A Collaborative Web Archiving Project on Chinese Social Media and Anti-Corruption Campaign

Yunshan Ye (East Asian Studies Librarian, Johns Hopkins University)

Justin Littman (Software Developer and Librarian, George Washington University)
Acknowledgement

- Ms Ellen Hammond, PI of Mellon Foundation –CEAL Innovation Grant Program, and all members of the grant review committee
- Library staff and management, faculty and grad students (professors Erin Chung, Joel Andreas of JHU; Andrew Wedeman of Georgia State U; Jackson Woods, then PhD candidate at GW)
The (final) Team

- Yunshan Ye (Project Lead), Johns Hopkins University
- Ding Ye, Georgetown University
- Cathy Zeljak, George Washington University
- Justin Littman, George Washington University
- Daniel Kerchner, George Washington University
- Yan He, George Washington University
- Chunman Zhang (student assistant), Johns Hopkins University
- Yecheng Tan (student assistant), George Washington University
What We Accomplished

* Built two digital archival collections of Chinese web and social media content on Chinese anti-corruption campaign (2012-);
* Enhanced software (SFM) with capacity to capture and store Chinese Sina Weibo content and able to capture and store social media data in other East Asian languages as well;
* Documented and published the project’s technical implementation, challenges and solutions to help future similar projects (JEAL Oct.2017);
* Presented and promoted our work at national conferences (ACRL 2017, CEAL 2016/2018).
Chinese Social Media and the Anti-Corruption Campaign (2012-)

Collected by: Johns Hopkins University

 Archived since: Sep, 2015

 Description: This archive is part of a collaborative web archiving project by Johns Hopkins University, George Washington University, and Georgetown University. It is funded by the Mellon Foundation-Council on East Asian Libraries Innovation Grants for East Asian Librarians. The goal of the project is to preserve Chinese blogs and micro-blogs related to the ongoing Chinese Anti-Corruption Campaign (2012-).

 Subject: Blogs & Social Media, China, Anti-Corruption, Social Media

 Creator: Milton S. Eisenhower Library, Johns Hopkins University

Narrow Your Results

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Enter search terms here

Sites  Search Page Text

Page 1 of 13 (1,300 Total Results)

Sort By: Title (A-Z)  Title (Z-A)  URL (A-Z)  URL (Z-A)

Title: 原河南省副省长吕德彬生命的最后18天/ The last 18 days of former Henan province vice governor Lv Debin.

URL: http://bbs.tianya.cn/post-no06-136171-1.shtml

Description: This article discloses the secret life of the former vice governor of Henan province Lv Debin.
Re: Chinese Social Media and the Anti-Corruption Campaign (2012-) Online in Chinese

This archive is part of a collaborative web archiving project by Johns Hopkins University, George Washington University, and Georgetown University. It is funded by the Mellon Foundation Council on East Asian Libraries Innovation Grants for East Asian Librarians. The goal of the...

Online Access

www.archive-it.org
### Chinese Social Media and Anti-corruption Campaign (2012+)

#### The Project

This project is a joint effort and collaborations by [John Hopkins University libraries](https://www.hopkinslibrary.jhu.edu/), the [George Washington University libraries](https://library.gwu.edu/), and [Georgetown University libraries](https://library.georgetown.edu/). It is funded by a Mellon innovation grant administered by the Council on East Asian Libraries to use and adapt web archiving tools - [Social Feed Manager](https://socialfeedmanager.org/) and [Archive It](https://archive-it.org/) to build two collections of Microblogs (Sina Weibo) and blogs respectively. The content is focused on reaction to and commentary about the current Chinese government’s anti-corruption campaign from within China.

#### The Chinese Blog Collection at JHU

**Scope**: Using [Archive It](https://archive-it.org/), a web archiving tool from [Internet Archive](https://archive.org/), we collected 13,000 blog posts covering from 1991 to 2017. The blog posts came from both national and regional blogging sites, addressing corruption cases and issues regarding various government branches and industries, including different levels of government administrations, military, energy, education and so on.

**Access**: Users can find the link to the collection via the [JHU online catalog](https://catalog.jhu.edu/) or go to the [Archive It](https://archive-it.org/) site. For any questions or assistance, please contact Mr. Yunshan Ye at yye@jhu.edu.

