10. Alternative Data Practices in China

Yolanda Jinxin Ma

Abstract

This chapter gives an insider view of the landscape of data journalism in China, its key players and data culture, as well as some practical tips.

Keywords: China, data culture, citizen participation, open data, data journalism, data visualization

A couple of years ago, I delivered a presentation introducing data journalism in China at the Google News Summit, organized by Google News Lab. It was a beautiful winter day in the heart of Silicon Valley, and the audience comprised a packed room of a hundred or so senior media professionals, mainly from Western countries. I started by asking them to raise their hands if they think, firstly, that there is no good data in China, and secondly, that there is no real journalism in China. Both questions got quite some hands up, along with some laughter.

These are two common beliefs, if not biases, that I encounter often when I attend or speak at international journalism conferences. From my observations over the past six years, far from there being no data, in fact a vast quantity of data is generated every day in China, and of rapidly improving quality and broader societal relevance. Instead of no “real” journalism being done, there are many journalists producing important stories every day, although not all of them are ultimately published.

Issue-Driven Data Creation

Data stories were being produced even before the term “data journalism” was introduced in China. While nowadays we normally use the term “data-driven stories” in China, there was a period when we saw the contrary: Instead of

data being the driver of stories, we witnessed stories, or particular issues, driving the production of data. This typically occurred in relation to issues that resonate with regular citizens, such as air pollution.

Since 2010, the Ministry of the Environment has published a real-time air pollution index, but one important figure was missing.¹ The data on particulate matter (PM), or pollutants that measure less than 2.5 micrometres in diameter, which can cause irreversible harm to human bodies, was not published.

Given the severity of air pollution and the lack of official data on PM_{2.5}, a nationwide campaign started in November 2011 called “I test the air for the motherland.” The campaign advocated for every citizen to contribute to monitoring air quality and to publish the results on social media platforms.² The campaign was initiated by an environmental non-profit. The testing equipment was crowd-funded by citizens, and the non-profit organization provided training to interested volunteers. This mobilization gained broader momentum after a few online influencers joined forces, including Pan Shiyi, a well-known business leader, who then had more than 7 million followers on Sina Weibo, one of China’s most widely used social media platforms (Page, 2011).

After two years of public campaigning, the data on PM_{2.5} was finally included in the government data release. It was a good start, but challenges remained. Doubts about the accuracy of the data were prompted by discrepancies between the data released by the government and that released by the U.S. embassy in China (Spegele, 2012).

The data was also not journalist-friendly. Despite hourly updates from more than a hundred cities, the information was only provided on a rolling basis on the web page, with no option to download a data set in any format. Although data has been centralized, historical data is not publicly accessible. In other words, without being able to write a script to scrape the data every hour and save it locally, it is impossible to do any analysis of trends over time or undertake comparisons between cities.

That is not the end of the story. Issue-driven data generation continues. When the data is not well structured and when data journalists struggle due to limited technical skills, civil society and “tech geeks” step in to provide support.

One early example back in 2011 was PM25.in, which scrapes air pollution data and releases it in a clean format. The site claims more than 1 billion

1 http://www.gov.cn/jrzq/2010-11/25/content_1753524.htm (Chinese language, Xinhua News Agency, National air quality real-time release system launched in Beijing, November 25, 2010)

2 <http://www.bjep.org.cn/pages/Index/40-1699?rid=2782> (Chinese language)

search queries since they started operating.³ Another example is Qing Yue, a non-governmental organization which collects and cleans environmental data from government websites at all levels, and then releases it to the public in user-friendly formats. Their processed data is widely used not only by data teams in established media outlets but also by government agencies themselves for better policymaking.

The generation of data and the rising awareness around certain issues have gone hand in hand. In 2015, a documentary investigating the severity of air pollution took the country by storm. The self-funded film, entitled *Under the Dome*, exposed the environmental crisis of noxious smog across the country and traced the roots of the problem and the various parties responsible (Jing, 2015). The film has been compared with Al Gore's *An Inconvenient Truth* in both style and impact. The storytelling featured a lot of scientific data, charts explaining yearly trends, and social network visualizations of corruption within environment and energy industries. As soon as it was released online, the film went viral and reached 200 million hits within three days, before it was censored and taken down within a week. But it had successfully raised public awareness and ignited a national debate on the issue, including around the accessibility and quality of air pollution data. It has also successfully made the country's leadership aware of the significance of the issue.

