-intensive news production, data govenance, data literacy, journalism education and training, to the shift of discourse of power from social elites to grassroots (Zeng and Hou 2015) as well as role perceptions of media practitioners (Boyles and Mayer 2016). While some regard data journalism as a socio-discursive practice involving the discursive efforts of all the actors involved (De Maeyer, et al. 2015) or a rhetorical craft using Aristotle's concept of techne (Karlsen and Stavelin 2014), actor-network theory mostly serves as the dominant paradigm for understanding data journalism, yet with the drawback of lacking critical perspectives (Hammond 2017).

Overall, an instructive and practice-oriented approach is commonly applied to examine the mechanics of data journalism (see Bradshaw 2010; Knight 2015) and furnish a justifying analysis of its inevitability in newsrooms. Additional attempts are needed to explore its impact, value and desirability as media practice and assess the ways in which audience perceive it in terms of credibility, newsworthiness and engagement. There have also been calls for scholarly interrogation of data and its journalistic value and to apply quantitative methods and cross-disciplinary insights to this inquiry (Ausserhofer, et al. 2017; Lesage and Hackett 2014; Lewis 2015).

3. AN INQUIRY ABOUT OCCURENCES OF DATA JOURNALISM IN CHINA

Although academic efforts have increasingly been examining this new genre of quantitatively-oriented journalism (Borges-Rey 2016; Fink and Anderson 2016; Knight 2015; Oberai 2015; Tabary, Provost and Trottier 2015; Appelgren and Nygren 2014), the US and Scandinavian newsrooms are often the research focus of attention, forming a detailed yet fragmented publication landscape. Outside of the western scholarship, very little has been published in English (Ausserhofer, et al. 2017), likely because such practices are more established in English-speaking countries(Loosen, Reimer and Schmidt 2015). Thus, it would be constructive to conduct a cross-national investigation on the prevalence of data journalism, open data and citizen involvement in countries like China, where these practices are new and subject to experimentation (Appelgren and Nygren 2014).

China, well-known for obfuscation at the government level, is nevertheless seeing an increased demand for open data. Both Chinese citizens and senior leadership are starting to call for greater access to public information (Ross 2015). Although the thirst for data-driven storytelling is pushing innovative transformations in Chinese newsrooms (Chen 2015; Peng 2013), the hard fact is that the Chinese government has opened its data sources quite slowly despite the directive to do so (see Liu, Mo and Fan 2016). As shown by the Global Open Data Index1, China's overall index ranking fell to No. 93 in 2015 among the 122 sample countries, with a calculated 18 percent of data sets published by the government. Additionally, investigative reporting tradition in China is far from being deep-rooted (Zhang and Shen 2011) and is still struggling to survive (Chan 2016) as the government steps up its control of traditional and digital media. Moreover, Chinese journalists' adherence to precision journalism norms lagged behind their US counterparts (Shen and Luo 2016). All of these calls into question the way data is used in the practice of data journalism in China.

This study seeks to identify major data journalism practitioners in China and examine their contemporary explorations in this emerging journalistic field. While paying special attention to the way data sits in the workflow of data journalism production, this study intends to examine its adherence to shared professional logic and norms, as well as the major constraints faced by Chinese practitioners. It hopes to shed some light on any possible national characteristics of data journalism as it diffuses into media practice in China. Inspired particularly by Seth C. Lewis' (2015) three analytical lenses to trace the broad outline of journalism in an era of big data, i.e., empirical, conceptual and critical perspectives, this study seeks to critically examine how data is incorporated into journalistic practices in China and to answer the following questions:

RQ1: Who is currently producing "data journalism" in China?

RQ2: In what ways does data sit in the workflows of Chinese practitioners, particularlyin their core practices of *acquiring, analyzing, presenting and publishing data*?

RQ3: What are the main problems perceived by practitioners in their experiments with data journalism in China?

4. METHODS

To get a thorough overview of data journalism experiments in China, this study draws on a combination of qualitative research interviews to seek answers for the first and third research questions, and quantitative content analysis for the second. More specifically, the empirical basis of this study is, in part, composed of in-depth interviews with Chinese data journalism practitioners about their background and incentives to enter this field, their work methods and main problems they have encountered. It is further based on a content analysis of 268 data journalism products of selected Chinese media outlets in a year-long period from March 2016 to March 2017.

4.1. Interviews

Semi-structured interviews were conducted among data journalism practitioners working in various media outlets in China. According to Spradley's (1979) description of enculturation, or "the natural process of learning a particular culture", interviewees with an advanced level of enculturation were selected from: 1) pioneering media organizations with awarding-winning data journalism projects, e.g., Caixin Data Visualization Lab2 and CCTV; 2) media organizations with dedicated data sections on their digital platforms like Xinhuanet.com and three of the four major Chinese internet portals – Sohu.com, Sina.com and 163.com; and 3) independent business entities that have a clear mission of enthusiastically promoting data journalism development in China, such as Dyclub Data Media Lab and Data Works .

Altogether, 20 representatives from 18 Chinese media outlets agreed to be interviewed. The geographic segmentation consisted of organizations based in Beijing, Shanghai or Guangzhou, the three largest cities in China, as few newsrooms in smaller cities practice data journalism. Although the number of selected data journalism practitioners is not extensive, all the interviewees are early adopters and enthusiastic pursuers of such practices in China. Most started their experiments with data-driven practice in 2011 and have remained active in relevant social media communities since then. They are, therefore, well-positioned to provide detailed information about the work practices and main constraints that they have encountered as pioneering Chinese practitioners.

