Web Archiving to Excavate Hidden Collections

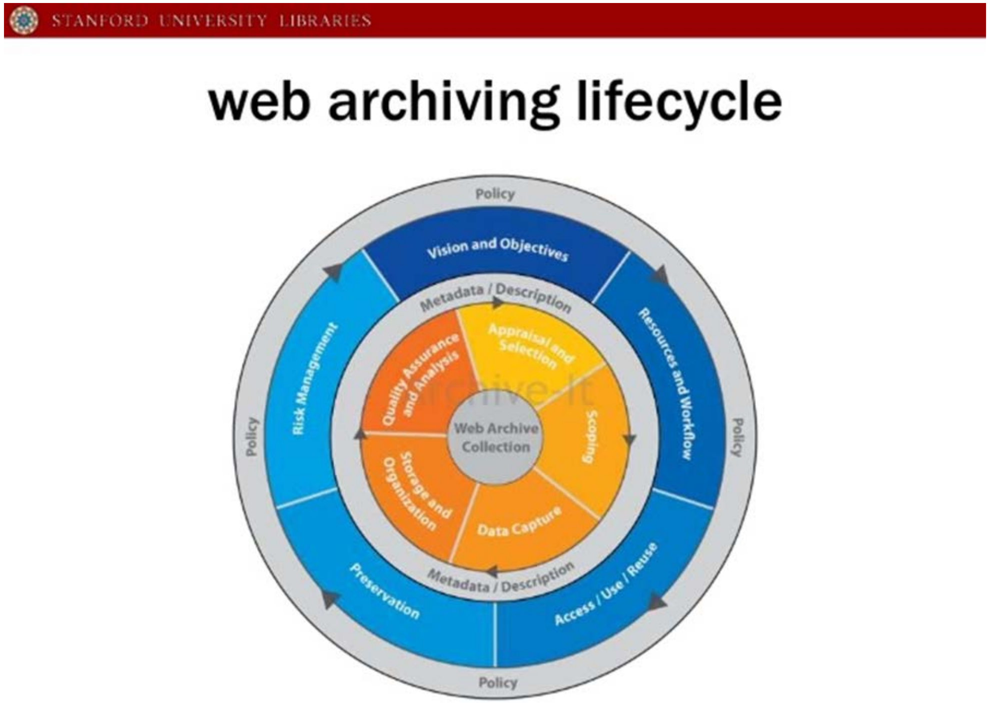
Daniel C. Tsang, Distinguished Librarian, UC Irvine

Prepared for presentation at PRDLA 2014
University of Macau
2 December 2014

Web Archiving Service
Collect, manage & preserve websites
A SERVICE OF THE UNIVERSITY OF CALIFORNIA CURATION CENTER — UC3
PRESERVE YOUR INSTITUTIONAL MEMORY AS YOUR SITES EVOLVE AND CHANGE
BUILD YOUR OWN COLLECTIONS OF ANY SIZE
CAPTURE POLITICAL & SOCIAL EVENTS AS THEY HAPPEN
SAVE AT-RISK ONLINE INFORMATION BEFORE IT IS LOST

http://webarchivingrt.files.wordpress.com/2013/05/wasimage.jpg

Web Archiving Process



STANFORD UNIVERSITY LIBRARIES

web archiving lifecycle

http://image.slidesharecdn.com/considerationsforstrategicwebarchivecollectiondevelopment-140530120129-phpapp02/95/considerations-for-strategic-web-archive-collection-development-2-638.jpg?cb=1407347520

Internet Archive

http://blogs.library.jhu.edu/wordpress/wp-content/uploads/2014/07/wayback.bmp

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).