Two weeks after the release of the documentary, at a press conference held by the National People's Congress, Premier Li Keqiang addressed a question about air pollution which referred to the film, admitting that the government was failing to satisfy public demands to halt pollution. He acknowledged some of the problems raised by the documentary, including lax enforcement of pollution restrictions, and emphasized that the government would impose heavier punishments to cut the toxic smog (Wong & Buckley, 2015). At the end of August 2015, the new Air Pollution Prevention and Control Law was issued, and it was implemented January 2016 (Lijian et al., 2015).

Air pollution is only one example illustrating that even when data availability or accessibility pose a challenge, public concern with issues can lead to citizen contributions to data generation, as well as to changes in government attitudes and in the availability of public sector data on the issues at hand. In more established ecosystems, data may be more readily available and easy to use, and the journalist's job more straightforward: To find data and use it as a basis for stories. In China the process can be less

3 <https://chrome.google.com/webstore/detail/pm25by-bestapp-labs/bgjclhmlafjipakmliln-loihemmlfndj>

linear, and citizens, government, civil society and the media may interact at multiple stages in this process. Data, instead of just serving as the starting point for stories, can also come into the picture at a later stage to enable new kinds of relations between journalists and their publics.

Evolving Data Culture

The data environment in China has been changing rapidly in the past decade. This is partly driven by the dynamics described thus far in this chapter, and partly due to other factors, including the global open data movement, rapidly growing Internet companies and a surprisingly high mobile penetration rate. Data culture has been evolving around these trends as well.

Government legislation provides the policy backbone for data availability. To the surprise of many, China does have laws around freedom of information. The State Council Regulations on the Disclosure of Government Information was adopted in 2007 and came into force on May 1, 2008. The law has a disclosure mandate and affirms a commitment to government transparency. Following the regulation, government agencies at all levels set up dedicated web pages to disclose information they hold, including data sets.

However, although it gave journalists the right to request certain data or information from the authorities, in the first three years since the law was enforced, there are no publicly known cases of any media or journalists requesting data disclosure, according to a 2011 study published by Caixin, a media group based in Beijing and known for investigative journalism.⁴ The study revealed that, in 2010, the *Southern Weekly*, a leading newspaper, only got a 44% response rate to a request sent to 29 environmental bureaus to test their degree of compliance with the law. Media organizations do not usually have a legal team or other systems to support journalists to advance their investigations and further their information requests. In another instance, one journalist who, in his personal capacity, took the government to the court for not disclosing information, ended up losing his job. The difficulties and risks that Chinese journalists encounter when leveraging legal tools can be much greater than those experienced by their Western peers.

China is also responding to the global open data movement and increasing interest in big data. In 2012, both Shanghai and Beijing launched their own

4 <http://finance.ifeng.com/leadership/gdsp/20110901/4512444.shtml> (Chinese language)

open data portals. Each of them holds hundreds of data sets on issues such as land usage, transportation, education and pollution monitoring. In the following years, more than a dozen open data portals have been set up, not only in the biggest cities, but also in local districts and less-developed provinces. The development was rather bottom-up, without a template or standard structure for data release at the local level, which did not contribute to the broader comparability or usability of this data.

By 2015, the State Council had released the Big Data Development Action Plan, where open data was officially recognized as one of the ten key national projects, and a concrete timeline for opening government data was presented.⁵ However, official data is not always where journalists start, and also not always aligned with public interests and concerns.

On the other hand, the private sector, especially the technology giants such as Alibaba or Tencent, have over the years accumulated huge amounts of data. According to its latest official results, Alibaba's annual active consumers reached 601 million by September 30, 2018 ("Alibaba Group Announces," 2018). The e-commerce data from such a strong user base—equivalent to the entire Southeast Asian population—can reveal lots of trading trends, demographic shifts, urban migration directions, consumer habit changes and so on. There are also vertical review sites where more specific data is available, such as Dianping, the Chinese equivalent of Yelp. Despite concerns around privacy and security, if used properly, those platforms provide rich resources for data journalists to mine.