As the interviewees mentioned the competitive pressures of their media outlets and proprietary approaches to data-driven news work, all of them were granted anonymity. Among them were seven data journalists or journalists working with data (coded as DJ1-DJ7), five data editors (DE1-DE5) and eight news product managers (PM1-PM8). Four interviewees also trained other journalists in their own or other newsrooms, in addition to working as product managers and data editors. In-depth interviews were conducted between March—May 2017 by telephone or on Wechat4, a Chinese social media platform on which video and voice calls can be made.

The interview guide consisted of four relatively open questions in different aspects of data journalism practice, leading to more follow-up questions. Interviewees were first asked to briefly describe their current position, educational background and team compositions (for example, whether the team is stable, and how many people worked on each project). Then they were asked about the reasons why they decided to enter this journalistic field. Follow-up questions were raised related to their work dynamics of using data to produce stories. Then they were asked about the problems or obstacles in their professional practices. The interviews lasted between 30 to 60 minutes, all of which were recorded, transcribed and arranged into general themes upon further analysis.

4.2. Content Analysis

In order to find more empirical evidence of the data-driven news production in Chinese media outlets, this study also examined the output of their data journalism experiments. A pragmatic and inductive approach was adopted since data journalism is still a diffuse and evolving term which adds to

difficulties of identifying related news items for content analysis (Loosen, Reimer and Schmidt 2015). Six of the above-mentioned Chinese media that fall within the purview of interview studies were selected, and pieces that were published within dedicated data sections were chosen with a constructed-week sampling method.

4.3. Samples

The six Chinese online media – Xinhua.net, the Beijing News, The Paper.cn, Caixin.com, Sohu.com and People.com were selected because all of them: 1) are early adopters exploring data-driven journalism in China; 2) have set up data sections on digital platforms and produce data stories on a regular basis; and, 3) represent different types of media organizations in China, ranging from traditional print publications to online media and internet portal companies. Relevant sections have been grouped into data journalism categories on their websites, such as the Matrix section on Sohu.com and the Data News section on Caixin.com. Data stories presented in these sections were selected during a six-constructed-week period, Monday through Sunday during the above-mentioned year-long period.

This early stage of sample compilation resulted in a corpus of 290 products during the investigated period. However, not all the products presented under a broad umbrella of data blog or interactive column would qualify for this study as some didn't produce journalistic content or didn't serve journalistic purpose. Individual projects posted on personal blogs were also excluded from the study, as it is unclear if they adhere to professional norms as in the field of data journalism. In total, 268 products were included in the study sample.

4.4. Variables

To examine the role of data in journalism, clear benchmarks are required to properly categorize and measure samples. Guided by the categorization of data journalism stream, form and content (Uskali and Kuutti 2015; Knight 2015), and based upon the work flows outlined by Lorenz (2010), this study examined the following seven variables during the work process of acquiring, analyzing, presenting and publishing data:

Variable 1: Subject

The selected data-driven stories were categorized by the subjects they covered: 1) politics; 2) finance and business; 3) social issues, including education, housing, health and the environment; 4) livelihood, covering culture, entertainment, tourism, sports and stories relevant to lifestyle; 5) science and military; 6) worldwide issues or foreign affairs.

Variable 2: Stream

Two main streams of data journalism were examined in the sampling projects: 1) investigative data journalism; 2) general data journalism. This analysis aimed to find the proportion of these streams practiced by Chinese media, and uncover evidence of whether Chinese data journalism is coupled with investigative reporting, like with many western counterparts.

Variable 3: Amount of data sources

Given the claim that "data journalism is all about diverse sources" (Rogers 2011), this study sought to examine the amount of sources used in data-driven news production in Chinese media outlets. Sampling products were put into three groups: 0) Unsourced; 1) One source; 2) Multi-source,as journalists collate pieces of information from more than one source.

Variable 4: Identification of data sources

The origin of data in each story was also examined. Generic data sources were identified as: 0) N. A., with no attributed data source; 1) government; 2) non-government institutions, i.e. business enterprises or research institutions; 3) in-house data collected by, or archived within the media group; or 4) mixed, in case where more than one type of data sources are included.

Variable 5: Sophistication of analysis

This measured different levels of information processing and data analysis, a requirement for data journalism. The following criteria were set to determine the complexity levels of data analysis: 1) basic—simply presenting the raw data; 2) intermediate—applying descriptive statistical analysis to provide quantitative descriptions of distribution, central tendency and dispersion; and 3) advanced usage of inferential statistical analysis involving large data sets, data mining and programming skills.

Variable 6: Modes of visualization

Four modes of data visualization were assessed in an ascending hierarchical order: 0) N/A, i.e., no exhibition of visual element; 1) text-heavy, only using tables, lists, charts or graphs; 2) static visualization such as maps, timelines, and infographics; and 3) interactive representation where dynamic maps, videos, interactive timelines, and news applications are designed for greater visual appeal and user engagement.

Variable 7: Data transparency

This examined whether Chinese practitioners make their data sets public in the analysis and invite readers to participate in the production of data-driven stories. Each story was flagged either 0) absence, or 1) presence of direct links or original data sets shown or published on the web page.

