

UC Irvine
LAUC-I and Library Staff Research

Title

Web Archiving: Preserving the Umbrella Movement

Permalink

<https://escholarship.org/uc/item/8b3893ch>

Author

Tsang, Daniel C

Publication Date

2015-09-16

WEB ARCHIVING: PRESERVING THE UMBRELLA MOVEMENT

Daniel C. Tsang, Distinguished Librarian
Data Librarian & Politics, Economics & Asian
American Studies Bibliographer, UC Irvine
dtsang@uci.edu

Prepared for presentation & revised after presentation
at University of Hong Kong Library
16 September 2015

DEDICATION



This talk is dedicated to my dad, Kenneth, a former HKU medical student, shown in this inset from a HKU Medical Society 1937 group photo currently mounted on the new HKU MTR Station wall.

Photo source: HKU Communications and Public Affairs Office.

WEB ARCHIVING: DEFINITION

“Web archiving is the process of collecting portions of the World Wide Web to ensure the information is preserved in an archive for future researchers, historians, and the public. Web archivists typically employ web crawlers for automated capture due to the massive size and amount of information on the Web.”

Source: https://en.wikipedia.org/wiki/Web_archiving Wikipedia

INTERNET ARCHIVE: ARCHIVE-IT

[HOME](#)[EXPLORE](#)[LEARN MORE](#)[CONTACT US](#)

The leading web archiving service
for collecting and accessing
cultural heritage on the web
Built at the Internet Archive



Welcome to Archive-It!
Attend a live informational webinar and demo
to learn more about the service

Contact Us to sign up for an upcoming session:
Jun 02 2015, 11:30 AM PDT
Jun 16 2015, 11:30 AM PDT

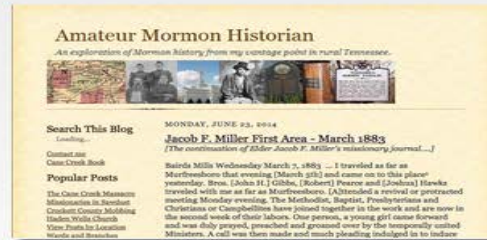
Explore Collections

[Show All Collections](#)

Smithsonian Institution Websites

By Smithsonian Institution

Over 200 websites archived related to Smithsonian museums, galleries, and programs.



Mormon Blogs Collection

By Brigham Young University

Features the lifestyle and culture of Mormons through self published blogs.



Virginia's Political Landscape, 2010

By Library of Virginia

By Library of Virginia

A collection of Web sites that document Virginia's 2010 Congressional elections (primary and general). All 11 members of Virginia's Congressional Delegation were...

Explore Collecting Organizations

[Show All Organizations](#)

SELECTION CRITERIA

According to Jinfang Niu (University of South Florida):
“Existing web archiving efforts use the following selection criteria to determine what to preserve: domain (such as .gov or .edu), topic or event, media type and genre. Many European countries archive the web in their country domain. The library of the NASA Goddard Space Flight Center (GSFC) captures pages in the Goddard domain...”

Source: Jinfang Niu. “An Overview of Web Archiving.” D-Lib Magazine (March/April 2012).

INCLUSION CRITERIA...

“The Library of Congress has created various event-based web collections, such as the September 11, 2001 web archives, the election web archives and the Iraq War 2003 web archives...

“Media-type based selection includes or excludes certain media types. The Goddard library, for example, avoids crawling large video files and software products ...The web archiving project conducted by Chirag Shah and Gary Marchionini (2007), on the other hand, focused on preserving election videos on Youtube. Some web archives select based on genres such as blogs, newspapers, virtual worlds, etc. The National Library of France created a web collection of e-diaries...The Internet Archive has a software archive and an archive of videogame videos...”

Source: Jinfang Niu. “An Overview of Web Archiving.” D-Lib Magazine (March/April 2012).

IIPC: COLLECTION DEVELOPMENT POLICIES



IIPC netpreserve.org | INTERNATIONAL INTERNET PRESERVATION CONSORTIUM

HOME ABOUT IIPC **WEB ARCHIVING** PROJECTS MEMBER ARCHIVES EVENTS FOR MEMBERS

COLLECTION DEVELOPMENT POLICIES

IIPC members

- > [Bibliothèque nationale de France](#)
- > [Library of Congress - 2013](#)
- > [British Library - 2014](#)
- > [The National Archives \(UK\) Records Collection Policy and Operational Selection Policy 27: UK Government Web Estate](#)
- > [National Library of Finland - 2011](#)
- > [Portuguese Web Archive](#)
- > [Swiss National Library](#)
- > [Austrian National Library](#)
- > [Columbia University Libraries](#)
- > [Stanford University Libraries](#)

