Web Archiving for Librarians:
A Three-University Collaboration to Preserve Non-Official Voices in China's Anti-Corruption Campaign

Ding Ye, Daniel Kerchner, Yan He, Cathy Zeljak, and Yunshan Ye[*]

Introduction

In 2014, librarians from Johns Hopkins University, the George Washington University, and Georgetown University recognized the emergence of social media and microblogs as an opportunity to capture the everyday voices from a unique period in Chinese history. Concurrently, tools were maturing to enable the capture of both web and social media content to build collections of archived web and social media content. Seizing this opportunity, we joined forces, applied for, and were awarded a Mellon innovation grant administered by the Council on East Asian Libraries (CEAL) to use and adapt these technologies to build a collection of Sina Weibo and web-based blog content focused on reaction to and commentary about the current Chinese government's anti-corruption campaign from within China. We also saw this as an opportunity to establish a model project for librarians in East Asian studies and beyond that can be replicated, adapted and sustained for other future social media archiving projects.

This paper describes the significance of the project and its intended contributions, and provides an update of progress to date.

Why is Web Archiving of Chinese Social Media Important?

Web content in general, and social media in particular, tends to disappear quickly even in free and open societies due to non-political reasons. Various studies have looked into the average lifespan of web pages, with results ranging from 44 to 75 days.[1] Specific content on social media disappears even more frequently. In some cases, content can be 'relocated' when sites are re-designed.

In countries where government authorities exercise censorship, web content is all the more at risk. Under politically sensitive and tightly controlled regimes, content on social media can simply be deleted by government censors. In China, the current government continues to tighten access to the Internet and corresponding social media platforms. Unsanctioned access using virtual private network ("VPN") connections, until recently a popular method for bypassing some access restrictions, has been the focus of a new Chinese government clampdown.[2]

Notwithstanding certain reforms, the ruling Communist Party of China (CPC) practices censorship on all forms of media for a variety of perceived reasons, ranging from political to moral to economic. Traditional me-

** Ding Ye is Asian Studies and Linguistics Bibliographer, Georgetown University; Daniel Kerchner is Senior Software Developer and Librarian, George Washington University; Yan He is China Studies Librarian, George Washington University; Cathy Zeljak is Director of Global Resources, George Washington University; and Yunshan Ye is Librarian for Political Science, International Studies and East Asian Studies, Johns Hopkins University.*

dia such as radio, television and newspapers are under tight government control and serve primarily as an extension of official Party communication. Under such conditions, social media has emerged as a significant public space where ordinary Chinese citizens can express their unedited views, including criticisms of the government. Access to Facebook, Twitter, YouTube and other major social media platforms is now blocked by the Chinese government; however, WeChat instant messaging, Sina Weibo—a microblogging platform similar to Twitter—and other local platforms offer the Chinese people local choices to meet their social media needs. Blogging and microblogging play a crucial role in China's socio-political activism. They represent non-official voices of history that often challenge the official narratives in government-run, tightly-controlled traditional media.

In addition to having the world's largest Internet user base—nearly 600 million people, more than double the 250 million users in the United States—China also has the world's most active environment for social media, with 100 million bloggers and 300 million microbloggers at present.[3] Despite the government's continual efforts to monitor and control Internet traffic, the sheer volume and scale of Chinese social media makes it technically impossible to censor everything. As a result, the Internet, and social media in particular, has "effectively eliminated the government's monopoly on information."[4]

An indicator of its lasting significance, over the last several years Chinese social media has become a favorite subject for scholarship by Chinese studies scholars worldwide. A search on "social media and China" in ProQuest's Dissertations and Theses produces 10,216 hits, with the majority of titles (9,926) published since 2010. A more focused search on Weibo (微博) in the Chinese Dissertations and Theses database (by Wanfang Data) results in an even more hits: 12,482, the majority of which are published since 2011.