Intercoder Reliability

Two coders analyzed the samples during one week in April 2017. Training was initially carried out using data stories produced by both prestigious western newsrooms and Chinese media outlets as examples. Each coder first worked separately to make independent judgment. Related discussion and an updated coding manual helped resolve their disagreements before a retest of reliability was conducted. This iterative process last until the intercoder reliability reached an acceptable level of 85 to 92 percent based on the Hosti reliability formula, and the computer program SPSS was used for data analysis.

5. FINDINGS

Who is producing data journalism in China? In-depth interviews with data journalism practitioners in China echoed our proposition that this new genre of journalism has already gained enthusiastic followers in Chinese media outlets. A chronological order of major producers in China showed that the pioneers have come from China's internet portals Sohu, Netease, Tencent, and Sina News. Most of them launched their first data journalism projects before the first half of 20123, a landmark year when big data became a new buzzword in the country. The reasons are two-fold. Firstly, China's online news websites are not authorized to produce original news content, forcing them to quote traditional media and most of the times, the official Xinhua News Agency, which inevitably lead to homogenization of news content. To attract larger audiences and grow traffic, online media organizations spared no efforts in exploring innovative and differentiated ways of news aggregation and reproduction. Consequently, data journalism was adopted as an initiative to use publicly available data to produce original news content. Secondly, online media have shown peculiar strengths in data mining and processing, a deeper understanding of the internet, and a more open mind toward innovation, all of which made it possible for them to become early adopters of data journalism in China.

In contrast, it was not until 2013 when Chinese traditional media followed suits, asthePeople.com, Caixin.com, the online websites of the People's Daily and Caixin Magazine, took their initial steps towards data journalism practice. Those resource-rich traditional media, though not first movers, quickly utilized their strengths in resource integration and channel distribution. As a program producer of the national television broadcaster CCTV put it,

> Data journalism is not used as the primary form in our news reporting. Our production of data projects are all project-based. We didn't tap into this new kind of data-driven storytelling method until 2014. But our debut was the successful launch of a series of big data-based Spring Festival specialprograms5 using user migration data from China's search engine Baidu and e-commerce giant Alibaba. (Original translation)

In terms of age composition, only two of the interviewees are above 45 years of age whereas 11 are under 30 with less than five years of work experience as data journalists or editors. Fourteen interviewees have journalism as their primary expertise, among which three hold data journalism degrees. Seven of them entered this field as it was deemed to be "the future of journalism" and suits the overall industry trend in the big data era. Five mentioned the possibility of avoiding trouble with government censorship by digging deeper into open data and using increasingly available tools to interpret the issues that Chinese audiences care about. Other reasons include the newsroom pressures to generate more in-depth news stories and visually appealing graphics as well as their personal interests in quantification and data analysis.

Conditions vary in terms of data team size and composition. It was found that a data team of less than ten people was commonly seen among Chinese media outlets, comprising over half of the total studied teams (M=8.31, SD=5.76, n=17). More specifically, six teams are comprised of five to ten people while another five reported a team of 11 to 20. Notably, online media outlets such as the Paper.cn and DT Finance both have a stable data team of over ten members, an indication that the production of data-driven news is personnel-intensive in Chinese newsrooms. The largest data team thus far in China is from Xinhuanet.com, with a dedicated data team of 18 to 20 people. CCTV once formed a tentative production team of over 80 people to work on the TV documentary of Telling the Story of Common Destiny with Big Data, but the team was dismissed once the project was finished.

When asked to describe the composition of their teams, nearly all interviewees talked of data journalist, data editor, graphic designer and product manager, without mentioning data analyst as a regular position in their respective teams. Ten interviewees said data scraping, cleaning, and analysis are mostly undertaken by data journalists or editors themselves who have expertise in more than one field. But sometimes they hire outside programmers or experts if a problem is beyond their ability or is overly time-consuming. The two interviewees from CCTV and the project leader of Renmin.com mentioned they tend to outsource the work of data analysis to data mining companies or collaborate with research institutions. The blurring usage of professional titles has also been brought up by Chinese practitioners. Among the most frequently used are data journalist, data editor, news application developer, programmer journalist, and project manager, which once more raises the questions of conceptualization and role perception among practitioners. Undervarious professional titles, those who are engaged in the collaboration process are pondering their roles and uniqueness within newsrooms.

In what way does data sit in the work process of data journalism? Of the total 268 data journalism outputs that constituted the study samples, 23 per cent were selected from Xinhua.net, the website of China's official newswire Xinhua News Agency; 15 per cent were chosen from the digital platform of People's Daily, a national newspaper with an extensive readership, and 20 percent were from The Beijing News, a municipal daily newspaper based in Beijing. Another 18 per cent were from Caixin.com, the online version of an influential Chinese weekly magazine on finance and business. The remaining 24 per cent were chosen from China's leading internet portal Sohu.com and a rising online media outlet, the Paper.cn.

Table1. *Distribution of data journalism product samples by media outlet*

		Frequency	**Percent**	**Cumulative Percent**
Valid	The Beijing News	54	20.1	20.1
	Caixin	49	18.3	38.4
	People's Daily	40	14.9	53.4
	Sohu.com	31	11.6	64.9
	the Paper.cn	32	11.9	76.9
	Xinhua	62	23.1	100.0
	Total	268	100.0	

The sampling data stories were first categorized by the main subjects they covered. Results show that domestic news accounts for a vast majority of nearly 90 percent in all samples, suggesting a local or national proximity in topic selection. Stories covering social issues (e.g., education, public housing, healthcare and environmental issues) constitute about 40 percent of total samples, followed by politics (19.8 per cent), livelihood (18.7 per cent), and finance & business (11.2 per cent) coverage. Few data news items (1.1 per cent) covered science and military, presumably due to lack of data accessibility and the difficulties of illustrating such complicated topics.

