10-2017

Web-Archiving Chinese Social Media: Final Project Report

Yunshan Ye
Ding Ye
Cathy Zeljak
Daniel Kerchner
Yan He

See next page for additional authors

Follow this and additional works at: https://scholarsarchive.byu.edu/jeal

BYU ScholarsArchive Citation
Available at: https://scholarsarchive.byu.edu/jeal/vol2017/iss165/11

This Report is brought to you for free and open access by the All Journals at BYU ScholarsArchive. It has been accepted for inclusion in Journal of East Asian Libraries by an authorized editor of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
Web-Archiving Chinese Social Media: Final Project Report

Authors
Yunshan Ye, Ding Ye, Cathy Zeljak, Daniel Kerchner, Yan He, and Justin Littman

This report is available in Journal of East Asian Libraries: https://scholarsarchive.byu.edu/jeal/vol2017/iss165/11
Web-Archiving Chinese Social Media: Final Project Report
August 2017

Yunshan Ye (East Asian Studies Librarian, Johns Hopkins University), Project Lead
Ding Ye (Asian Studies and Linguistics Bibliographer, Georgetown University)
Cathy Zeljak (Director of Global Resources, George Washington University)
Daniel Kerchner (Senior Software Developer and Librarian, George Washington University)
Yan He (China Studies Librarian, George Washington University)
Justin Littman (Software Developer and Librarian, George Washington University)

I. PROJECT OVERVIEW

Recognizing the critical roles of social media in Chinese socio-political activism and importance and the urgency of preserving these born digital, ephemeral records, librarians from three universities, including Johns Hopkins University (JHU), George Washington University (GW), and Georgetown University (GU) decided to join forces and collaborate on a web-archiving project in 2014. Using the latest technology available, the project intended to preserve Chinese social media responses to the ongoing anti-corruption campaign, a major political campaign launched by the Chinese President Xi Jinping in 2012. Besides preserving an invaluable part of history, this collaborative effort also serves as a model for future endeavors in archiving "at risk" web sources in the field of East Asian Studies and beyond. The project received enthusiastic endorsement from scholars and library administrators. It was funded over a two-year period (July 1, 2015 to June 30, 2017) by the Mellon Foundation-CEAL Innovation Grant Program. At the close of the grant term, we are proud to have accomplished all the major goals we set out in the initial proposal, including:

- Developing 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;
- Building a collection of Chinese web content and social media data with finding aids and metadata to enable discovery via the library catalog;
- Documenting the project’s technical implementation, technical challenges and solutions to help future similar projects;
- Presenting and promoting the collections at national professional venues such as the annual conferences of CEAL (March, 2016) and ACRL (March, 2017).
Below is the final report of the project. We will start with a backgrounder on Chinese social media and the anti-corruption campaign, a description of our general approach and outcome, followed by detailed reports on two components of the project: the Archive-It collection at JHU and Social Feed Manager collection at GW.

1. Chinese Social Media and the Anti-corruption Campaign: the Project Rationale
The project focused on archiving online postings related to the Chinese anti-corruption campaign on Chinese social media platforms such as Sina Weibo, Sina blog and other blogging and news sites. Before we go into the details of the project, we would like to give a brief summary of our research on Chinese social media and the anti-corruption campaign that led to the idea and eventual proposal and implementation of the project.

1.1. Why social media?
The past decade has witnessed an explosion of social media, thanks to the rapid advancement of the internet and mobile technology. Social media have become an integral part of our life today. Yet, because of the transient, ephemeral nature of social media postings, these born digital materials are prone to disappearing quickly, becoming lost forever to future researchers. For this reason, how to preserve social media records for future studies has presented a challenge to archivists and librarians.

In the case of China, preserving social media records has become even more critical. The ruling Communist Party of China (CPC) practices censorship of all forms of media for a variety of perceived reasons, ranging from political to moral to economic. Traditional media 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.¹ 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.”

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 have been published since 2011.