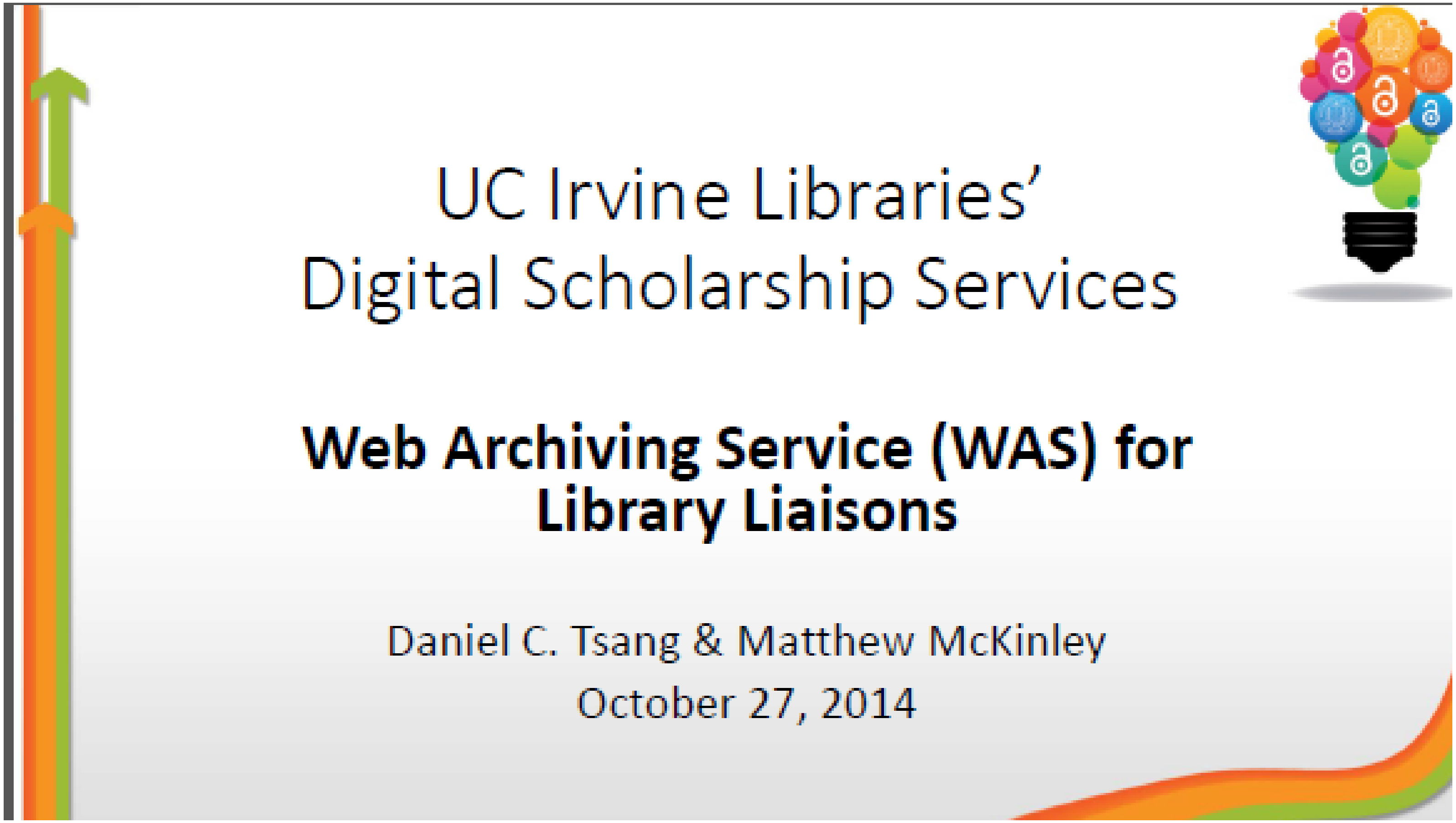
Selection Criteria …

“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).

