

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)

Acknowledgement (cont.)

- * 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).

The Archive It Collection at JHU



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

Subject

Sort By: **Count** | [\(A-Z\)](#)

[Tigers \(362\)](#)
[Flies \(178\)](#)
[Military \(73\)](#)
[Guangdong \(66\)](#)
[Hebei \(57\)](#)

[More ▼](#)

Date

Sort By: **Count** | [\(A-Z\)](#)

[December 2015 \(80\)](#)
[September 2016 \(68\)](#)
[February 2017 \(56\)](#)
[December 2016 \(55\)](#)
[November 2015 \(52\)](#)

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.

Sites

Search Page Text

Page 1 of 13 (1,300 Total Results)

[Next Page ►](#)

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.

Discovery tool(1): Catalog Record

Catalog+Articles **Catalog** Articles Reserves

Any Field Search

Show only items available online
[Advanced Search](#)

Refine your search

- Format >
- Library Location >
- Publication Year >
- Author >
- Organization >
- Language >
- Subject >
- Region >

You searched for:

[Start Over](#)

« Previous | 1 - 10 of 474 | Next »

Sort by relevance ▾

10 per page ▾

1. Chinese Social Media and the Anti-Corruption Campaign (2012-) Online in Chinese

Bookmark

2015 .

[Political corruption — China](#). [Political corruption — China — Prevention](#). [Corruption — China — Prevention](#). ...

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

Discovery Tool (2): LibGuide page

Home

Find Articles

Find Books

East Asian Sources

Primary Sources in English

English/Translated Journals
from Asia

Student Research Help

Faculty Research Help

Local Digital Collections

Chinese Social Media and Anti-corruption Campaign (2012-)

The Project

This project is a joint effort and collaborations by [John Hopkins University libraries](#), the [George Washington University libraries](#), and [Georgetown University libraries](#). It is funded by a Mellon innovation grant administrated by the Council on East Asian Libraries to use and adapt web archiving tools - [Social Feed Manager](#) and [Archive It](#) to build two collections of Microblogs (Sino 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](#), a web archiving tool from [Internet Archive](#), we collected 13,00 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.

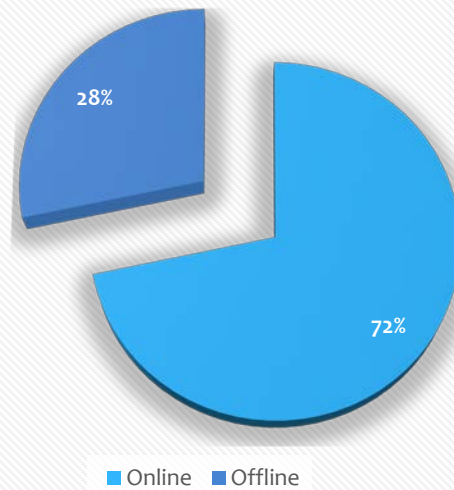
Access: Users can find the link to the collection via the [JHU online catalog](#), or go to the [Archive It](#) site. For any questions or assistance, please contact Mr. Yunshan Ye at yye@jhu.edu.

The Chinese Weibo Collection at GW

Scope: Dynamic social media sites such as Weibo requires more specialized software tools. Social Feed Manager (SFM), 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

Significance (1): Why Social Media

Articles still online as of March 8, 2018



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

A Weibo



西安国际港务区  

今天 15:43 来自 搜狗高速浏览器

#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](#) 