#### The Chinese Weibo Collection at GW

**Scope**: Dynamic social media sites such as Weibo requires more specialized software tools. [Social Feed Manager (SFM)](https://socialfeedmanager.org/), open-source software developed by the [George Washington University Libraries](https://library.gwu.edu/), with the support of grants from the Institute for Museum and Library Services and the National Historical...
Significance (1): Why Social Media

Articles still online as of March 8, 2018

- 72% Online
- 28% Offline
Significance (2): Why Anti-corruption campaign

* Anti-corruption campaign (2012-): So far
  * 1.53 million CCP members disciplined
  * 278,000 prosecuted, including 440 ministerial or provincial officials, 43 Central Committee members (about 11.4% of that body)
  * 13,000 military officers dismissed; more than 50 generals imprisoned for corruption
Significance (cont.): consequently
Web Archiving at GW

- Archiving Weibo via API
- Social Feed Manager (SFM)
- The anti-corruption campaign collection
西安国际港务区

#民生服务# 【中央机关公开遴选选调360名公务员】2017年中央机关公开遴选和公开选调公务员工作今日开始报名。此次公开遴选和公开选调共有56个中央机关参加，计划选拔360名公务员。

报名时间截止5月22日18:00。笔试时间为2017年6月25日，考试地点设在北京、上海、西安、兰州等17个城市。详情...

展开全文
【中央机关公开遴选选调360名公务员】2017年中央机关公开遴选和公开选调公务员工作今日开始报名。此次公开遴选和公开选调共有56个中央机关参加，计划选拔360名公务员。[心]报名时间截止5月22日18:00。[心]笔试时间为2017年6月25日，[心]考试地点设在北京、上海、西安、兰州等17个城市。详情 ，

"source": "<a href="http://app.weibo.com/t/feed/6ghA0p" rel="nofollow">搜狗高速浏览器</a>"，
"user": {
  "screen_name": "西安国际港务区",
  "name": "西安国际港务区",
  "location": "陕西",
  "followers_count": 22224,
  "created_at": "Mon Mar 24 13:42:11 +0800 2014"
},
"comments_count": 1,
APIs & Web data

That data is combined with presentation information, using a Content Management System (CMS), API, and/or other tools... Publicly accessible

...to produce a website, which combines content data with presentation information.

Behind most websites is content stored as structured data, not visible to the public.

That data—or a subset or version of it—can be directly queried if the platform has a public API.

Often, people use another application—possibly a third party tool—to interact with the API.

...to get web content as pure, structured data.
“friends timeline” method
- Returns last 150 weibos of the user and user’s friends.
- Part of basic API

“topic search” method
- Returns most recent 200 weibos matching a topic, e.g., “#keyword#” or “#你好#”
- Part of advanced API
Social Feed Manager

- Open source software developed by George Washington University Libraries
- Supports Twitter, Tumblr, Flickr & Sina Weibo.
- For researchers, faculty, students, and archivists.
- Lower the barriers to collecting social media data:
  - User-friendly website to create, manage & export collections
  - Handles technical aspects of working with APIs

http://go.gwu.edu/sfm
SFM: Create collection

- Add Twitter user timeline
  Tweets from specific accounts

- Add Twitter search
  Recent tweets matching a query

- Add Twitter filter
  Tweets in real time matching filter criteria

- Add Twitter sample
  A subset of all tweets in real time

- Add Tumblr blog posts
  Blog posts from specific blogs

- Add Flickr user
  Posts and photos from specific accounts

- Add Weibo timeline
  Posts from a user and the user's friends

- Add Weibo search
  Recent Weibo posts matching a query
SFM: Describe collection

Add Weibo timeline
* indicates a required field.

Collection name*

Description

Credential*
CNWeibo's weibo credential

Image sizes
- Thumbnail
- Medium
- Large

For harvesting images, select the image sizes.