A further breakdown of the topic distribution among media groups reveals differences in their editorial selection. The online platforms of traditional media such as People's Daily and Xinhua both show a clear tendency of covering straight data news such as political and social stories, whereas online news media such as thePaper.cn and Sohu News tend to use data to produce soft news that has close proximity to the general readership. The Beijing News also put emphasis on social news coverage while Caixin devoted a greater effort in using data to cover financial and business topics, following its self-portrait as a financial and business publication. This reaffirms the scholarly claim that data journalism retains an emphasis on editorial selection and conventional judgment of news values (see Coddington 2015) in their data-driven news-making process.

Each product was then classified into one of the two main streams of data journalism suggested by Uskali and Kuutti (2015). Examination of their links to the ethos of investigative reporting showed that general data journalism projects represented an overwhelming majority (93.7 per cent) in sampling products, while investigative projects were rarely seen, making up a mere 6.3 percent of the entire samples. This indicates that data journalism experiments in China are far less coupled with investigative reporting than in western newsrooms, where it is often used within the context of investigative projects. Such a disproportionately uneven distribution is largely due to the fact that investigative data journalism is often time-consuming and requires extensive resources to perform advanced data skills. Moreover, unofficial and confidential information and data leaks often play a significant role in starting and directing investigations, making it hard to pursue such stories for Chinese media outlets struggling with government censorship and facing pressure to meet deadlines and realize profitability.

Table2. *Coding results of sample content analysis*

	Frequency (n=268)	Percent (%)	Cumulative Percent (%)
Subject			
Politics	53	19.8	19.8
Finance & Business	30	11.2	31.0
Social Issues	107	39.9	70.9
Livelihood	50	18.7	89.6
Military & Science	3	1.1	90.7
World	25	9.3	100.0
Stream			
Investigative datajournalism	17	6.3	6.3
General data journalism	251	93.7	100.0
Amount of data sources			
Unsourced	12	4.5	4.5
One source	75	28.0	32.5
Multi-source	181	67.5	100.0
Identification of data sources			
N/A	10	3.7	3.7
Government	20	7.5	11.2
Non-official institution	122	45.5	56.7
In-house	27	10.1	66.8
Mixed	89	33.2	100.0
Modes of visualization			
N/A	5	1.9	1.9
Text-heavy	25	9.3	11.2
Static representation	210	78.4	89.6
Interactive visualization	28	10.4	100.0
Data transparency			
Absence of raw data sets	259	96.6	100.0
Presence of raw data sets	9	3.4	

Results showed nearly two thirds of data-driven projects were multi-sourced, as can be seen from journalists collating pieces of information from more than one source, while approximately one third of data-driven projects quoted only one source. Only less than five percent of studied samples were where no source was identified. It could be interpreted as an increasing awareness of, and adherence to the professional benchmarks of attributing data sources and using diverse sourcing to confirm, crosscheck and validate the data.

The identification of data sources in each sample was then examined. Of all the generic data sources measured in this study, there was a clear reliance on available open data. Specifically, among the 258 projects that attributed their data source, about half were sourced via non-government institutions like business entities (for example BAT companies6 with abundant user data) and research institutions, both at domestic and international levels. About seven percent quoted public documents released by government agencies while nearly one third used mixed sources. A further breakdown of projects with mixed sources showed that 65 items included open data from government agencies. That makes a total of 85 projects, or over one third of the studied samples that has sourced via government agency at least once. Notably, in-house data, or the data gathered by the media organizations through traditional interview techniques, investigations or construction of their own database, only accounts for a small slice of 10 percent.

Further examination of the data sources by subject revealed that stories covering social and economic issues showed a heavy reliance on both official and public releases. A large portion of business (78 per cent) and social stories (60 percent) sourced their data through government agencies or research institutions. In nearly half of those instances, the data was from survey findings or research reports issued by those institutions and used without further comment or challenge by the journalist. This doesn't fit with the claim that data journalism provides an optional approach to "break away from the dominance of official sources and public releases" (Knight 2015).

Evaluating sophistication levels of data analysis in sampling news items caused the most difficulty. As a result, data specialists were consulted to understand data structures, software tools, and analytical methods. About one quarter (28 percent) of total samples remained at a basic level of simply presenting data or comparing values by calculation of the average or percentage. More than half (60 percent) used descriptive statistical skills to help readers and users understand the features of a specific data set. Only one sixth of total projects applied advanced level of data mining or inferential statistics in their analysis of data sets. One possible reason might be the absence of data analysts as regular team members or shortage of hands within newsrooms, making it hard for sophisticated data mining and statistical analysis. In addition, the data sets made public by institutions are often pre-processed and aggregate, leaving limited access to raw data of journalistic interest.