UCI In-service Training…



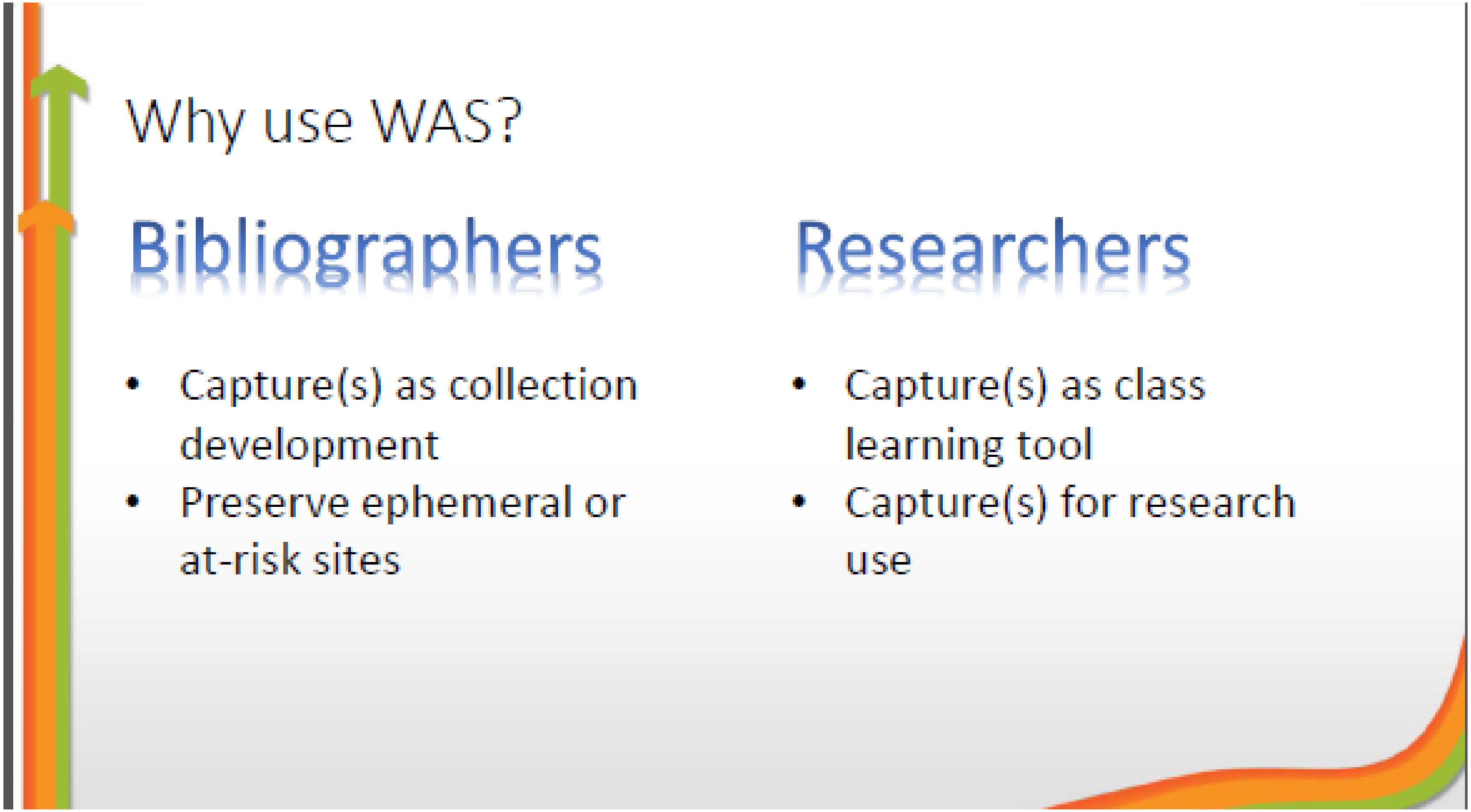
UC Irvine Libraries'
Digital Scholarship Services

Web Archiving Service (WAS) for Library Liaisons

Daniel C. Tsang & Matthew McKinley
October 27, 2014

Source: https://escholarship.org/uc/item/1sr2w7vs

Web Archiving Service for what use?



Why use WAS?