Initiated in July 2015, this web archiving project marks a first step toward archiving Chinese social media. With joint resources and technical expertise, librarians from three premier research institutions have collaboratively developed tools and strategies to archive selected Chinese social media sites. The project immediately benefits scholars interested in the use of social media as a tool for Chinese socio-political activism and expression within civil society.

Many scholars at the three collaborating universities have recognized the importance of the work of this project. For example, Joel Andreas, Associate Professor of Sociology at Johns Hopkins University, upon learning of the project, shared his manuscript paper just recently submitted for publication with the project team. Entitled "Mass Supervision and the Bureaucratization of Governance in China," the article studies cyber-activism as a new form of "mass supervision" that was once popular during the Maoist era (1949-1976). Similarly, Jackson Woods, a PhD student in Political Science at George Washington University, while studying political discourse on the web in contemporary China was excited to be able to utilize the collection in his dissertation project.

In addition to benefiting researchers, this project also creates tools and documents methods that can be used by librarians who wish to pursue similar projects to archive Chinese blogs and microblogs. One of the project outcomes is a white paper documenting the web archiving process and lessons learned from the experience, with particular emphasis on the unique challenges encountered when collecting China-based social media content. Given the scarcity of such information, this document will fill a major gap and likely be of significant value to the wider library community.

Why Was the Anti-Corruption Campaign Chosen as the Theme?

This project focuses on China's ongoing anti-corruption campaign. Even though fighting corruption has been a perennial theme throughout Chinese history, the Anti-Corruption Campaign launched by the Chinese President Xi Jinping right after he was named General Secretary of the CPC in November 2012 has evolved into a major political movement and has generated great interest from scholars both inside and outside China.[5] More significantly, social media has played a crucial role in the ongoing campaign.[6]

What distinguishes China's anti-corruption legal framework from other nations' is that there is not a specific anti-corruption law in China. More generally, the criminal law of China serves as the leading legal weapon to define and punish corrupt behaviors. According to China's Criminal Statutory Law, corruption is generally viewed as any activity which employs public power to earn private benefits.[7] This is a very loose definition of corruption in the Chinese setting but is widely accepted by the Chinese government.

Corruption among China's government servants was historically believed to be the major cause of dynastic collapse, including the collapse of the last rulers of China before the communists, the KMT Nationalists, who left for Taiwan in 1949. After Mao-Tse Tung's rise to power in 1949 and until his death in 1976, a series of political campaigns took aim at cadre corruption, reaching a climax during the decade of the Cultural Revolution (1966-1976).

The next fifteen years from 1976 to 1992 saw China open its doors during the Deng Xiaoping era. The wide-ranging reforms central to this economic transformation brought corruption in China to new heights. Corrupt practices allowed China's top leaders to profit handsomely as they engaged in various business enterprises. Their roles as middlemen or brokers (*Guandao* 官倒 bureaucrat-profiteers) became the major cause of China's pro-democracy movement in 1989.

Under the leadership of Jiang Zemin, Deng's successor, the decade from 1992 to 2002 witnessed a period of increased emphasis on anti-corruption. In 1993, the term "anti-corruption" officially entered into government documents and the Party Charter. Although the central government took heightened actions to crack down on corruption, the absence of an effective legal framework hindered results. Additionally, ineffective laws and regulations permitted top corporate officers of state-owned enterprises to avoid repercussions for a wide range of corrupt practices.

The first decade of the 21st century saw the advent of more institutional means to curb corruption. According to the National Bureau of Corruption Prevention, more than 43,000 officials were put on trial between 2003 and 2011. In 2013, under Xi Jinping's new leadership, the CPC launched a new, more effective "war on corruption." To date, the Central Discipline and Inspection Committee (CDIC) led by one of Xi's most trusted colleagues, Wang Qishan, has announced the investigation of hundreds of high-ranking officials commonly known as "tigers (大老虎)" and hundreds of thousands of low-ranking officials commonly known as "flies (小苍蝇)." The tigers who were prosecuted and sentenced in 2014 and 2015 included China's top leaders including Bo Xilai (薄熙来), Zhou Yongkang (周永康), Xu Caihou (徐才厚) and Ling Jihua (令计划), known as the new Gang of Four.