One outstanding example in leveraging big data is the Rising Lab, a team under the Shanghai Media Group, specializing in data stories about urban life.⁶ The Lab was set up as an answer to the emerging trend of urbanization: China has more than 600 cities now, compared to 193 in 1978, with 56% of the population living in urban areas, according to a 2016 government report ("Gov't Report: China's Urbanization," 2016). Shifting together with the rapid urbanization is the rise of Internet and mobile use, as well as lifestyle changes, such as the rapid adoption of sharing economy models. These trends are having a big impact on data aggregation.

With partnership agreements and technical support from tech companies, the Lab collected data from websites and apps frequently used by city dwellers. This data reflected various aspects of urban life, including property prices, numbers of coffee shops and bars, numbers of co-working spaces, and quality of public transportation. Coupled with its original methodology,

5 http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm (Chinese language)

6 <https://zhuanlan.zhihu.com/therisinglab> (Chinese language)

Figure 10.1. Cities Beyond Data 2018: Ranking of business attractiveness of cities in China. Source: *The Rising Lab*.

the Lab has produced a series of city rankings taking into account aspects such as commercial attractiveness, level of innovation and diversity of life (Figure 10.1). The rankings and the stories are updated every year based on new data, but follow the same methodology to ensure consistency. The concept and stories have been well received by the public and have begun to influence urban planning policies and companies' business decisions, according to Shen Congle, director of the Lab (Shen, 2018).

The Lab's success illustrates the new dynamics emerging between data providers, journalists, and citizens. It shows how softer topics have also become a playground for data journalism, alongside other pressing issues, such as the environmental crisis, corruption, judicial injustice, public health and money laundering. It also explores new potential business models for data journalism, as well as how data-based products can bring value to governments and businesses.

Readers' news consumption practices have also had an impact on the development of data journalism. Two aspects deserve attention here, one being visual news consumption and the other, mobile news consumption. Since 2011, infographics have become popular thanks to a few major news portals' efforts to build dedicated vertices with infographics stories, mostly driven by data. In 2014, the story of the downfall of the former security chief Zhou Yongkang, one of the nine most senior politicians in China, was the

Figure 10.2. An interactive piece showing Zhou's networks. Source: Caixin.

biggest news of the year. Together with the news story, Caixin produced an interactive social network visualization (Figure 10.2) to illustrate the complex network around Zhou, including 37 people and 105 companies or projects connected to him, and the relationship between these entities, all based on the 60,000-word investigative piece produced by its reporting team. The interactive received 4 million hits within one week, and another 20 million views on social media, according to Caixin.⁷ The wide circulation of this project brought new kinds of data storytelling to new publics, and created an appetite for visual stories which didn't exist before.

Almost at the same time, the media industry was welcoming the mobile era. More and more data stories, like any other online content in China, are now disseminated mostly on mobile. According to the China Internet Network Information Center (CNNIC), more than 95% of Internet users in the country used a mobile device to access the Internet in 2016 (Chung, 2017). WeChat, the domestic popular messaging app and social media platform, reached 1 billion users in March 2018 (Jao, 2018). The

7 <http://bit.ly/storybench-caixin> (Chinese language, Storybench, How a 4-million-hit data journalism product was produced, October 10, 2017)

dominance of mobile platforms means data stories in China are now not only mobile-first, but in many cases mobile-only. Such market demand led to a lot of lean, simple and sometime creative interactives that are mobile friendly.

In short, the data culture in China has been evolving, driven by various factors from global movements to government legislation, from public demand to media requests, from new generations of data providers, to new generations of news consumers. The interdependent relationships between players have created very complex dynamics, where constraints and opportunities coexist. Data journalism has bloomed and advanced along its own path in China.

Practical Tips

This final section is aimed at readers of this book who are looking to work on China-related stories and wondering where to get started. It will not be easy. If you are not a Chinese language speaker, you will be faced with language barriers, as most data sources are only available in Chinese. Next you will be faced with common issues pertaining to working with data: Data accuracy, data completeness, data inconsistency, etc., but we will assume that, as a reader of this book, you have the skills to deal with these issues, or at least a willingness to learn. A good way to start would be to identify the biggest players in data journalism in China. Quite a few of the leading media outlets have data teams, and it is good to follow their stories and talk to their reporters for tips. Here are a few you should know: The Data Visualisation Lab (Caixin), Beautiful Data Channel (The Paper), The Rising Lab (Shanghai Media Group), and DT Finance.⁸

The second question pertains to where to find data. A comprehensive list of data sources would be a separate book, so here are just a few suggestions to get started. Start with government websites, both central ministries and local agencies. You would need to know which department is the right one for the data you are looking for, and you should check both the thematic areas of ministries (for example, the Ministry of Environmental Protection) and the dedicated data website at the local level, if it exists.