Bibliographers

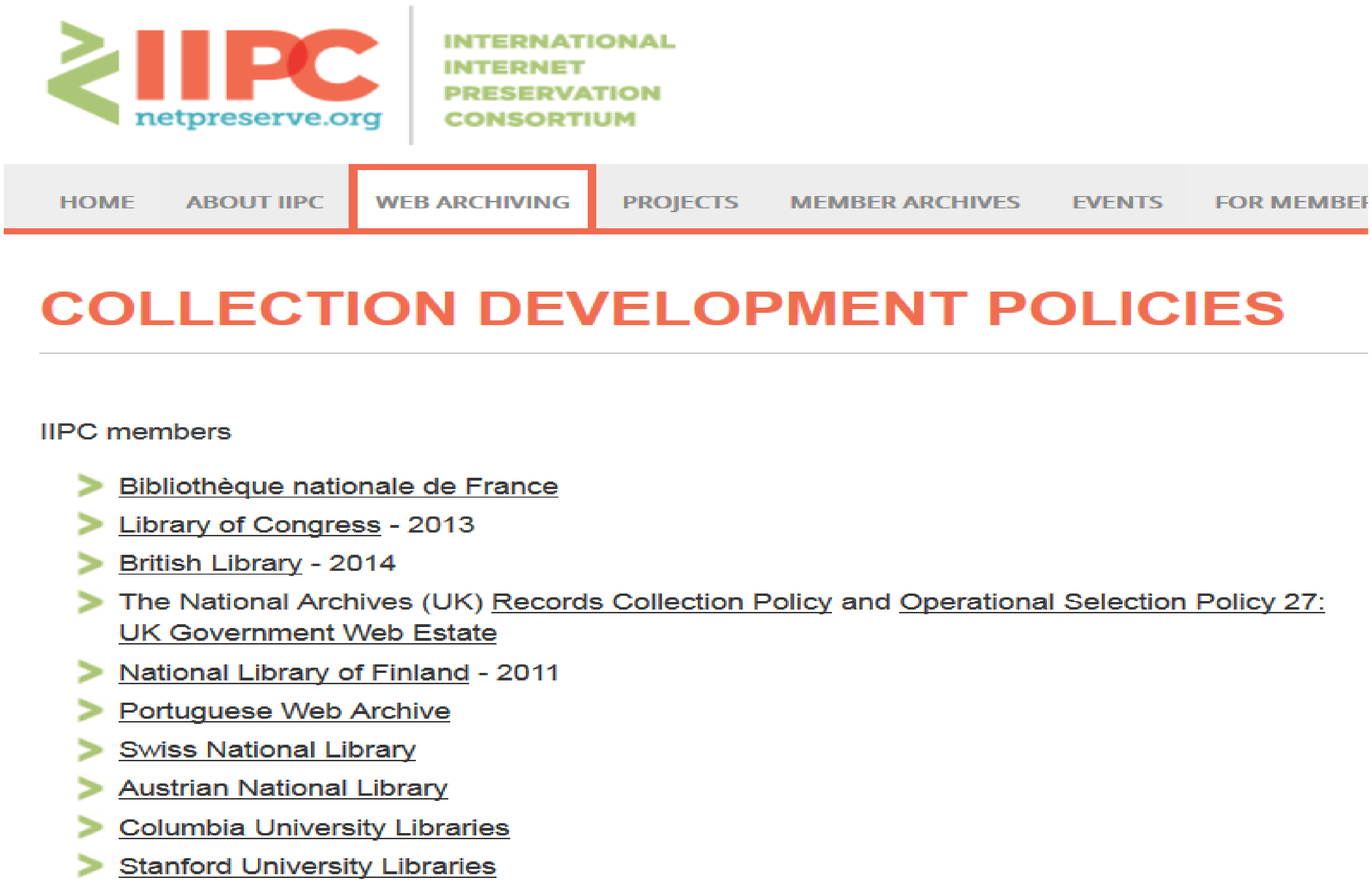
- Capture(s) as collection development
- Preserve ephemeral or at-risk sites