This study also examined Chinese practitioners' visualization practices based on the visual element and interactive feature shown in each sample. Even though data-driven pieces often use several elements, this study only focused on the most prevalent element with an aim of identifying the predominant visualization approach adopted by Chinese practitioners. It is found that some products were labeled in relevant data sections despite the fact that no "visual" elements or merely conventional "textual analysis6"wasdeployed. This type of production constitutes one tenth of the entire sample. A majority of projects used graphic elements to enhance visual appeal, of which nearly 80 per cent used static maps, timelines or static infographics. A further investigation into the projects that featured static representation found that static infographics take up a notably major portion of 78 percent. A more advanced level of interactive visualization was also observed (including dynamic maps, interactive timelines, videos, and adding animations), only on a small scale(about 10 percent) as compared with static representations.

Additionally, this study set out to track whether the data sets used in sampling projects were published to encourage the readers to take a participatory role. Results showed a dominating proportion of projects neither provided direct links of original data sets nor explained the methodology and code used in the analysis. After taking a further look at the very few projects (n=9) that contained direct links to original sources, the methodology that led to their published conclusions were not available. Moreover, some data sources appear to be ambiguously attributed, making it hard to track and cross-check. The opacity of raw data sets and the reluctance to open up data-driven news production process suggest that participatory openness is still far from being the mainstream practice among Chinese practitioners.

6. MAIN CONSTRAINTS

Although the Open Source movement across the globe has provided vital opportunities for data journalism, in-depth interviews with Chinese practitioners revealed mixed professional views on its future, as limited access to quality data, low levels of data literacy among practitioners, and lack of business models for sustainability all hinder further development.

Limited access to quality data. The greatest challenge of practicing data journalism in China, as most interviewees in this study agreed, lies in the difficulty of obtaining data. According to them, gaining public data by submitting a request is rarely successful due to insufficient "national laws requiring the mandatory release of a certain number of documents." As a former data team leader of South Reviews pointed out, access to public information is largely dependent on the good intentions of political authority since "government officials can always cite a variety of exceptions to prevent documents being released or disclosed." (Original translation)

Interviewees also expressed great concern over the quality of data obtainable. When data sources are incomplete, unavailable, proprietary, or in formats that are not appropriate for computer analysis, journalists must spend extensive time scraping, wrangling and crosschecking data. Two interviewees from Xinhuanet.com and Renmin both mentioned the highest quality data comes from the National Bureau of Statistics while other sources are considered questionable. Particularly, data provided by business entities require great scrutiny, as "those companies normally don't provide an API to make it openly accessible to journalists." (Original translation) Even if some businesses choose to allow access to journalists, their motivation and data collection methods must be carefully investigated to maintain journalistic independence and objectivity. As the former team leader of Caixin.com and now a co-founder of Data Works mentioned, although the ascent of a "big data" era brings opportunities to journalism, the real power lies in the hands of business entities, particularly BAT7 companies with their own extensive user data.

> In China, so-called big data is mostly owned by large internet companies and e-commerce businesses that make use of big data to meet user demand and explore the business value of user data. From this perspective, media groups are only passive trend followers. (Original translation)

Low level of data literacy. Data literacy, defined as "the ability to process, sort and filter vast quantities of information" (Koltay, 2016), remains another barrier to widespread adoption of data journalism in China. Three interviewees disclosed difficulties in hiring qualified analysts that have both data mining and programming skills on top of a "nose for news," partly because the hiring cost is too high and partly because of an insufficient supply of such experts. A deeper cause lies in the current curriculum design of journalism majors in China. As one project manager of DT Finance said:

> It is clearly observed that most journalism graduates from Chinese universities are arts majors who are afraid of algorithms and computer science, and are reluctant to perform statistical analysis, let alone programming and coding. If they are incapable of poring over data, how can it be possible for them to find stories from data? That partly explains why data journalists that we hire are science majors who are better able to derive meaningful information from data. At least they won't panic when dealing with data. (Original translation)

The lack of data literacy further hinders critical thinking as journalists are unable to interrogate and verify data. And it may also pose ethical problems with regards to privacy, trust, and responsibility concerning data collection, analysis and dissemination. It also leads to technical inefficiencies for practitioners, as "there is only a small group of journalists in China who boast a hybrid command of data scraping, cleaning, analysis and visualization skills." Only a limited number of media groups maintain regular data teams, hire people with data expertise and provide regular on-the-job training opportunities. For most journalists, online courses or workshop trainings are the only ways to learn how to use the right tools to meet the growing demand for data-driven news production. Three interviewees described their frustrations and confusion upon entering this new area. One said she had signed up for some data journalism workshops but only found it to be a week-long course for starters, which neither suited her self-study needs nor provided a systematic learning opportunity afterwards. A project manager in a traditional newspaper expressed her concerns on the transitional cost of training traditional reporters to become data journalists as follows:

> To those reporters who have worked in traditional media for over three years, the cost of making a (professional) transition (to data journalism) is huge. For starters, he/she needs to get the approval of a department head to provide time and opportunity to take training or learn from data experts. He/she must learn in his/her spare time while performing routine tasks. It would take more than a year just to enter this field, and that only covers the basics of data acquisition, processing and visualization. That's why I seriously doubt it would be possible. (Original translation)

Doubt towards the value of data journalism. Nearly all interviewees expressed concerns about the cost of producing data-driven stories. Eleven of them emphasized that high-quality projects are time-consuming and require intensive human and capital resource as well as internal and external cooperation. Interviewees hold divided views on whether data journalism should follow a business model that generates revenue stream for media groups. Resource-rich traditional media can afford to maintain their data journalism initiatives. Four interviewees, all from traditional media, believe that data journalism is "merely a matter of journalistic value" and holds potential for impact rather than for revenue. Two interviewees, a producer at CCTV and a data project manager at the People.com, mentioned that producing high-profile, award-winning data-driven stories creates invisible value for media groups and is therefore also important in terms of accumulating the "cultural capital" of mainstream media.