Even though countering corruption has become the government's central objective, it is still not legally defined. Some critics have claimed that the current anti-corruption campaign is the result of a power struggle to cleanse and purge political opponents. Such discussions attracted wide attention on social media. While the central government has depended on reliable propaganda tools including newspapers, TV, magazines and other traditional media, they also engage new technologies such as social media to promote and defend the campaign. Government communication via social media has increased over the last five years to become a very important informational and interactive tool to combat corruption. Popular social media outlets include WeChat, blog, micro-blog, bbs, forums and online news. At the same time, since most Chinese citizens can access the Internet, spreading evidence of suspected corruption can create huge social pressure to force government agencies to intervene.

The advantages of using social media over traditional media for Chinese citizens wanting to communicate are manifold: it is more open and accessible; information spreads more quickly; the costs are much lower; and contributions can be initiated more anonymously. As a result, individuals are able to effectively expose corrupt practices or critique government action that would not otherwise be revealed, and can do so with less worry about retaliation. Social media plays a significant role in empowering individuals who otherwise would not have adequate protections, either legally or through CPC's practice in ruling the nation. The government in turns blocks access to certain sites, strengthens firewalls and launches other strict controls on social media in order to reestablish control over anti-corruption initiatives.

Based on the historical significance of the event, scholarly interest, and the crucial role of social media, we decided that the Anti-Corruption Campaign would be a perfect topic for our collective efforts. Our project aims to archive both the government's accounts and private accounts focused on the anti-corruption campaign with the concern that some social media contents might forever get "lost" due to government censorship and the passage of time. Despite its clear historical and scholarly importance, archiving social media has remained a major challenge for librarians and archivists worldwide. The Internet Archive and the Twitter Archive Project at the Library of Congress represent two major steps in the West in Internet archiving. However, in the case of Chinese Internet sources, especially Chinese social media, little has been done to preserve this valuable source of Chinese history, either inside or outside China.[8]

For this reason, we saw the project as an opportunity to contribute to the field of China Studies as well as librarianship and web archiving. In addition to preserving an invaluable part of Chinese history, this collaborative project will also serve as new model for future endeavors in archiving "at risk" web sources, and will create software for collecting Weibo content that will be made available to others with an interest in creating Weibo collections in the future.

General Approach

This project focuses on identifying and archiving content from two sources: Blog sites on the Internet, and social media posts on Sina Weibo.

A primary web content archiving technology available today is the Heritrix web crawler. The Internet Archive now provides a hosted web archiving platform (based on Heritrix) called Archive-It. Archive-It offers a convenient interface for librarians and archivists to build collections by identifying sites to crawl and by specifying a crawling schedule; curators also specify collection metadata such as subject tags, to make the collection more discoverable. This project's Archive-It collection, curated primarily by Johns Hopkins University Libraries, has so far archived over 1,000 Chinese microblog sites related to the Anti-Corruption campaign. The collection is publicly viewable at https://archive-it.org/collections/6314.

Collecting social media content can be viewed as a type of web archiving, but it has some distinguishing features that call for more specialized software tools.[9] Social Feed Manager (SFM), open-source software developed by the George Washington University Libraries with the support of grants from the Institute for Museum and Library Services and the National Historical Preservation and Records Commission, is designed specifically to enable archivists and researchers to build collections of social media data.[10] SFM provides the ability to collect content from Twitter, Tumblr, and Flickr; this project's grant from CEAL supported the GW's development of a new Sina Weibo module for SFM to enable this and potential future Weibo archiving efforts. After selecting and following particular Weibo accounts that were likely to have posts discussing the anti-corruption campaign, SFM has harvested roughly 48,000 Weibo posts since mid-2016; the collection continues to grow.