Researchers

- Capture(s) as class learning tool
- Capture(s) for research use

Source: https://escholarship.org/uc/item/1sr2w7vs

IIPC: Collection Development Policies



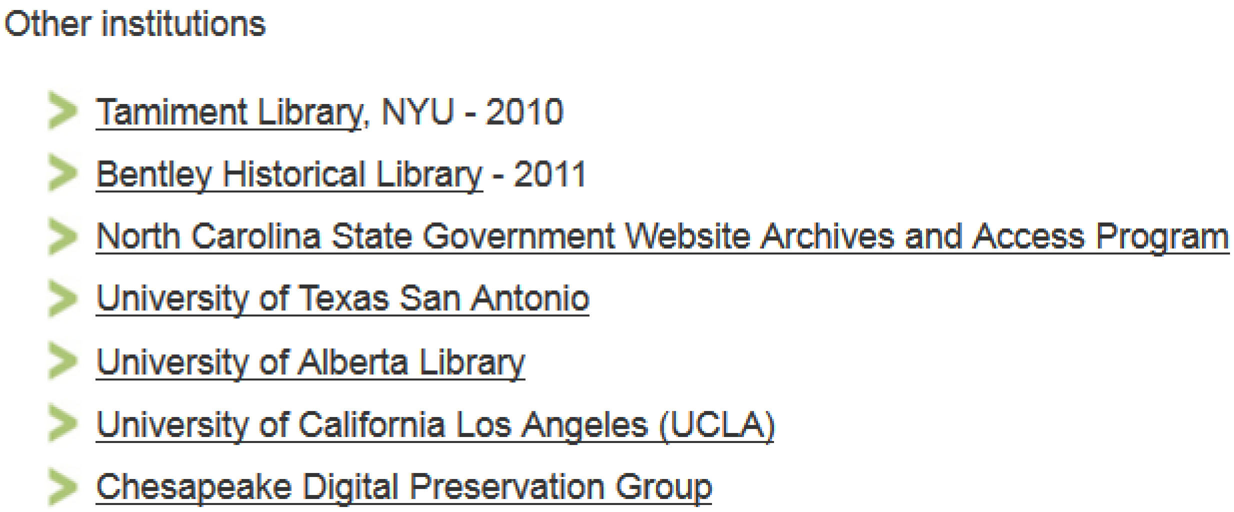
COLLECTION DEVELOPMENT POLICIES

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).

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

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. An enhancement to this approach is finding ways to provide seamless access to those external resources to our users, such as through topic guides, SearchWorks, or Memento.

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, California Digital Library's Web Archive Service 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

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

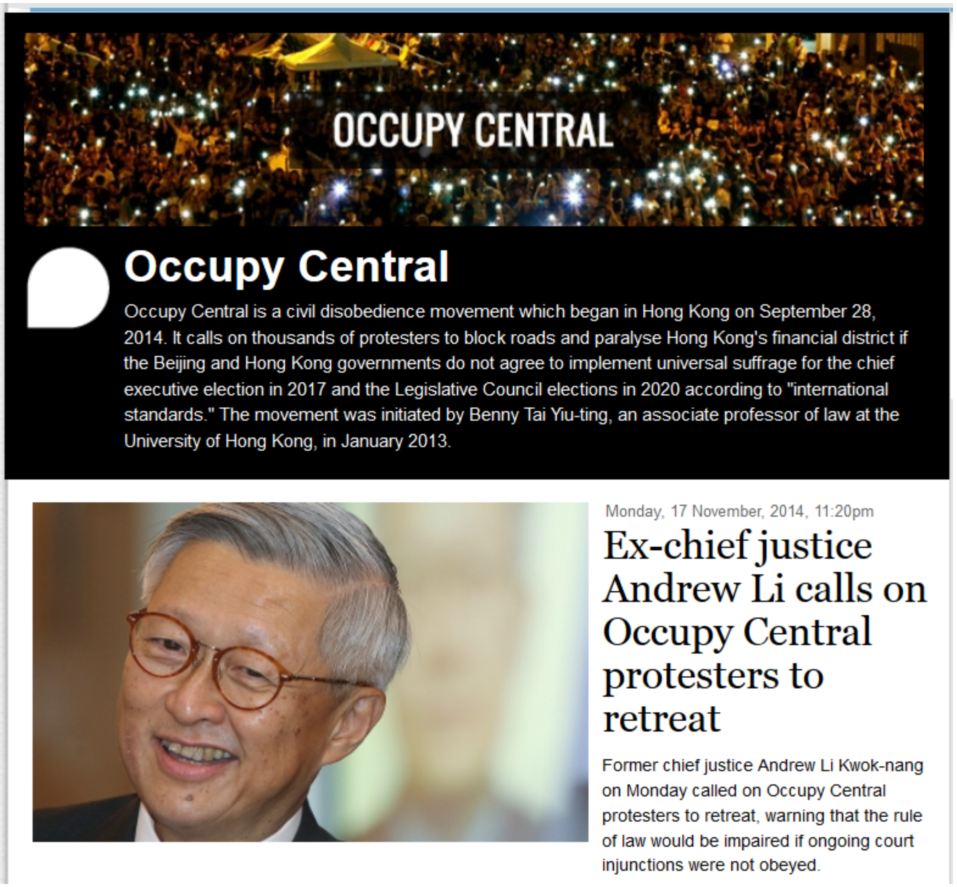
Collection Development Policy:
Key Guidelines

Here are some of the key selection criteria you might include:

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

Event Archiving - Exhibit A: *South China Morning Post*

Behind firewall?