 收藏

 转发

 1

 1

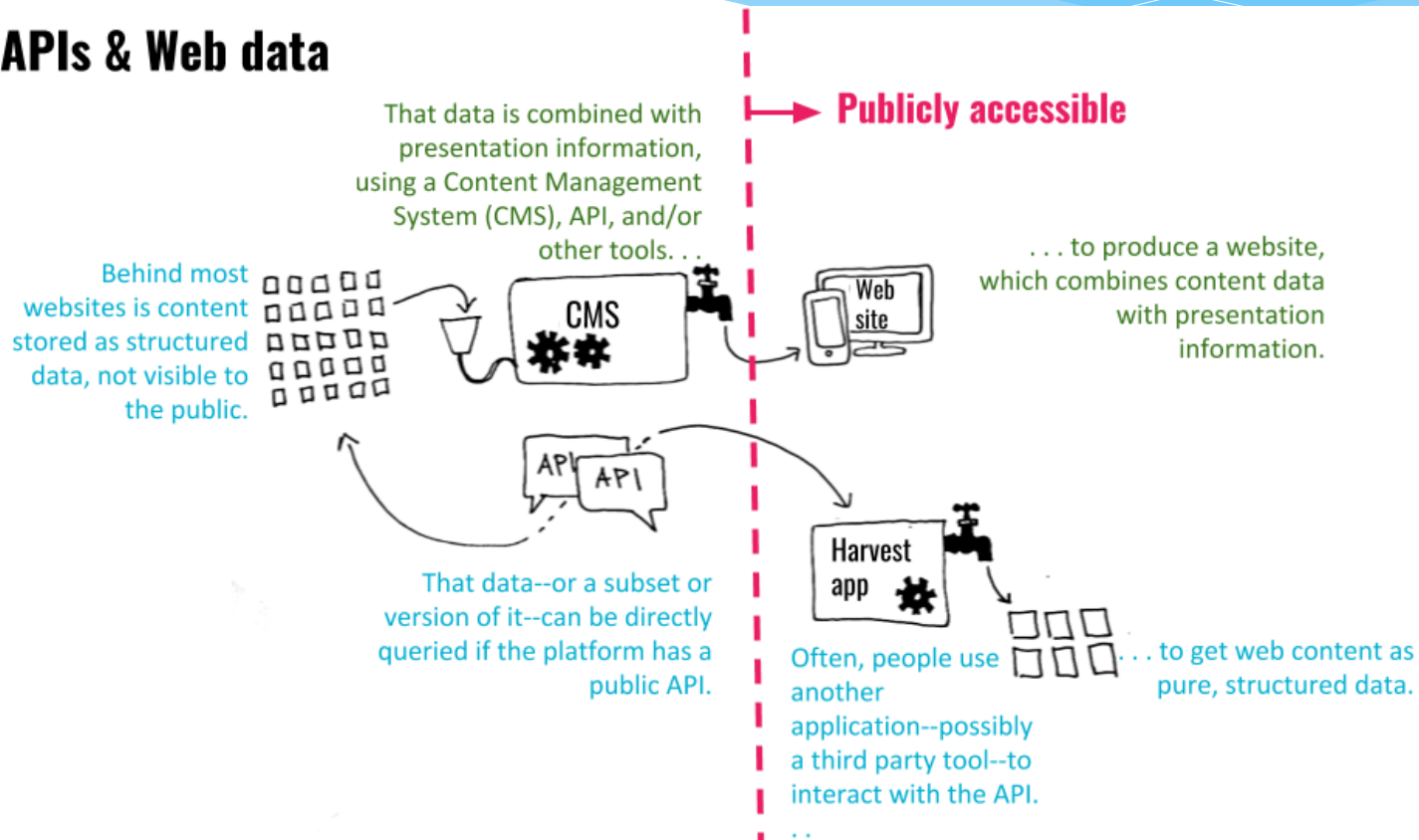
A Weibo retrieved from the API

```
{
  "created_at": "Thu May 11 15:43:13 +0800 2017",
  "id": 4106245138832598,
  "text": "#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。[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,
}
```

Full Weibo: <https://gist.github.com/justinlittman/db0280402a2ded54cc3d539a798c16e0>

API

APIs & Web data



Affordances of the Weibo API


- * “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 Collection 

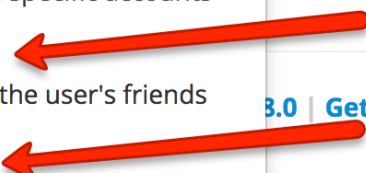
- Add Twitter user timeline**
Tweets from specific accounts
- Add Twitter search**
Recent tweets matching a query
- Add Twitter filter**
Tweets in real time matching filter criteria
- Add Twitter sample**
A subset of all tweets in real time
- Add Tumblr blog posts**
Blog posts from specific blogs
- Add 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

6, 1:07:59 p.m.