Initiated in July 2015, this web archiving project marks a first step towards archiving Chinese social media. With joint resources and technical expertise, librarians from GWU, JHU and GU 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 Ph.D. 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 librarians who wish to pursue similar projects can use to archive Chinese blogs and microblogs. One of the project outcomes is a companion 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.

1.2. Why the Anti-corruption Campaign?
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 generating great interest from scholars both inside and outside China. More significantly, social media has played a crucial role in the ongoing campaign.

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. 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 Zedong’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 (Bulletin-board Sites), 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 the 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 Internet-based/social-media accounts and private accounts focused on the anti-corruption campaign due to the concern that some social media contents might forever be “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 to archive web content. 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.6

For this reason, we view 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 a 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.

2. General Approach and Outcome: a Summary

This project focuses on identifying and archiving content from two sources: postings on “static” blogging and news sites on the Internet, and microblogging from “dynamic” social media sites such as Sina Weibo.

For the “static” blogging sites, we use Archive It (https://archive-it.org/) from the Internet Archive (https://archive.org/) to collect and preserve selected records. Archive-It offers a convenient interface for librarians and archivists to build collections by identifying sites to crawl and then specifying a crawling schedule; curators also add collection metadata such as subject tags, to make the collection more discoverable. This project’s Archive-It collection, curated by Johns Hopkins University Libraries, has so far archived 1,300 Chinese blogs entries and news articles related to the Anti-Corruption campaign. The collection is publicly viewable at https://archive-it.org/collections/6314.

Collecting “dynamic” 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.7 Social Feed Manager (SFM), an 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.8 SFM provides

---


8 Further information about the Social Feed Manager project and software application can be found at https://gwu-libraries.github.io/sfm-ui/.
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 harvester for SFM to enable this and potential future Weibo archiving efforts. This will be discussed in more detail later in this paper.

In addition to the collection of data, our project has also accomplished the following goals:

2.1. Being 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 the Department of East Asian Languages and Cultures, Professor Philip Kafalas of the same department and several PhD students in Georgetown’s departments of History and Sociology all expressed strong enthusiasm and interest in the project. China Studies faculty at George Washington University are also strongly supportive, including Bruce Dickson, Chair of the Political Science Department and a specialist on Chinese politics, authoritarian regimes, regime change, and state-society relations.

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. GW 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)\(^9\) to develop the initial SFM prototype, and an innovation grant from National Historical Publications and Records Commission (2014),\(^10\) 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 re-architect 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

\(^9\) Institute of Museum and Library Services grant #LG-46-13-0257-13.
\(^10\) National Historical Publications and Records Commission grant #NARDI-14-50017-14.
envision the project. Daniel Kerchner, Senior Software Developer, the technical lead on the project and who 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.

2.2. Developing 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. 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. This challenge has yet to be satisfactorily resolved. 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.

2.3. Integrating the records for Chinese social media collections into the existing library discovery tools such as LibGuides and online catalog. In order to facilitate the use of the archived social media collections, we designed a LibGuide page dedicated to the project (http://libguides.gwu.edu/c.php?g=637999&p=4466423). The guide introduces the user to the collections, including an overview of the project, sample data, as well as link to collected data (in the case of Archive It). At Johns Hopkins, the Archive It collection is also discoverable through the library catalog, with links leading directly to the data.

---

11 Please see the companion paper for technical specifications and other details.
Because of legal and privacy concerns, it is more challenging to provide access to the Weibo data; the company specifically forbids republishing their data. After investigating the Library of Congress Twitter Archiving Project and consulting with GW 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\textsuperscript{13} that cannot be resolved during the grant term. As a result, we will rely on the LibGuide and other publicity channels to promote and facilitate the use of the collection.

II. ARCHIVE-IT AND CHINESE BLOG COLLECTION

As mentioned above, our web archiving project consists of two components. One is to collect data from Sina Weibo using Social Feed Manager; this is conducted by the team at George Washington University. The other is to collect data from various blogging and news sites using Archive It and is done by the team at Johns Hopkins. Below is a summary of building the Chinese social media collection using Archive It at JHU.