By contrast, nearly half of the interviewees expressed rather pessimistic views, questioning the sustainability of data journalism if it cannot even cover production costs, let alone bring in revenue. Furthermore, they argue that data journalism is regarded by most Chinese as an innovative counter to declining advertising revenue and shrinking readership, but results are relatively lukewarm. As a product manager of DT Finance put it:

> I once tried scraping WeChat user data from the mobile platforms of several data journalism producers in China and found a homogeneous group of readership in which university students and data journalism practitioners take up a large majority. User reading time, scrolling and interaction levels are rather low. In a word, it is far from being the most read stories on the platforms in which they appear. (Original translation)

7. FUNDING

8. DISCUSSION AND CONCLUSION

This study validated a few simple assertions. Obvious discrepancies were found between actual practices of Chinese data journalism practitioners and shared professional norms in scholarly literature. Results from content analysis showed the experiments of data-driven newswork by Chinese media outlets were only on a cursory level in terms of scope and depth. Some frontiers seem porous, as some projects under the umbrella of "data stories" were heavily reliant on open data, engaged only on a limited scale with statistical analysis and interactivity, and showed little commitment to data openness and transparency. This can result in oversimplified and even ungrounded findings and conclusions. Such findings serve as a wakeup call for Chinese media outlets and data journalism practitioners alike.

Data journalism, while still in its infancy, does not comprise of the mainstream of journalism practice in China. There is enthusiastic fandom to implement this genre of journalism as an innovative initiative in Chinese newsrooms, but also skepticism among those who doubt the sustainability of data journalism or remain unfamiliar with its working methods. Broadening access of data for journalistic production and the enhancement of media practitioners' data literacy and the exploration of potential business models are vital to the seemingly opaque future of data journalism in China.

This study has not yet taken the Chinese journalism conventions and local constraints into account, which could have exerted influence on the way data journalism is practiced within Chinese newsrooms. Another limitation of this study is the concentration on data-driven news items published on the digital platforms of Chinese media outlets, without exploring the variety of this emerging form of journalism in traditional newspaper, broadcast or mobile devices. Along those lines, an

ethnographic approach into Chinese newsrooms and an empirical analysis of how well data stories are perceived and received among massive internet and mobile users, could better illustrate the integration of data journalism practices into different types of media outlets in China.

Notes

1. The Global Open Data Index is not an official government representation of the open data offering in each country, but an independent assessment from a citizen's perspective. It is a civil society audit of open data. Available at http://2015.index.okfn.org/place/china/.

2. Caixin Data Visualization Lab was established in Feb. 2012 and has since then produced award-winning data-driven reports, including Report onQingdao Sinopec oil pipeline explosion, Colored pattern of Nobel Prize winners and visualized story-telling of Zhou family ties and relationships.

3. By June 2012, the "Big Four" Chinese internet portals, e.g. Sohu.com, Netnease.com, Tecent and Sina.com had all established dedicated data journalism sections.

4. WeChat, a dominant messaging service in China, is owned by Chinese internet giant Tencent with 700 million monthly active users. People can use this interactive social app to text, voice call and video chat with each other.

5. The official media CCTV invested over 7 million yuan in 2015 and produced a series of data-driven documentaries China Common Destiny.

6. Based on Knight's (2015) taxonomy of data visualization.

7. An abbreviation of the three major Chinese internet and E-commerce companies, i.e., Baidu, Alibaba and Tencent.

REFERENCES

Anderson, Christopher. 2015."Between the Unique and the Pattern: Historical tensions in our understanding of quantitative journalism."Digital Journalism3(3): 349-363. doi: 10.1080/21670811.2014.976407.

Appelgren, Ester, and Gunnar Nygren. 2014."Data journalism in Sweden: Introducing new methods and genres of journalism into 'old' organizations."Digital Journalism2(3): 394-405.

Ausserhofer, Julian, Robert Gutounig, Michael Oppermann, Sarah Matiasek, and Eva Goldgruber. 2017. "The datafication of data journalism scholarship: Focal points, methods, and research propositions for the investigation of data-intensive newswork."Journalism.doi:10.1177/1464884917700667.

Borges-Rey, Eddy. 2017. "Towards an epistemology of data journalism in the devolved nations of

the United Kindom: Changes and continuities in materiality, performativity and reflexivity."Journalism. doi: 10.1177/1464884917693864.

Borges-Rey, Eddy. 2016."Unravelling Data Journalism: A study of data journalism practice in British newsrooms."Journalism Practice 10(7): 833-843.doi:10.1080/17512786.2016.1159921

Boyd, Danah, and Kate Crawford. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon."Information, Communication & Society 15(5) : 662-679.

Boyles, Jan Lauren, and Eric Mayer. 2016. "Letting the Data Speak: Role Perceptions of Data Journalists in Fostering Democratic Conversation. "Digital Journalism 4(7): 944-954.

Bradshaw, Paul. 2010. "How to be a data journalist."The Guardian data Journalism. https://www.theguardian. com/news/datablog/2010/oct/01/data-journalism-how-to-guide.