Source: <http://netpreserve.org/collection-development-policies>

IIPC COLLECTION DEVELOPMENT: OTHER INSTITUTIONS

Other institutions

- > Tamiment Library, NYU - 2010
- > Bentley Historical Library - 2011
- > North Carolina State Government Website Archives and Access Program
- > University of Texas San Antonio
- > University of Alberta Library
- > University of California Los Angeles (UCLA)
- > Chesapeake Digital Preservation Group

Source: <http://netpreserve.org/collection-development-policies>

STANFORD UNIVERSITY

Collection development

Our collection development guidance is intended to fulfill the following objectives:

- complement discipline-specific collection development policies;
- help curators decide what and, more importantly, what *not* to collect; and
- ensure that comparatively limited web archiving resources are deployed only for the most valuable content.

Focus on at-risk content

All web content is in some sense at-risk; this is, in fact, the [raison d'être for web archiving](#). Particular categories of web content are more at-risk, however, because they are of time-limited interest or purpose, subject to government censorship, disseminated by immature organizations, or for other reasons. Spontaneous events, including [disasters](#), [revolutions](#), and [trending social topics](#) may briefly occupy the public spotlight, then fade from view. This unique and ephemeral content is especially deserving of our attention.

Source: <http://library.stanford.edu/projects/web-archiving/collection-development>

STANFORD UNIVERSITY...

Complement existing collecting strengths

We have collecting strengths in particular areas, reflected by the research we support, our staffing for different subjects, our Special Collections, our relationships with donors and alumni, our geography and our institutional history. We provide added value when we consider web archiving as a potential component of a broader collecting plan and create web archives to complement other extant and prospective collections.

Observe resource constraints

A format-agnostic collection development policy will more than likely designate a broader range of web content as in scope for collecting than is practically feasible, given available web archiving resources. We should be mindful of collection dimensions that are most likely to increase costs. This includes not just the number of nominated websites but also their complexity (i.e., demanding additional staff time for crawl configuration and quality assurance) and contents (i.e., large files like video balloon storage requirements).

STANFORD...

Consider what others are collecting

We are a member of an international community whose collective goal is collecting, preserving, and providing access to the historical web. Considering the cumulative and growing volume of information that has ever existed on the Web, even our aggregated efforts represent but a small fraction. We should therefore strive to identify existing web archives that overlap with areas where we intend to archive the Web ourselves and minimize duplication of effort....

Web archive holdings are not documented systematically, in terms of subject area, temporal coverage, language, top-level domain, or other identifiers, though research is underway that should simplify this. In the meantime, places to consult to discover existing web archives include: Archive-It's collections portal, ... the International Internet Preservation Consortium's list of member archives, the Wikipedia List of Web archiving initiatives, the Internet Archive Wayback Machine, and the UK Web Archive Memento aggregator service. Curators may often learn about and/or contribute to planned web archives through their discipline-specific communities of practice. If overlap with another web archive is discovered, we should additionally consider the depth and frequency of their archiving to determine whether it is still worthwhile for us to archive it.

Source: <http://library.stanford.edu/projects/web-archiving/collection-development>

ACCESS AND VALUE

Consider the access conditions of what others are collecting

National libraries, in particular, create web archives under legal frameworks that only permit limited access (e.g., on-premise, for designated research, etc.). While generally we should avoid duplicatively archiving web content that has already been preserved by another organization, the prospect of their not making it accessible should count in favor of our archiving it, as well.

Assess value to researchers

A fundamental challenge for selecting content is that its potential utility increases over time, as the risk of change to or loss of the original content increases and the archive takes on historical context. Through their relationship with faculty and awareness of the web resources that have been vital to research within a given subject area, curators are best positioned to identify the content that matters for future research.

Source: <http://library.stanford.edu/projects/web-archiving/collection-development>

MORE FROM STANFORD UNIVERSITY



STANFORD UNIVERSITY LIBRARIES

Boiling the Ocean, Together: Web Archive Collection Development in a Global Context

Nicholas Taylor
Web Archiving Service Manager
[Digital Library Systems and Services](#)

[Chalk Talk](#)
May 12, 2014



COLLECTION DEVELOPMENT POLICY: KEY GUIDELINES

Here are some of the key selection criteria I've adapted from these policies you might include:

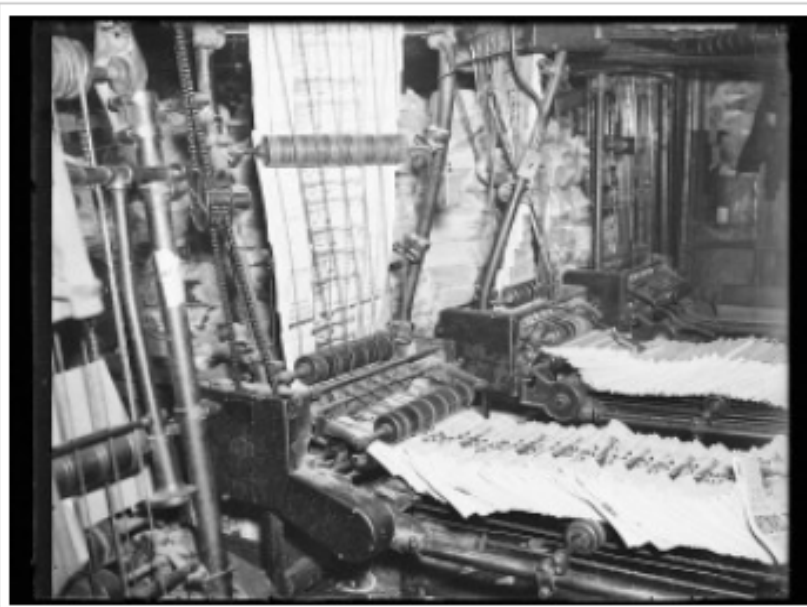
- Complement collection strengths OR weaknesses
- Focus on more at-risk online content
- Do not duplicate unless necessary
- Assess potential research value
- Assess content language
- Keep in mind resource limits
- Be cognizant of copyright issues
- Be aware of what is accessible for crawling

KEEP IN MIND RESOURCE LIMITS

Users, Use Cases and Adapting Web Archiving to Achieve Better Results

May 8, 2015 by [Butch Lazorchak](#)

*The following is a guest post from **Michael Neubert**, a Supervisory Digital Projects Specialist at the Library of Congress.*



Newspapers coming off press, [1936]. Courtesy of the Library of Congress Prints and Photographs Division.

That is large.

In a [blog post](#) about six months ago I wrote about how the Library of Congress web archiving program was starting to harvest “general” internet news sites such as Daily Kos, Huffington Post and Townhall.com, as well as newer sites such as news.vice.com and verge.com.

Many of these sites are extremely large. How large? While not an exact count (and in fact, far from it), use of the “site” limiter in Google will provide a count of digital objects found and indexed by Google (which is a far larger number than the number of web pages, but gives some sense of relative scale to other sites). A “site:huffingtonpost.com” search in Google returns “about 3,470,000 results.”

Source: *The Signal: Digital Preservation* blog, 8 March 2015

COMPLEMENT COLLECTION STRENGTHS OR WEAKNESSES

Group wants to save Occupy protesters' artwork

Lawmaker urged to ensure police don't destroy pieces that have drawn worldwide attention

Vivienne Chow
vivienne.chow@scmp.com

PUBLISHED : Friday, 31 October, 2014, 4:33am

UPDATED : Friday, 31 October, 2014, 2:10pm



The colourful Lennon Wall notes in Admiralty. Photo: EPA

An arts and culture group wants to make sure Hong Kong police do not destroy the creative works of Occupy protesters.

The group - including representatives from the Umbrella Movement Visual Archives and Research Collective, Hong Kong Shield, copyright concern group Keyboard Frontline and theatre artists - raised the issue with Ma Fung-kwok, a lawmaker representing culture and sports, in a meeting yesterday.

Artist Wen Yau, a member of the collective which wants to archive the items, said the group urged Ma to talk to the government and police to urge them not to demolish the works when they are clearing protest sites.

"We asked him to tell the police that these works are worth keeping," said Wen.

SHARE

56

Like

56

Share

8

Tweet

submit

reddit

1

Share

0

+1

2

Comments

Email

Print

RELATED TOPICS

Occupy Central

RELATED ARTICLES

HKU head defends Education Bureau's calls for investigation into Occupy co-founder Benny Tai

27 May 2015 - 6:36am

POSTER FROM ADMIRALTY PROTEST SITE, 2014



FOCUS ON MORE AT-RISK ONLINE CONTENT

中文	  Republic of Hong Kong  
Home	
Legal Bases	Proposed Hong Kong National Anthem
Bandits	Is it a great country? Yes, it is.
Democracy	Is it a great country? Yes, it is.
Our Visions	Is it a great country? Yes, it is.
Future HK	Hong Kong is really great.
Localization	Is it our great country? Yes, it is.
FAQ	Is it our great country? Yes, it is.
Criticisms	Is it our great country? Yes, it is.
Separation	I love Hong Kong.
HK Anthem	I love it forever and forever.
Join Us	I love it forever for sure.
Links	Is it our great country? Yes, it is.
Donald Tsang	I love it forever and forever!
	Why not we call ourselves a nation?
	Yes, I think we should do so for sure.
	PS: Songs for Taiwanese People and Chinese People. Click here for lyrics and melody.