In order to increase utilization of both aspects of this unique collection by researchers, the project also calls for the creation of finding aids to integrate this digital content into library collections, to aid in its discoverability by researchers.

Anticipated Outcomes

The project breaks new ground in several key regards:

1. The project is the first collaborative effort on the institutional level to archive Chinese social media.

With expertise in China studies librarianship and in social media web archiving technology from three top research universities, we are in a unique and advantageous position to take on the challenge of Chinese social media archiving.

The project benefits from strong faculty support in China studies from all three institutions. Professor Erin Chung, chair of East Asian Studies Department at Johns Hopkins University, as well as other core faculty members in the department such as Professor Lingxing Hao and Joel Andreas, all have pledged their full support and agreed to serve on the Advisory Committee. Core faculty at Georgetown University, such as Professor Jingyuan Zhang, Chair of Department of East Asian Languages and Cultures, Professor Philip Kafalas of the same department and several PhD students at Georgetown's departments of History and Sociology all expressed strong enthusiasm and interest in the project.

The expertise of George Washington University Libraries in social media archiving is unique and has been essential to creating a Weibo archiving capability; GW also contributes experience using Archive-It to build web archive collections. The GWU Libraries is a national leader in the area of social media archiving and has several years of experience using its locally-developed software, Social Feed Manager, to build collections for researchers mainly from Twitter, which is similar to Weibo in key respects. GW earned the support of two major grants: an IMLS Sparks! Ignition Grant (2013)[11] to develop the initial SFM prototype, and an innovation grant from National Historical Publications and Records Commission (2014),[12] to extend SFM to automate capturing and preserving data from Twitter, Tumblr, Flickr, and now Weibo, to build a user interface for SFM that enables researchers to build collections directly, and to rearchitect the application to make it easy for other institutions to set up. Daniel Chudnov, former manager of the Twitter Project at the Library of Congress, and Principal Investigator of GWU Libraries social media archiving grants through 2015, helped envision the project. Daniel Kerchner, Senior Software Developer, the ongoing technical lead on the project, has contributed to the SFM code, serves as a consultant to GW researchers building collections with SFM to analyze social media content in the contexts of a variety of academic fields, and has participated in web and social media preservation collaborations at a national level. Christie Peterson, formerly Digital Archivist at JHU and currently GW's University Archivist, has long been using Archive-It to capture and preserve web sources.

2. The project is the first to develop open access software usable by librarians and researchers that is adapted to archive social media content from Sina Weibo.

Archiving social media content from any platform generally requires retrieving data using the platform's Application Programming Interface (API). Each platform's API is different, and each provides varying levels of access to the data. Only institutions with special arrangements, such as in the case of the Library of Congress Twitter Archive Project, have access to all content from a social media platform[13]. Even in the rare cases when all of the platform's content can be obtained, the volume of data presents a formidable challenge in storing and providing access to the data.

For more realistic scenarios, social media APIs provide a service to make more focused requests for data. These APIs generally provide the ability for an application to request the content of specific public microblogs (accounts), to request posts mentioning specific terms (such as hashtags, names, or keywords), or to request a small sample of all posts; other types of requests are often available as well. As SFM has been developed to implement these interactions to retrieve data from Twitter and other platforms, SFM was a natural choice to extend in a similar manner to retrieve social media data using Weibo's API. Furthermore, SFM fills a unique niche as open-source software for building social media collections, in contrast to commercial options for obtaining this data. More specifically, we are not aware of robust open-source software for collecting data from Weibo; as such, SFM enhanced with Weibo collection capability provides an important new option indeed, particularly for librarians and researchers.

3. The project will integrate the records for Chinese social media collections into the existing library catalogs.

Providing access to archived data is a most important aspect of web sources archiving, and also the most challenging part of the project. After investigating the Library of Congress Twitter Archiving Project and consulting with GWU Libraries staff responsible for the Social Feed Manager, we decided that providing open access to the collected data, while ideal, is out of scope of this project and would involve legal and ethical challenges[14] that cannot be resolved during the grant term. We will, however, integrate the metadata of the collected social media records into our regular collection indexing systems, so that users can discover and then request the data that they need for academic uses.