Chan, Ying. "China: Investigative Journalism Struggles but Survives."Uncovering Asia: The Second Investigative Journalism Conference. Accessed June 1, 2017. http://2016.uncoveringasia.org/2016/09/ 23/china-investigative-journalism-struggles-but-survives/.

Chen, Changfeng. 2015. "The Changes of Journalistic Practices in a Big Data Era."Young Journalist 16: 69-70.doi: 10.15997/j.cnki.qnjz.2015.16.047.

Coddington, Mark. 2015."Clarifying Journalism's Quantative Turn:A Typology for Evaluating Data Journalism, Computational Journalism, and Computer-assisted Reporting." Digital Journalism 3(3): 331-348.

Constantaras, Eva. 2015."Global Models for Starting Data Journalism in Developing Countries." In Data Journalism: Inside the Global Future, edited by Tom Felle, John Mair and Damian Radcliffe, 254-265. Suffolk: Arima Publishing.

Crawford, Kate, Kate Miltner, and Mary L. Gray. 2014."Critiquing Big Data: Politics, Ethics, Epistemology. "International Journal of Communication8: 1663-1672.

De Maeyer, Juliette, Manon Libert, David Domingo, Francois Heinderyckx, and Florence Le Cam. 2015. "Waiting for Data Journalism: A qualitative assessement of the anecdotal take-up of data journalism in French-speaking Belgium."Digital Journalism 3(3): 432-446.doi:10.1080/21670811.2014.976415.

Fang, Jie, and Lu Gao. 2015. "Data Journalism: An Area Urgently Need to be Regulated—A Quantitative Study Based on Five Domestic Data News Columns."Journal of International Communication 12: 105-124. doi: 10.13495/j.cnki.cjjc.2015.12.006.

Felle, Tom. 2015. "Accountability meets engagement: data journalism's watchdog role." In Data Journalism: Inside the Global Future, edited by Tom Felle, John Mair and Damian Radcliffe, 167-176. Suffolk: Arima Publishing.

Fink, Katherine, and Chris W. Anderson. 2016. "Data Journalism in the United States: Beyond the "Usual Suspects"."Journalism Studies 16(4): 467-481.

Gray, Jonathan, Liliana Bounegru, and Lucy Chambers. 2011. The Data Journalism Handbook: How Journalists Can Use Data to Improve the News.Sebastopol, CA: O'Reilly Media.

Guo, Xiaoke. 2013. "The Current Situation and Functions of Data Journalism as a Discipline."Editorial Friend8: 87-89. doi: 10.13786/j.cnki.cn14-1066/g2.2013.08.020.

Gynnild, Astrid. 2014."Journalism Innovation Leads to Innovation Journalism: The Impact of Computational Exploration on Changing Mindsets."Journalism 15(6): 713-730.

Hammond, Philip. 2017. "From computer-assisted to data-driven: Journalism and Big Data ."Journalism 18(4): 408-424. doi: 10.1177/1464884915620205.

Hewett, Jonathan. 2015. "Data journalism grows up." In Data Journalism: Inside the global future, edited by Tom Felle, John Mair and Damian Radcliffe, 27-38. Suffolk: Arima Publishing.

Howard, Alexander Benjamin. 2014."The Art and Science of Data-driven Journalism."New York: Tow Center for Digital Journalism, Columbia Unviersity. https://towcenter.org/the-art-and-science-of-data-driven-journalism/.

Karlson, Joakim, and Eirik Stavelin. 2014. Computational journalism in Norwegian newsrooms. Journalism Practice 8(1): 34-48.

Knight, Megan. 2015. "Data journalism in the UK: a preliminary analysis of form and content."Journal of Media Practice 16(1): 55-72.doi: 10.1080/14682753.2015.1015801.

Koltay, Tibor. 2016. "Data governance, data literacy and the management of data quality." International Library Associations and Institutions 42(4): 303-312. doi: 10.1177/0340035216672238.

Lesage, Frédérik, and Robert A. Hackett. 2014."Between Objectivity and Openness—The Mediality of Data for Journalism ."Media and communication 2(2): 42-54.

Lewis, Seth C. 2015."Journalism in an Era of Big Data: Cases, Concepts, and Critiques."Digital Journalism 3(3): 321-330.

Lewis, Seth C., and Oscar Westlund. 2015. "Big Data and Journalism: Epistemology, Expertise, Economics, and Ethics."Digital Journalism 3(3): 447-466.

Lewis, Seth C., and Usher N. 2014. "Code, collaboration, and the future of journalism: A Case Study of the hacks/hackers global network."Digital Journalism 2(3):383-393.

Liu, Haifang, Shihong Mo, and Bingbing Fan. 2016."The Latest Development and trends of open data movement."Journal of Intelligence9: 163-167.

Loosen, Wiebke, Julius Reimer, and Fenja Schmidt. 2015. "When Data Become News': an exploratory content analysis of data journalistic projects." In Procceddings of The Future of Journalism 2015 Conference: Risks, Threats and Opportunities, Cardiff, September 10-11.

Lorenz, Mirko. 2010. "Data-driven journalism: what is there to learn?"Paper presented at the data-driven journalism roundtable, Amsterdam: August 24.

Manyika, James, et al.2011.Big data: The next frontier for innovation, competition, and productivity.McKinsey Global Institute. http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation.

Meyer, Philip. 1973.Precision Journalism: A Reporter's Introduction to Social Science Methods. Bloomington , Indianapolis: Indiana University Press.