DO NOT DUPLICATE UNLESS NECESSARY



學民思潮 Scholarism
@Scholarismhk

學民思潮(Scholarism)是由一群九十後組成的學生團隊,主張以社會行動介入政府施政,堅信著「立於街頭,走進人群」。「學民」表示著學生也是社會公民的一部份,學生不是「社會未來的主人翁」,乃是現在社會公民,學生絕對有權力影響政府施政。

香港
scholarism.com

TWEETS 6,979 FOLLOWING 38 FOLLOWERS 9,437

Tweets Tweets & replies Photos & videos

學民思潮 Scholarism @Scholarismhk · 3h
旺角突然起哄 警方呼籲克制
警民正在對峙 fb.me/2A3a1dPpt

學民思潮 Scholarism @Scholarismhk · 3h
愛字頭辱罵黃之鋒實錄 fb.me/1ndhdaNuA

BE AWARE OF WHAT IS ACCESSIBLE FOR CRAWLING AND BE COGNIZANT OF COPYRIGHT ISSUES



Occupy Central

Occupy Central is a civil disobedience movement which began in Hong Kong on September 28, 2014. It calls on thousands of protesters to block roads and paralyse Hong Kong's financial district if the Beijing and Hong Kong governments do not agree to implement universal suffrage for the chief executive election in 2017 and the Legislative Council elections in 2020 according to "international standards." The movement was initiated by Benny Tai Yiu-ting, an associate professor of law at the University of Hong Kong, in January 2013.



Monday, 17 November, 2014, 11:20pm

Ex-chief justice Andrew Li calls on Occupy Central protesters to retreat

Former chief justice Andrew Li Kwok-nang on Monday called on Occupy Central protesters to retreat, warning that the rule of law would be impaired if ongoing court injunctions were not obeyed.

Behind
Firewall:
SCMP

FIRECHAT APP

Source:
<http://www.theatlantic.com/technology/archive/2014/10/firechat-the-hong-kong-protest-tool-aims-to-connect-the-next-billion/381113/>

TECHNOLOGY

What Firechat's Success in Hong Kong Means for a Global Internet

The app now connecting political protesters could soon connect people in the developing world.



In an online video, a Hong Kong protester explains how the Occupy Central movement is using Firechat. (VanessaGler/YouTube)



ROBINSON MEYER | OCT 6, 2014

Look at pictures of any protest and you'll see a mix of high and low technology. The Occupy Central protests in Hong Kong are no different. [As the futurist Georgina Voss noticed](#), you'll see umbrellas to deflect tear gas cans and saran wrap to protect from pepper spray. You'll see bamboo threaded between metal barricades to strengthen them.

And you'll hear about one thing more—a piece of software protesters are downloading to their phones. It's helping them communicate digitally across the miles-long protest site, asking for supplies or reinforcements, and it stays useful even when the Internet is blocked or down. It's called Firechat.

ASSESS CONTENT LANGUAGE



[首頁](#) [信念書](#) [意向書](#) [最新消息](#) [文章](#) [媒體專訪](#) [問與答](#) [請捐款支持](#) [商討日](#) [全民投票](#) [English](#) [聯絡我們](#)

和平佔中籲港府尊重學生表達自由



Occupy
Central
web site.
Note:
English
tab

<http://oclp.hk/>

OCCUPY CENTRAL ENGLISH-LANGUAGE BLOG



[Introduction](#) [About](#) [Resources](#) [News Clippings](#) [For Journalists](#) [Get Involved](#)

[Multilingual](#)

Hong Kong protesters carry out 'yellow ribbon' march

Posted on [November 10, 2014](#)

Hundreds of pro-democracy protesters in Hong Kong have marched to the office of China's top representative in the city.

Activists are angry about a decision by China to screen candidates for Hong Kong's 2017 leadership election. They want direct talks with Beijing. [Continue reading →](#)

Posted in [Era of Peaceful Resistance](#) | Tagged [Beijing](#), [CY Leung](#), [March](#)

ABOUT US

OCLP is a nonviolent direct action movement that demands genuine universal suffrage in Hong Kong in compliance with international law, in particular one-person-one-vote and the right to run and be elected to office without unreasonable restrictions.

FOLLOW US @OCLPHK

Tweets

[Follow](#)



Occupy Central

和平佔中

@OCLPHK

15 Nov

REAL HONG KONG NEWS

THE REAL HONG KONG NEWS

The news about Hong Kong you don't get to read in world's press



[about](#) / [supplement – columnists & commentators](#)

HONG KONG GENERAL POLITICS

HK & CHINA CONFLICTS

DEMONSTRATION/PROTEST /RALLY

FREEDOM OF SPEECH AND PRESS

ELECTION/UNIVERSAL SUFFRAGE

HONG KONG EDUCATION

HONG KONG POPULATION

ABOUT

There are only two local English-language daily newspapers in Hong Kong – The Standard and The SCMP. Given the number of expatriates in this international city, this is simply not enough. Some of them who see Hong Kong as their home want to know **what is really going on in Hong Kong**, but the English dailies available in the market don't seem to reveal the truth objectively often enough...