1. Why Archive-It?

Launched by the Internet Archive in 2006, Archive-It is a subscription-based web archiving service that allows organizations to collect and preserve their digital content. According to the “Learn More” page on their website, Archive-It currently serves over 400 partner organizations including college and university libraries, state libraries, museums and art libraries, federal institutions and NGOs, and public libraries.\textsuperscript{14}

Archive-It allows partners to capture and manage web published materials including websites, blogs, and social media sites. It could be used to capture postings on “dynamic” sites such as Twitter, but it seems to work best when used to capture postings on “static sites.” Compared with the public Wayback Machine, a free service provided by the Internet Archive, the subscription based Archive-It brings many important benefits. For example, the public Wayback Machine is just one huge collection holding all archived materials, the user having very little control over the collected data, whereas Archive-It allows the partner to build their own focused collections based on themes and types of contents. Archive-It gives full control over the scope of the collection and frequency. It comes with full-text search functionality, content cataloged with metadata, and complete access control. The archived data are duplicated with copies residing at Internet Archive data centers on multiple servers. The service works with partners who may wish to have redundant copies for preservation.

\textsuperscript{13} 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.

\textsuperscript{14} Archive-It, ”Learn More”, retrieved at https://archive-it.org/learn-more/.
More importantly, the technical support for Archive-It is excellent and crucial, especially for those who just start out.

In short, because it offers a package solution to our collecting, cataloging, access, and preservation needs, we chose Archive-It as the tool of choice for archiving Chinese blogs.

2. Using Archive-It: Issues and Solutions
Anyone can learn and use Archive-It; no programming experience is required. That said, it does have a learning curve and needs some training and practice in order to use it effectively. For prospective partners, the Archive-It hosts twice-monthly webinars that provide an overview and demonstrations of the product. The webinar consists of a PowerPoint presentation that thoroughly discusses the features, customizable options of the product, archiving needs of the trainee, as well as some sample work done by other institutions. We would highly recommend starting with the training webinar for anyone unfamiliar with the tool.

Next comes the experimentation phase. It took us about three weeks to thoroughly familiarize ourselves with the tool. Eventually we came up with a three-step process that worked fine for us.

Step One: Identify the target material.

Finding a target blog to archive seems an easy step. However, it became an issue for us almost as soon as we started. The Archive-It is designed to crawl a specific website or page (via its URL, or “seed” as it is called in the Archive-It), at a frequency designated by the user. For our project, since we only collect postings related to the theme of “anti-corruption”, it became obvious to us that we could not use the default approach, crawling the same site or page and archiving whatever may appear on the page at the moment. Instead, we opted for a “curation” approach, searching and finding the relevant content and then archiving that article only. Fortunately, the Archive-It does give the option to allow us to just crawl one page at a time and save that page.

In order to find relevant content to archive, we built a Google Blog Search Engine that searches over 50 major blogging sites and news sites in China. We also came up with a list of keywords on various themes of the anti-corruption campaign as well as names of key people and high profile cases. Every week, we conduct the same searches via the special search engine and identify target articles for archiving. The graduate assistant who worked on the project was a Ph.D. student in political science specialized in contemporary Chinese politics. So we have high confidence in the quality of data collected.

Step Two: Collecting Data

This probably is the most challenging part for people who never used the tool before. It will
be confusing and frustrating at times. For example, even if we find the right content and feed the “seed” (URL) to the Archive-It, the Archive-It may or may not like it. Not all crawls will be successful. Also, the Archive-It automatically collects all the “documents” including any advertisements, extraneous links and images that happen to be on the page. If not careful, you may quickly use up the storage capacity that you have purchased. We would recommend starting with one page at a time, and always use “test crawl” first before permanently archiving a file.

After you get to know the process well and feel comfortable using it, you may increase crawling up to three or five “seeds” at time. As the collect grows, it is important to keep track of the collected documents and avoid duplication. That is because the production mode of Archive-It only gives you a record number and partial URL for each document, and it is difficult to know which record corresponds to which document. Once the collection grows to more than 100, it becomes increasingly difficult to remember whether a certain blog has already been archived or not. For this reason, we also recommend using a word document or spreadsheet to keep track of the entire document URLs and their titles, dates of their crawling, and whether the crawl was successful.