Oberai, Sanjit. "Baby steps: a slow start but data journalism is India is gathering pace." In Data Journalism: Inside the global future, edited by Tom Felle, John Mair and Damian Radcliffe, 266-272. Suffolk : Arima Publishing .

Parasie, Sylvain, and Eric Dagiral. 2013."Data-driven journalism and the public good: 'Computer-assisted reporters' and 'programmer-journalists' in Chicago."New Media & Society 15(6): 853-871. doi:10.1177/ 1461444812463345.

Peng, Lan. 2013."New challenges faced by journalism in an era of big data ."Editorial Friend, 1: 6-10.doi:10.13786/j.cnki.cn14-1066/g2.2013.01.012.

Petre, Caitlin. 2013. "A Quantitative Turn in Journalism?"New York:Tow Center for Digital Journalism, Columbia Unviersity.http://towcenter.org/a-quantitative-turn-in-journalism/.

Rogers, Simon. 2011. Facts are sacred: The power of data . London: The Guardian.

Ross, Eleanor. 2015. "Why Open Data Doesn't Mean Open Governance?" Media & Tech Network. Accessed June 4, 2017. https://www.theguardian.com/media-network/2015/dec/02/china-russia-open-data-open-government

Shen, Hao, and Chen Luo. 2016. "Data Journalism: Historical Prospect from the Perspective of Modernity. "Journalism Bimonthly, 2:1-5.

Splendore, Sergio. 2016. "Quantitatively Oriented Forms of Journalism and Their Epistemology."Sociology Compass, 10(5): 343-352.doi: 10.1111/soc4.12366.

Spradley, James P. 1979. The Ethnographic Interview. New York: Holt, Rinehart & Winston.

Stavelin, Eirik. 2013. "Computational Journalism: when journalism meets programming." PhD diss., University of Bergen.

Stray, Jonathan. 2011 "A computational journalism reading list." Accessed June 1. http://jonathanstray.com/a-computational-journalism-reading-list.

Tabary, Constance, Anne-Marie Provost, and Alexandre Trottier. 2015."Data journalism's actors, practices and skills: A case study from Quebec ."Journalism, 17(1): 66-84.

Tandoc Jr., Edson C., and Soo-Kwang Oh. "Small Departures, Big Continuities? Norms, values, and routines in The Guardian's big data journalism"Journalism Studies,18(8): 997-1015.

Uskali, Turo, and Heikki Kuutti. 2015."Models and Streams of Data Journalism."The Journal of Media Innovations, 2(1): 77-88.

Wang, Qiong, Liu Zhenzhen, Tian Qing, and Wang Wenchao. 2015. Chinese Data Journalism Development Report. Hubei: Wuhan University.

Weinacht, Stefan, and Ralf Spiller. 2014. "Datejournalismus in Deutschland: Eine explorative Untersuchung zu Rollenbildern von Datenjournalisten."Publizistik 59(4):411-433.

Young, Mary Lynn, Alfred Hermida, and Johanna Fulda. 2017. "What Makes for Great Data Journalism? A content analysis of data journalism awards finalists 2012-2015."Journalism Practice.doi: 10.1080/ 17512786.2016.1270171.

Zeng, Qingxiang, and Hou Xueqi. 2015. "Data journalism: the diminishing power of discourse of soical elites ."Exploration and Free Views 3: 83-86.

Zhang, Zhian, and Shen Fei. 2012. "A Broad-based Survey of Investigative Reporters in China ."Journalism & Communication 4: 51-55.

AUTHOR'S BIOGRAPHY

Shuling Zhang, is Associate Professor at the School of International Studies, University of International Business & Economics. Her research interest lies in data journalism, cross-cultural and media studies. Contact info: No. 10 Huixin Dongjie, Chaoyang District, Beijing, China, 100029. E-mail: adazsl@126.com ORCID:https://orcid.org/0000-0002-3553-7276

APPENDIX 1: LIST OF INTERVIEWEES

Interviewee (Coded based on job title)	Media outlet	Data section	Foundingtime	Media Type
DJ1	Sohu.com	Matrix	May-2011	Internet portal
DJ2	Netease.com	Data Blog	Jan-2012	Internet portal
PM1	Xinhua.net	Data News	Nov-2012	Traditional media

DJ3	Sina.com.cn	Infographics News	Jun-2012	Internet portal
PM2	Caixin.com	Data news	Jun-2013	Traditional media
DE1	People.com	Infographic News	Jan-2013	Traditional media
PM3 DJ4	Yicai.com	CBNData	Jul-2013	New media
DE2	CBNweekly	The Rising Lab	Sep-2015	New media
DJ5	The Beijing News	Graphic News	Jul-2013	Traditional media
PM4 DJ6	CCTV	N/A	N/A	Traditional media
DJ7	The paper.cn	Meishuke	Jul-2014	New media
DE3	Dyclub Data Media Lab	Dyclub	Jan-2015	Independent producer
DE4	CNPolitics	China Politics	Feb-2015	Traditional media
PM5	Data Works	Data works	Aug-2016	Independent producer
PM6	Ifeng.com	Dayu	Aug-2015	New media
DE5	South Reviews Magazine	Tuzheng	Oct-2013	Traditional media
PM7	Jiefang Daily	Graphics	Dec-2016	Traditional media
PM8	Metrodata	Metrodata	Feb-2015	Independent producer