We feel that these newspapers don't really cover the matters that [Hongkongers](#) (including the non-ethnic-Chinese community in HK) care about, and very often they write from China's perspective and the news is simply China centric. Shouldn't local papers focus more on home affairs?

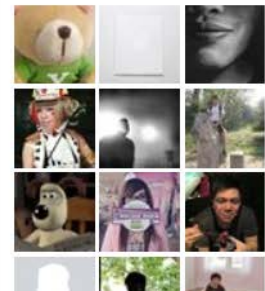
REAL HONG KONG NEWS



Real Hong Kong News



5,437 people like Real Hong Kong News.



Translated content site:
Translating from Original Cantonese or Chinese

<http://therealnewshk.wordpress.com/about/>

ASSESS POTENTIAL RESEARCH VALUE



抗命  不認命

香港已經進入抗爭的大時代，你又如何抉擇？



SocREC 社會記錄頻道
Media/News/Publishing

👍 Liked ▾

✓ Following

💬 Message



Timeline

About

Photos

Events

More ▾

PEOPLE >

184,181 likes



Post



Photo / Video

Write something on this Page...

Facebook site: <https://www.facebook.com/socrec>

SOCREC CAPTURE?




Alternative potential capture via SocRec for selective SCMP content

Monday, September 22

A cartoon timeline: Harry's View on Occupy Central | South China Morning Post

<http://www.scmp.com//news//article//1616107//cartoon-timeline-harrys-view-occupy-central>


AT UC IRVINE: POLITICAL LITERATURE WEB ARCHIVE




HOME | EXPLORE | LEARN MORE

CONTACT US

The leading web archiving service for collecting and accessing cultural heritage on the web
Built at the Internet Archive



Explore >> University of California, Irvine >> Political Literature Web Archive



Political Literature Web Archive

Collected by: [University of California, Irvine](#)

Archived since: Apr, 2015

Description: Collection of digital content of global political movements, activism, political pamphlets, images, videos and social media

Subject: [Politics & Elections](#), [Activism](#), [Alternative Media](#), [Democracy Movements](#)

Coverage: [International](#)

Language: [English](#), [Cantonese](#)

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.

Group Sort By: Count | (A-Z)

Occupy Central (3)

Subject Sort By: Count | (A-Z)

Hong Kong (9)

Democracy Movements (5)

Alternative Media (4)

Student Movements (4)

News (3)

[More ▾](#)

Creator Sort By: Count | (A-Z)

SOCIAL RECORD CO (1)

Wong Chi Fung (1)

Language Sort By: Count | (A-Z)

English (11)

Cantonese (9)

Chinese (1)

Coverage Sort By: Count | (A-Z)

International (1)

Tag Sort By: Count | (A-Z)

Hong Kong (5)

Protest movements (4)

Occupy Central Hong Kong 2014 (3)

Coffee Brands (1)

Labor unions (1)

Sites Search Page Text

Page 1 of 1 (16 Total Results)

Sort By: [Title \(A-Z\)](#) | [Title \(Z-A\)](#) | [URL \(A-Z\)](#) | [URL \(Z-A\)](#)

Title: International Political Forum
URL: <http://internationalpoliticalforum.com/>
Description: News site
Captured 3 times between Jun 16, 2014 and Aug 7, 2015
Subject: [Alternative Media](#), [News](#)
Language: [English](#)
Tag: [Protest movements](#)

Title: New Bloom Magazine
URL: <http://newbloommag.net/>
Description: "An online magazine covering youth culture and politics in the Sinosphere. We seek to put local voices in touch with international discourse, beginning with Taiwan."
Captured once on Aug 10, 2015
Subject: [Democracy Movements](#), [Taiwanese](#), [Sunflower Student Movement](#), [Student Movements](#), [Protests](#), [Anti-Textbook Revision Movement](#)
Language: [English](#), [Chinese](#)

Title: Occupy Central with Love and Peace 和平佔中
URL: <http://oclp.hk/>
Description: Web site of Occupy Central
Captured 6 times between Oct 17, 2014 and Aug 6, 2015
Subject: [Occupy Central](#), [Democracy Movements](#), [Hong Kong](#)
Group: [Occupy Central](#)
Language: [Cantonese](#), [English](#)
Tag: [Occupy Central Hong Kong 2014](#), [Hong Kong](#), [Protest movements](#)

Source:

<https://archive-it.org/collections/5609>

EXCERPT FROM LIST OF SITES

Source: <https://archive-it.org/collections/5609>

Title: Dictionary of Politically Incorrect Hong Kong Cantonese: Politically Incorrect Views from Hong Kong

URL: <https://badcanto.wordpress.com/>

Captured 5 times between Jun 16, 2014 and Aug 10, 2015

Subject: Hong Kong Identity, Hong Kong - China Relations, Cantonese

Language: English, Cantonese

Tag: Hong Kong

Title: The Real Hong Kong News | The news about Hong Kong you don't get to read in world's press

URL: <https://therealnewshk.wordpress.com/>

Description: "We want to make this blog a helpful little tool for you to "read" Cantonese newspaper without having to learn the language. We are not journalists, but a bunch of true Hong Kongers who want to protect our home: We will translate Cantonese news articles from Hong Kong's newspapers into English to help the world ..."

Captured 3 times between Oct 17, 2014 and Oct 22, 2014

Subject: Alternative Media, Translations to English from Cantonese, Hong Kong News, Activism - Hong Kong, Hong Kong, News - Hong Kong

Language: English

Title: Occupy Central 和平佔中 (@OCLPHK) | Twitter

URL: <https://twitter.com/OCLPHK/>

Description: Twitter site of Occupy Central

Captured once on Nov 15, 2014

Subject: Occupy Central, Protest Movements - Hong Kong, Hong Kong, Twitter

Group: Occupy Central

Language: Cantonese

Tag: Occupy Central Hong Kong 2014, Hong Kong, Protest movements

Title: 學民思潮 Scholarism (@Scholarismhk) | Twitter

URL: <https://twitter.com/scholarismhk/>

Description: Twitter site of student activism group, Scholarism

Captured 2 times between Nov 15, 2014 and Aug 17, 2015

Subject: Scholarism, Twitter, 學民思潮, Student Movements, Hong Kong, Umbrella Movement, Umbrella Movement

Language: Cantonese

Title: 黃之鋒 Joshua | Facebook

URL: <https://www.facebook.com/joshuawongchifung/>

Description: Facebook pages of student activist Joshua Wong

Captured 3 times between Oct 21, 2014 and Aug 17, 2015

Subject: Joshua Wong, Facebook, Student Movements, Hong Kong, Umbrella Movement

Language: Cantonese

Title: SocREC 社會記錄頻道 | Facebook

URL: <https://www.facebook.com/socrec/>

Description: Facebook site of Social Record

Captured 3 times between Oct 20, 2014 and Nov 20, 2014

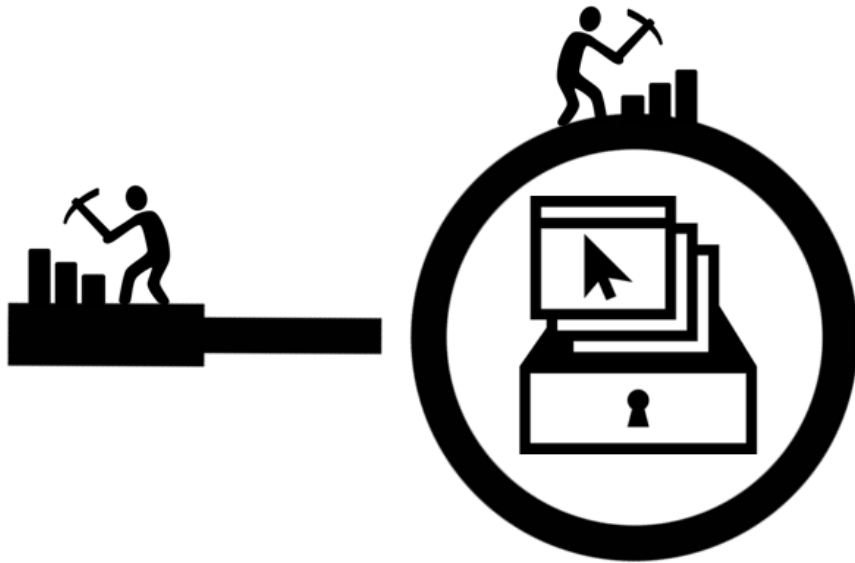
Subject: Social Movements - Archive, Hong Kong, Facebook, Umbrella Movement

Creator: SOCIAL RECORD CO

Language: Cantonese

DATA MINING WEB ARCHIVES

Web Archives as Research Datasets



Jefferson Bailey, Internet Archive
IIPC GA 2015 | Stanford/IA | [@jefferson_bail](https://twitter.com/jefferson_bail)
April 28, 2015





Archive-It Research Services

Archive-It Research Services (ARS) is expanding the ways that Archive-It partners can provide access to their archives by providing datasets extracted from partner collections. The service will enable any Archive-It partner to give users, researchers, scholars, developers, and other patrons easily-analyzed datasets that contain key metadata elements, named entities, and other data derived from the resources within their collections. Currently available datasets are listed below along with a brief description.