016, 11:49:14

3.0 | [Getting Started](#)

GW L



SFM: Describe collection

Add Weibo timeline

* indicates a required field.

Collection name*

Description

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

 Turn on

 Edit

 Export

Harvests (20,980)

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

[View all 20,980 harvests](#)

SFM: Export

Weibo_Anti-Corruption_Followers



Weibo timeline

Collection is active. Turn off to edit.

Turn off

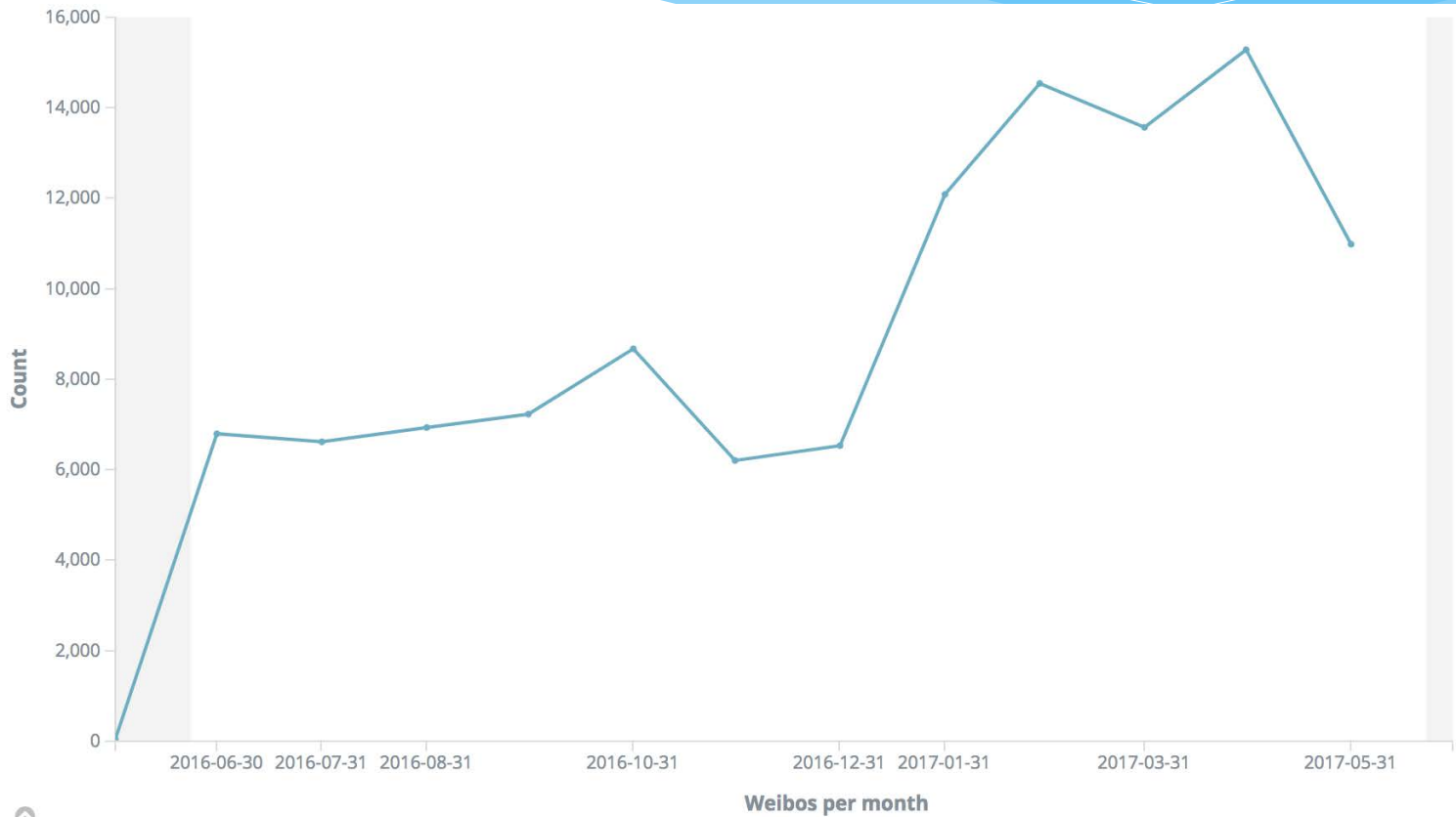
Edit

Export



	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_nar	followers_c	friends_cou	reposts_co	topics	in_reply_to	weibo_url	text	url1	url2	retweeted_retweeted	retweeted_url2					
2	2016-07-05 399402098	人民网	31512740	5000	11				http://m.w	【今晚，致敬这位男篮老将】今晚，北京奥体中心，斯坦科维奇杯赛前，现场举行了王治郅国家队退役仪									
3	2016-07-05 399401771	人民日报	45903532	1853	613				http://m.w	【微议录：致敬，愿平安！】@[致敬！抗洪抢险中，那些无所畏惧、最可爱的人]大雨滂沱，他们来									
4	2016-07-05 399401621	人民网	31512740	5000	268				http://m.w	【最常见的工伤：胖！[衰]】因为忙，乱吃早餐成虚胖；因为忙，久坐电脑大腿粗；因为忙，应酬多啤酒									
5	2016-07-05 399401242	人民网	31512740	5000	50				http://m.w	【母女蹲】http://t.cn/R5ugBqD									
6	2016-07-05 399401125	人民日报	45903532	1853	247				http://m.w	【人民微评：太湖不是垃圾桶】：【震惊！】http://t.cn/R5u6DZp									