Step Three: Creating Metadata

One nice feature of Archive-It is the ability to create our own metadata specific to the collection that we are trying to build. For each record collected, there is a meta form to tag the record. These metadata will then be used on the public site to facilitate browsing. We decided that we would use “subject,” “date,” “type,” language,” “collector,” and “source” as our main categories. Under “subject,” we used various descriptors that pertain to the data we are collecting. For example, we knew we wanted to collect data from both the national level and local levels; we wanted our data to cover both high profile cases (so-called “tigers”) and smaller cases (“flies”); and we wanted our collection to focus on government corruption, but also include cases from major industries such as banking, energy, education and so on. Accordingly, we came up with subject descriptors that reflected the target contents that we tried to collect. In the end, when browsing our collection online, one can either search for the content they are interested in, or click on the corresponding subject filters on the menu bar and easily access the materials.

3. Outcome

The official project start date was July 1, 2015. It happened to be in the summer and most students were out. Recruiting a qualified student assistant proved to be difficult. We did not hire a student assistant until October 2015, three months after the project kicked off. Because of the delay, plus time for training, we were only able to collect about 300 records in our collection. Then in the second year, we were in full production mode and added about 1,000 more entries to our collection. In addition to the metadata, each entry also has both
Chinese title and its English translation, and a brief description in English of the content. The entire collection is full text searchable, and easily browsable through subject filters. Below is a (partial) screenshot of the first page of the collection.

To facilitate discovery and access, we created a catalog record and added to the JHU library catalog (See below). Anyone who searches the online catalog with keywords such as “China social media” or “China corruption” will be able to find the record and link to the collection.
Although the project term comes to an end, because JHU has an institutional subscription to the Archive-It, we will continue to monitor the anti-corruption campaign in China and add new entries to the collection in the years to come.

III. SOCIAL FEED MANAGER (SFM) AND SINA WEIBO COLLECTION

Collecting social media data from social media platforms such as Twitter and Sina Weibo are governed by the affordances of the Application Programing Interfaces (APIs). In the case of Sina Weibo, those affordances are quite limited. The Sina Weibo API supports two methods that are relevant for the topical collecting of Weibo posts: the “friends timeline” method which is available from the basic API and the “topic search” method which is available from the advanced API. The “friends timeline” returns a subset of the Weibo entries posted by a user and the other users followed by the user. The “topic search” method returns a subset of tweets matching a topic. A topic is a single keyword bracketed by “#” in the text of a Weibo, allowing users to generate a folksonomy on the Sina Weibo platform, similar to the role of hashtags on the Twitter platform. (The API methods are explained in greater details in a companion piece to this article. The companion piece covers API-based social media collecting and the Sina Weibo platform and API in greater detail.)

1. Collecting social media data with SFM
Step One: Identify the target material.

Like collecting blogs, the first step in collecting data from Sina Weibo is to identify target material. For a collection built from the “friends timeline” method this involves identifying a list of users that post about a relevant theme. For a collection built from the “topic search” method this involves identifying a list of topics that are used to mark Weibo posts that are relevant to the theme.

In both these cases, the most straightforward way to identify users or topics is via the Sina Weibo website (http://Weibo.com). First, in order to view full content of the Sina Weibo website, one must be a registered user. Second, the selection of accounts is based on their role, accountability and relevancy to a specific topic. We generally follow three types of accounts: government accounts, media accounts and private accounts. Government accounts and media accounts with a label of “官方微博 Official Weibo” indicate that Sino Weibo has authenticated their identity. A private account with “微博认证 Weibo certification)” or a label of V(IP) could prove its accountability. In terms of identifying accounts with a focus on Anti-Corruption Campaign, government and media accounts post about diverse topics, of which the anti-corruption campaign may be only one of the topics. For example,澎湃新闻 is a newly created online news site which offers timely and relatively objective reports; one of its foci is exposing corrupt officials and their cases. Private accounts, however, must be highly focused on the anti-corruption campaign and have a respectable number of active followers. In the meantime, private accounts also play an important role in revealing potential corruption cases and individuals. Third, identifying these accounts requires professional knowledge in contemporary China’s politics and familiarity with Chinese media, government, and Sina Weibo platforms. Our China Studies librarians who possess the knowledge and skills first identified the most well-known government and media accounts. The second step is to identify more accounts based on the identified accounts’ posts. The most effective way is to check most authoritative accounts’ Weibo. If an account is tagged or its Weibo is being reposted by the authoritative accounts, it could be a potential candidate. In terms of identifying topics, some of them were found from hashtags that appeared in the identified accounts’ Weibo. The others need librarians’ knowledge on Anti-Corruption through reading scholarly publications and news.