For more information on the service, program details, information on datasets, sample usage scenarios, and more please [visit the ARS wiki site](#).

Available Dataset Types

WAT Dataset

Web Archive Transformation files feature key metadata elements that represent every crawled resource in a collection and are derived from a collection's (W)ARC files.

LGA Dataset

Longitudinal Graph Analysis files contain a complete list of timestamped hyperlinks from every text document in an entire collection.

WANE Dataset

Web Archive Named Entities files contain a list of all of the people, places, and organizations in every text document of a collection, along with the timestamp of capture, and are derived from a collection's (W)ARC files.

(W)ARC Files

Access to (W)ARC files is automatically included for free in an Archive-It subscription. (W)ARC files contain the complete, raw content of a partner collection and the WARC format is the ISO standard preservation format for web archives. Archive-It partners are encouraged to download their (W)ARCs for local storage and redundancy.

Archive-It Research Services

Created by Jefferson Bailey, last modified on Mar 17, 2015

Archive-It Research Services (ARS) is expanding the ways that Archive-It partners can enable access to their archives by providing datasets extracted from partner collections. The service will allow any Archive-It partner to give users, researchers, scholars, developers, and other patrons easily-analyzed datasets that contain key metadata elements, link graphs, named entities, and other data derived from the resources within their collections. By supporting access in aggregate to partner archives, ARS will facilitate new types of use, research, and analysis of the significant historical records from the web that Archive-It partners are working to collect, preserve, and make accessible.

The ARS supporting documentation describes the type of datasets currently available, provides guides to requesting and acquiring these datasets, and describes some example use cases and types of analysis these datasets enable.

Why Archive-It Research Services?

This page describes the [goals](#), [objectives](#), and [origins](#) of the program.

Types of Datasets Currently Available

- **WAT: Web Archive Transformation** files feature key metadata elements that represent every crawled resource in a collection and are derived from a collection's WARC files.
 - [WAT Overview and Technical Details & WAT Example Use Cases](#)
- **LGA: Longitudinal Graph Analysis** files feature a complete list of what URIs link to what URIs, along with a timestamp, within an entire collection.
 - [LGA Overview and Technical Details & LGA Example Use Cases](#)
- **WANE: Web Archive Named Entities** uses named-entity recognition tools to generate a list of all the people, places, and organizations mentioned in each URI in a collection along with a timestamp of URI capture.
 - [WANE Overview and Technical Details & WANE Example Use Cases](#)

Service Details

This page offers [information on ARS service details](#) for current Archive-It subscribers as well as for independent researchers, patrons, and users.

Source: Archive-it Research Services Wiki:

<https://webarchive.jira.com/wiki/display/ARS/Archive-It+Research+Services>

WHAT DATA JOURNALISTS ARE DOING



Getting Data from the Web

You've tried everything else, and you haven't managed to get your hands on the data you want. You've found the data on the web, but, alas — no download options are available and copy-paste has failed you. Fear not, there may still be a way to get the data out. For example you can:

- Get data from web-based APIs, such as interfaces provided by online databases and many modern web applications (including Twitter, Facebook and many others). This is a fantastic way to access government or commercial data, as well as data from social media sites.
- Extract data from PDFs. This is very difficult, as PDF is a language for printers and does not retain much information on the structure of the data that is displayed within a document. Extracting information from PDFs is beyond the scope of this book, but there are some tools and tutorials that may help you do it.
- Screen scrape web sites. During screen scraping, you're extracting structured content from a normal web page with the help of a scraping utility or by writing a small piece of code. While this method is very powerful and can be used in many places, it requires a bit of understanding about how the web works.

With all those great technical options, don't forget the simple options: often it is worth to spend some time searching for a file with machine-readable data or to call the institution which is holding the data you want.

In this chapter we walk through a very basic example of scraping data from an HTML web page.



Web Scraping: A Journalist's Guide

By: NAEL SHIAB | August 11, 2015

Tweet 126 Like 504

Print

Do you remember when **Twitter lost \$8 billion in just a few hours** earlier this year? It was because of a web scraper, a tool companies use—as do many data reporters.

A web scraper is simply a computer program that reads the HTML code from webpages, and analyze it. With such a program, or “bot,” it’s possible to extract data and information from websites.

Let’s go back in time. Last April, Twitter was supposed to announce its trimestrial financial results once the stock markets closed. Because the results were a little bit disappointing, Twitter wanted to avoid a brutal confidence loss from the traders. Unfortunately, because of a mistake, the results were published online for 45 seconds, when the stock markets were still open.