8 <http://datanews.caixin.com/> (Chinese language), https://www.thepaper.cn/list_25635 (Chinese language), <https://www.cbnweek.com/topics/10> (Chinese language), <https://www.dtcj.com/> (Chinese language)

There will be data that you don't even expect—for example, would you expect that the Chinese government published millions of court judgements after 2014 in full text? Legal documents are relatively transparent in the United States but not in China. But the Supreme People's Court (SPC) started a database called China Judgments Online doing just that.

Once you find some data that could be useful online, make sure to download a local copy. It is still common that data is not available online. Sometimes the data is published in the form of annual government reports which you can order online, or available only in paper archives. For example, certain government agencies have the records of private companies but not all of these are available online.

If the data is not released at all by the government, check if any user-generated content is available. For example, data on public health is very limited, but there are dedicated websites with information on hospital registrations or elderly centres, among others. Scraping and cleaning this data would help you gain a good overview of the topic.

It is also recommended to utilize databases in Hong Kong, anything from official ones like the Hong Kong Companies Registry, to independent ones such as Webb-site Reports. As mainland China and Hong Kong are becoming politically and financially closer, more information is available there, thanks to Hong Kong's transparent environment and legal enforcement, which may be valuable for tracing money.

There is also data about China not necessarily held in China. There are international organizations or academic institutions that have rich China-related data sets. For example, The Paper used data from NASA and Harvard University in one of its latest stories.

Last but not least, while some challenges and experience are unique to China, a lot of them could potentially provide useful lessons for journalists in other countries, where the social, cultural and political arrangements have a different shape but similar constraints.

Works Cited

- Alibaba Group announces September quarter 2018 results. (2018, November 2). *Business Wire*. <https://www.businesswire.com/news/home/20181102005230/en/Alibaba-Group-Announces-September-Quarter-2018-Results>
- Chung, M.-C. (2017, February 2). More than 95% of Internet users in China use mobile devices to go online. EMarketer. <https://www.emarketer.com/Article/More-than-95-of-Internet-Users-China-Use-Mobile-Devices-Go-Online/1015155>

- Gov't report: China's urbanization level reached 56.1%. (2016, April 20). *CNTV*. http://english.www.gov.cn/news/video/2016/04/20/content_281475331447793.htm
- Jao, N. (2018, March 5). WeChat now has over 1 billion active monthly users worldwide. *TechNode*. <https://technode.com/2018/03/05/wechat-1-billion-users/>
- Jing, C. (2015, February 28). Chai Jing's review: Under the dome—investigating China's smog. <https://www.youtube.com/watch?v=T6X2uwlQGM>
- Lijian, Z., Xie, T., & Tang, J. (2015, December 30). How China's new air law aims to curb pollution. *China Dialogue*. <https://www.chinadialogue.net/article/show/single/en/8512-How-China-s-new-air-law-aims-to-curb-pollution>
- Page, J. (2011, November 8). Microbloggers pressure Beijing to improve air pollution monitoring. *The Wall Street Journal*. <https://blogs.wsj.com/chinarealtime/2011/11/08/internet-puts-pressure-on-beijing-to-improve-air-pollution-monitoring/>
- Shen, J. (2018, October). Data journalism in China panel. Uncovering Asia, Investigative Journalism Conference, Seoul. <https://2018.uncoveringasia.org/schedule/>
- Spegele, B. (2012, January 23). Comparing pollution data: Beijing vs. U.S. embassy on PM2.5. *The Wall Street Journal*. <https://blogs.wsj.com/chinarealtime/2012/01/23/comparing-pollution-data-beijing-vs-u-s-embassy-on-pm2-5/>
- Wong, E., & Buckley, C. (2015, March 15). Chinese premier vows tougher regulation on air pollution. *The New York Times*. <https://www.nytimes.com/2015/03/16/world/asia/chinese-premier-li-keqiang-vows-tougher-regulation-on-air-pollution.html>

About the Author

Yolanda Jinxin Ma is a Digital Transformation Specialist with the United Nations, and before that she was a data journalism practitioner and evangelist who introduced data journalism to China and across Asia.