Incremental harvest
Only collect new items since the last data retrieval.
## SFM: Harvesting

### Weibo_Anti-Corruption_Followers

- **Weibo timeline**

---

### Harvests (20,980)

<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>

[View all 20,980 harvests](#)
<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
<th>R</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>created_at</td>
<td>2016-07-05</td>
<td>399402098</td>
<td>人民网</td>
<td>31512740</td>
<td>5000</td>
<td>11</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，他们来讨生活了！</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
</tr>
<tr>
<td>2</td>
<td>2016-07-05</td>
<td>399401771</td>
<td>人民网</td>
<td>45903532</td>
<td>1853</td>
<td>613</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，他们来讨生活了！</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
</tr>
<tr>
<td>3</td>
<td>2016-07-05</td>
<td>399401621</td>
<td>人民网</td>
<td>31512740</td>
<td>5000</td>
<td>268</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，他们来讨生活了！</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
</tr>
<tr>
<td>4</td>
<td>2016-07-05</td>
<td>399401242</td>
<td>人民网</td>
<td>31512740</td>
<td>5000</td>
<td>50</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，他们来讨生活了！</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
</tr>
<tr>
<td>5</td>
<td>2016-07-05</td>
<td>399401125</td>
<td>人民网</td>
<td>45903532</td>
<td>1853</td>
<td>247</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，他们来讨生活了！</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
</tr>
<tr>
<td>6</td>
<td>2016-07-05</td>
<td>399400865</td>
<td>人民网</td>
<td>31512740</td>
<td>5000</td>
<td>229</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，他们来讨生活了！</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
</tr>
<tr>
<td>7</td>
<td>2016-07-05</td>
<td>399400425</td>
<td>人民网</td>
<td>31512740</td>
<td>5000</td>
<td>197</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，他们来讨生活了！</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
</tr>
<tr>
<td>8</td>
<td>2016-07-05</td>
<td>399400487</td>
<td>人民网</td>
<td>31512740</td>
<td>5000</td>
<td>374</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，他们来讨生活了！</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
</tr>
<tr>
<td>9</td>
<td>2016-07-05</td>
<td>399399984</td>
<td>人民网</td>
<td>31512740</td>
<td>5000</td>
<td>85</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，他们来讨生活了！</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
</tr>
<tr>
<td>10</td>
<td>2016-07-05</td>
<td>399399950</td>
<td>南方都市</td>
<td>8430838</td>
<td>449</td>
<td>879</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，他们来讨生活了！</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
</tr>
<tr>
<td>11</td>
<td>2016-07-05</td>
<td>399399574</td>
<td>人民网</td>
<td>45903532</td>
<td>1853</td>
<td>666</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，他们来讨生活了！</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
</tr>
<tr>
<td>12</td>
<td>2016-07-05</td>
<td>399399481</td>
<td>人民网</td>
<td>31512740</td>
<td>5000</td>
<td>85</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，他们来讨生活了！</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
</tr>
<tr>
<td>13</td>
<td>2016-07-05</td>
<td>399399463</td>
<td>红色腰带</td>
<td>4532</td>
<td>2168</td>
<td>0</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，他们来讨生活了！</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
<td><a href="http://m.w">http://m.w</a></td>
<td>大雨倾盆，</td>
</tr>
</tbody>
</table>
Anti-corruption collection by date

June 29, 2016 - May 10, 2017: 91,584 weibos
## Anti-corruption collection by topic

<table>
<thead>
<tr>
<th>Topic</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>纪检监察</td>
<td>545</td>
</tr>
<tr>
<td>反腐倡廉</td>
<td>457</td>
</tr>
<tr>
<td>2017两会</td>
<td>417</td>
</tr>
<tr>
<td>2017看两会</td>
<td>402</td>
</tr>
<tr>
<td>北电学生举报教授</td>
<td>401</td>
</tr>
<tr>
<td>央视新闻微直播</td>
<td>380</td>
</tr>
<tr>
<td>关注里约奥运</td>
<td>373</td>
</tr>
<tr>
<td>微博看两会</td>
<td>294</td>
</tr>
<tr>
<td>北电学生实名举报女教授</td>
<td>264</td>
</tr>
<tr>
<td>一起看奥运</td>
<td>245</td>
</tr>
</tbody>
</table>
Challenges of the Weibo API

- Few methods available for collecting weibos.
- Poor documentation
- Difficult to get credentials, especially for Advanced API.
- For security reasons, some parts of Sina Weibo website are blocked.

Special thanks to Victor Tan.
Thanks!

* Yunshan Ye (Johns Hopkins University)
  yye@jhu.edu
* Justin Littman (George Washington University)
  justinlittman@gwu.edu ; @justin_littman.