7	2016-07-05 399400865	人民网	31512740	5000	229				http://m.w	#随手转发，宝贝回家# 【[话筒]急转寻人！辽宁13岁女孩走失已2天】胡若彤，女，13岁。7月3日9时许，									
8	2016-07-05 399400487	人民网	31512740	5000	197				http://m.w	【浙江丽水】http://t.cn/R5uDtrA									
9	2016-07-05 399400425	人民日报	45903532	1853	1222				http://m.w	【注意！】http://t.cn/R5n0nI9									
10	2016-07-05 399399984	人民网	31512740	5000	374				http://m.w	【武警救】http://t.cn/ http://t.cn/R5uWVoo									
11	2016-07-05 399399950	南方都市报	8430938	449	879	喜感新闻			http://m.w	#喜感新闻# 昆明一男子为追女孩，开着兰博基尼拖了一卡车牛奶在大学校园公然求爱，据说表白的对象									
12	2016-07-05 399399574	人民日报	45903532	1853	666				http://m.w	【[话筒]急转寻人！辽宁13岁女孩走失已2天】胡若彤，女，13岁。7月3日9时许，从辽宁朝阳市双塔区文									
13	2016-07-05 399399481	人民网	31512740	5000	85				http://m.w	【80岁“最”http://t.cn/R53yKGk									
14	2016-07-05 399399463	红色稻恶	4532	2168	0				http://m.w	继续转发！			多么漂亮，	http://t.cn/R53z9sr					

Anti-corruption collection by date



June 29, 2016 - May 10, 2017: 91,584 weibos

Anti-corruption collection by topic

Topic	Count
纪检监察	545
反腐倡廉	457
2017两会	417
2017看两会	402
北电学生举报教授	401
央视新闻微直播	380
关注里约奥运	373
微博看两会	294
北电学生实名举报女教授	264
一起看奥运	245

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.