4. At the conclusion of this two-year project, a survey of faculty and students on their experience of finding and using the collected social media data will be conducted. The survey results shall be documented and used for improving the products. The final products of the project will include:

- Web archiving software with the capability to capture and store Sina Weibo social media content, and able to capture and export social media data in other East Asian languages;
- A collection of Chinese web content and social media data with finding aids and metadata to enable discovery via the library catalog;
- Integration of the collections with library discovery systems, so that the collections will be discoverable via keywords, subjects, dates;
- Documentation on collection development policies and best practices for storing, describing, and retrieving social media data;
- Documentation on the project's technical implementation and technical challenges encountered that were unique to Weibo collecting.
- All of the project's products will be open access and freely available. Upon completion of the project, the project's products will be promoted via professional email lists and listservs, presentations at CEAL and other library conferences and a publication submitted to the Journal of East Asian Libraries.

Notes

1. Marieke Guy. "What's the Average Lifespan of a Web Page?" *JISC PoWR: Preservation of Web Resources*, accessed January 31, 2017, https://jiscpowr.jiscinvolve.org/wp/2009/08/12/whats-the-average-lifespan-of-a-web-page/
2. "China Tightens Great Firewall by Declaring Unauthorized VPN Services Illegal," *South China Morning Post*, January 23, 2017, accessed January 31, 2017, http://www.scmp.com/news/china/policies-politics/article/2064587/chinas-move-clean-vpns-and-strengthen-great-firewall.
3. Cindy Chiu, Chris Ip, and Ari Silverman, "Understanding Social Media in China," *McKinsey Quarterly* 2 (2012): 78-81.
4. Jonathan Sullivan, "China's Weibo: Is Faster Different?," *New Media & Society* 16, no. 1 (2014): 24-37.
5. Roderic Broadhurst and Peng Wang, "After the Bo Xilai trial: Does Corruption Threaten China's Future?," *Survival* 56, no. 3 (2014): 157-178.
6. Xiaoyang Meng and Caixia Tan. "Fan fu xin zhan chang (New Battle Field against Corruption)," *Xin wen shi jie (News World)* 11 (2014): 72-73.
7. "Chapter VIII Crimes of Embezzlement and Bribery," *Criminal Law of People's Republic of China*, accessed April 6, 2017, https://www.cecc.gov/resources/legal-provisions/criminal-law-of-the-peoples-republic-of-china#2%20Chapter%20VIII.
8. Pan Liao and Guomin Liu. "Weibo chang qi bao cun de ke xing xing yan jiu (On Feasibility of Permanently Preserving Weibo)," *Tu shu guan lun tan (Library Tribune)* 2 (2013): 45-62.
9. Justin Littman, Daniel Chudnov, Daniel Kerchner, Christie Peterson, Yecheng Tan, Rachel Trent, Rajat Vij, and Laura Wrubel. "API-based Social Media Collecting as a Form of Web Archiving." *International Journal on Digital Libraries*: 1-18.

10. Further information about the Social Feed Manager project and software application can be found at https://gwu-libraries.github.io/sfm-ui/, accessed January 31, 2017.
11. Institute of Museum and Library Services grant #LG-46-13-0257-13.
12. National Historical Publications and Records Commission grant #NARDI-14-50017-14.
13. Library of Congress. "Update on the Twitter Archive at the Library of Congress," accessed January 31, 2017, https://www.loc.gov/static/managed-content/uploads/sites/6/2017/02/twitter_report_2013jan.pdf.
14. A short list of recent publications on the ethical and legal considerations in collecting and using social media data can be found at https://gwu-libraries.github.io/sfm-ui/resources/ethics, accessed January 31, 2017.