These 45 seconds allowed a bot programmed to web scrape to find the results, format them and automatically publish them on Twitter itself. (Nowadays, even bots have scoops from time to time!)



GIJC15



**GLOBAL
INVESTIGATIVE
JOURNALISM
CONFERENCE**

9th Global Investigative
Journalism Conference,
Lillehammer, October
8th-11th 2015.

COMING EVENTS



BIRN SUMMER SCHOOL
OF INVESTIGATIVE REPORTING

NEWSLETTER



RESOURCES

WHAT RESEARCHERS ARE ALREADY DOING

The HKU Scholars Hub

The University of Hong Kong

香港大學學術庫



Title	Networked collective action in the 2014 Hong Kong Occupy Movement: analysing a Facebook sharing network
Author(s)	Fu, KW; Chan, CH
Citation	The 2nd International Conference on Public Policy (ICPP 2015), Milan, Italy, 1-3 July 2015.
Issued Date	2015
URL	http://hdl.handle.net/10722/211040
Rights	Creative Commons: Attribution 3.0 Hong Kong License

FACEBOOK CONNECTIONS (FU & CHAN, 2015)

Table 1

Online communities

Community Number	Number of pages	Name	Examples
1	338	Mainstream pro-activists	Apple Daily (mainstream media) Dash (online media affiliated with the Scholarism) USP United Social Press (online media) Keyboard Frontline (online media)
2	274	Activists	Hong Kong Federation of Student (student organization) Scholarism (student organization) Hong Kong Inmedia (online media) Occupy Central with Love and Peace (activist) Polymer (online media)
3	94	Autonomists	Passion Times (activist/online media) Hon9 Kon9 (online media) Dadazim (online media)
4	18	Environmentalist /Conservationists	Grebbish (community organization) Yue Man square (community organization) Help Tai Wei (community organization)
5	20	Irrelevant	Hong Kong Jokes
6	24	Pro-Beijing #1	Salute to Hong Kong Police (online media) Hong Kong Good News (online media)
7	11	Pro-Beijing #2	Silent Majority (online media) Support Hong Kong Police (online media)
8	14	Pro-Beijing #3	We are Chinese and proud of it (online media) One Man One Vote, Anti-Occupy (online media)

AT HKU: WEIBOSCOPE

Weiboscope

🕒 April 17, 2014

📁 Research at the JMSC

Weiboscope is a Chinese social media data collection and visualization project. One project objective, among many, is to make censored Sina Weibo posts of a selected group of Chinese microbloggers publicly accessible. Since January 2011, the system has been regularly sampling timelines of a set of selected Chinese microbloggers who have more than 1,000 followers or whose posts are frequently censored.



In year 2012, Weiboscope collected 226 million weibo posts, among which more than 10.9 million were no longer publicly accessible because of either being censored by the authorities or being deleted voluntarily by the user. The Year 2012 weibo dataset is available [here](#).

To learn more about this project, please visit its [website](#).

CROWD SOURCING FOR SELECTION INPUT

China Social Media and Anti-Corruption Web Archiving

We are soliciting your input on a collaborative web archiving project by librarians of Johns Hopkins University, George Washington University, and Georgetown University. Entitled "Blogging and Micro-blogging: Preserving Non-Official Voices in China's Anti-Corruption Campaign," this project aims at searching and preserving online social media records related to the ongoing Chinese anti-corruption campaign. The project is funded by a grant as part of the Mellon Foundation-Council on East Asian Libraries Innovation Grants for East Asian Librarians.

Please take a moment and answer the five questions below. Your assistance will be highly appreciated. Thanks!

1. When you search Weibo and other Chinese social media sites on anti-corruption posts, what personal names or political figures would you search?

2. When you search Weibo and other Chinese social media sites on anti-corruption posts, what keywords in Chinese would you use? Please list as many as possible.

3. What Chinese social media sites and/or websites would you recommend searching for posts on Chinese anti-corruption campaign?

4. Who are the most important bloggers and micro-bloggers in China, especially related to anti-corruptions and/or Chinese politics in general?

5. Do you have any other recommendations or suggestions? Please leave your contact info if you want us to get back to you.

Done

ON THE SCENE @ CUHK 22 SEPTEMBER 2014



Source: <http://www.neontommy.com/news/2014/10/what-does-protests-mean-mainland-student-reporter-hong-kong>
Photography by Sam Tsang

唔該晒 | THANK YOU VERY MUCH



SEMINAR

WEB ARCHIVING

Mr. Daniel C. Tsang

Distinguished Librarian, Data Librarian,
Asian American Studies, Political Science, Economics,
(interim) Orange County Documents Bibliographer,
University of California, Irvine

WHEN

**September 16, 2015
10:30am-12:00noon**

WHERE

**Multi-purpose Zone,
Level 3, Main Library**