For the project, initial work identifying topics and users relevant to the anti-corruption campaign in China was performed by Jing Zhong (former China Studies Librarian, George Washington University) and updated by her successor, Yan He. Thirty-two users were identified. (For privacy reasons, the list of users will not be disclosed here.) Forty-five topics were identified. These include 中央电视台, 举报, 党内监督, and 党纪国法. For various reasons, we were not able to collect all of the users and topics, e.g., some accounts were closed and some topics were no longer being used.
Step 2: Create a Sina Weibo account and get API credentials

Prior to beginning collecting Weibos, using the “friends timeline” method requires creating a “dummy” Sina Weibo account that is to be used exclusively for collecting. That account should follow only the users from which you want to collect Weibos. Using a personal account is not recommended.

Like other social media platforms, Sina Weibo requires applications such as SFM to have credentials to access the API. In turn, SFM requires users to acquire those credentials and supply them to SFM. Since the process of getting credentials from Sina Weibo can be confusing, we have documented the process to reduce the barrier for users at other institutions. Fortunately, once the credentials are acquired, the process of providing those credentials to SFM is via a web form and is a simple process.

Step 3: Create a collection

SFM is designed to allow users to directly create and administer their own collections without mediation. Thus, to create a collection, a user completes a web form supplying the name of the collection, a description of the collection, a collecting schedule, and some other options.

---

For a topics collection, the user must also provide the topic. (Like web harvesting, these are referred to as “seeds” in SFM.) The collection can then be turned on, which initiates harvesting in accordance with the schedule.

Here is part of the page showing the anti-corruption campaign timeline collection on GW’s SFM instance:
Step 4: Wait

SFM will repeatedly perform harvests in accordance with the schedule. In the case the anti-corruption campaign timeline collection, this is quite often -- every 30 minutes -- to try to
collect as many Weibo posts as possible. As SFM is proceeding with harvesting, users can review progress. Below is a list of harvests:

<table>
<thead>
<tr>
<th>Type</th>
<th>Requested</th>
<th>Updated/Completed</th>
<th>Status</th>
<th>Stats</th>
<th>Messages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weibo timeline</td>
<td>May 12, 2017, 10:23:51 a.m. EDT</td>
<td>May 12, 2017, 10:24:00 a.m. EDT</td>
<td>Success</td>
<td>24 weibos</td>
<td></td>
</tr>
<tr>
<td>Weibo timeline</td>
<td>May 12, 2017, 9:23:51 a.m. EDT</td>
<td>May 12, 2017, 9:24:00 a.m. EDT</td>
<td>Success</td>
<td>27 weibos</td>
<td></td>
</tr>
<tr>
<td>Weibo timeline</td>
<td>May 12, 2017, 8:23:51 a.m. EDT</td>
<td>May 12, 2017, 8:24:00 a.m. EDT</td>
<td>Success</td>
<td>23 weibos</td>
<td></td>
</tr>
<tr>
<td>Weibo timeline</td>
<td>May 12, 2017, 7:23:51 a.m. EDT</td>
<td>May 12, 2017, 7:24:00 a.m. EDT</td>
<td>Success</td>
<td>25 weibos</td>
<td></td>
</tr>
<tr>
<td>Weibo timeline</td>
<td>May 12, 2017, 6:23:51 a.m. EDT</td>
<td>May 12, 2017, 6:24:00 a.m. EDT</td>
<td>Success</td>
<td>51 weibos</td>
<td></td>
</tr>
</tbody>
</table>

and here is a diagram of the number of Weibo posts collected over the last month:

Step 5: Adjust users and topics

Sina Weibo is a very dynamic environment. To account for this, Sina Weibo should be monitored for new users to follow or topics to collect. (The reverse is true as well: it may be desirable to remove users or topics.) Adding a user is done by following the user from the
"dummy" account on the Sina Weibo website. To add a topic, a new topic collection is added in SFM.

2. Using Weibo data collected with SFM
There are a number of mechanisms for accessing social media data collected by SFM. However, by far the easiest and most common is to export it to a spreadsheet.

| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S |
| 1 | created_at | weibo_id | screen_name | followers_count | friends_count | reply_count | reblogs_count | favorite_count | topic | in_reply_to weibo_url | text | url | retweeted | retweeted_by | parent_weibo_url | parent | parent_reposted_url |

Spreadsheets are a flexible way of sharing with a scholar, as the Weibo data can be directly viewed or loaded into analytic software.

3. Comparison to web archiving with Archive-It
It is worth drawing attention to some of the salient differences between collecting web resources such as blogs or articles and collecting Weibo posts from Sina Weibo, as they have slightly different natures. First, collecting of Weibo posts is possible for current posts only; the ability to get historical Weibo posts is extremely limited. Web archiving, on the other hand, has a much larger collection window. While the life of a web page is notoriously short, it is measured in months not the minutes in which a Weibo can be collected from the API.

Second, collecting of social media data is done on an ongoing basis. For example, once a collection is started for a topic, that harvest is repeated over a period of time. By comparison, web harvesting of blogs or articles is generally a one-time activity. The curator of a web harvesting collection must regularly be looking for new blogs or articles to harvest. (One interesting approach that was not explored in this project would be to attempt to identify new blogs or articles from the links contained in harvested Weibo posts).

4. Outcome
The SFM and associated Weibo harvester work have made a number of contributions that we believe move the field forward in significant ways.

First, the SFM Weibo harvester is, to our knowledge, the first of its kind; the enhancements to SFM and creation of the Sina Weibo harvester were developed and documented with a broader goal of ensuring that they can be set up and run by other cultural heritage...
institutions interested in building collections of Sina Weibo data. After much work resolving a series of technical barriers and other impediments, GW’s SFM team first released the Sina Weibo harvester on May 31, 2016. Initially it only supported the “friends timeline” method, with support for the “topic search” method added on March 14, 2017. The SFM and Weibo harvester code is public and open-source, so that others looking to utilize and build on it have full transparency into its methodology, and can more easily build upon work to date.

Secondly, we have publicly documented a number of lessons learned in using the Weibo API for this type of work to help guide others seeking to embark on related work. This includes our creation of instructions for obtaining Weibo API credentials, and our writings on some of the challenges in working with the Weibo API and our solutions to those challenges.

Finally, we have used both of the above to successfully create the first Sina Weibo collection. The Sina Weibo collection was started on June 29, 2016. Since then, over 110,000 Weibo posts have been collected across all collections. We look forward to scholarly interest in analyzing the Weibo posts.

IV CONCLUSION

Since the start of the project in 2014, we have received strong support from library colleagues and administrators, faculty and graduate students from the three participating universities and beyond. Over the past two years, the collaborators from the three institutions, including East Asian Studies librarians, IT and administrative personnel, met multiple times to discuss the planning and implementation of the project. Despite extra work being added to our regular workload, all three librarian collaborators welcomed the opportunity to venture outside the realm of traditional librarianship and learn and use new technology to help develop new types of collections. And our effort has paid off. We have managed to build a digital collection of millions of Weibo posts and over a thousand blog articles, and helped to preserve a valuable part of Chinese history. In addition, we hope our project will inspire more similar projects in the field of East Asian Studies librarianship and beyond, to collect and preserve ephemeral online materials for future studies.