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

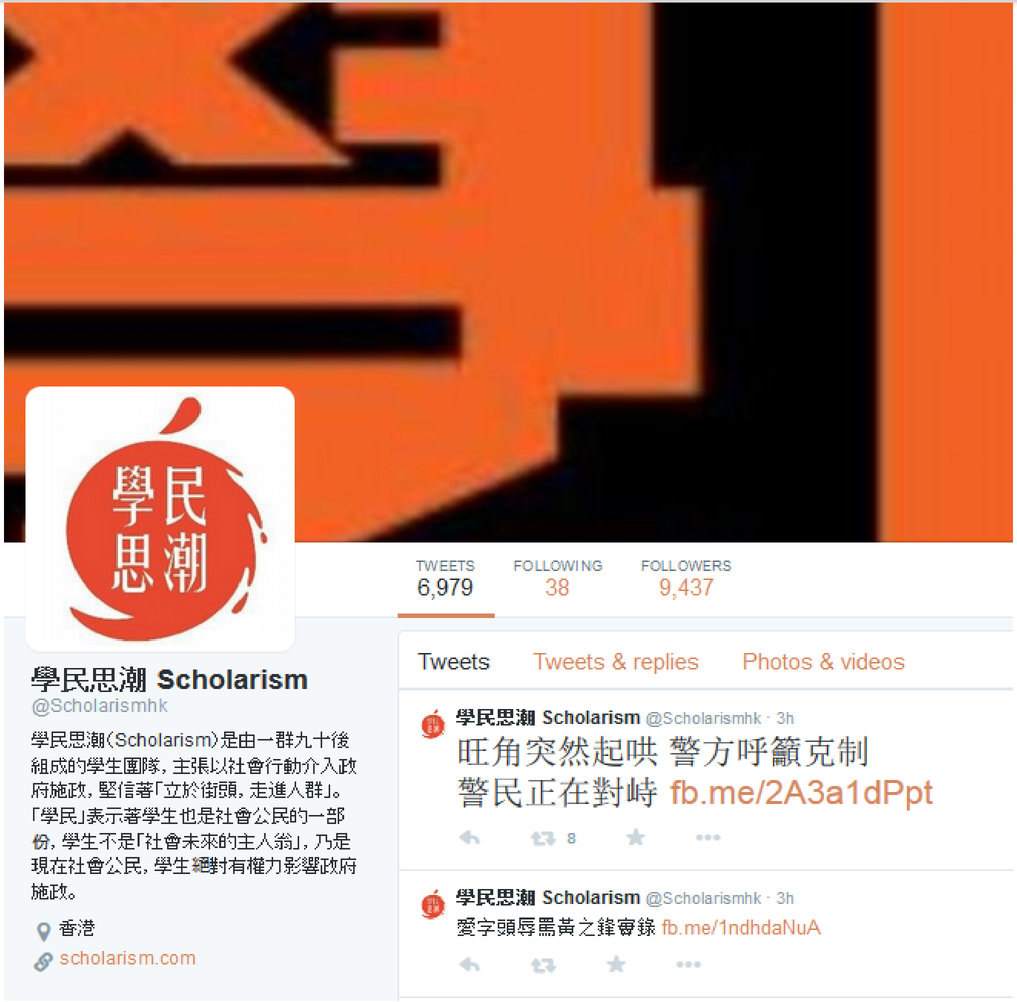
Ex-chief justice Andrew Li calls on Occupy Central protesters to retreat

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.

http://www.scmp.com/topics/occupy-central

Exhibit B: Scholarism

Twitter site



TWEETS 6,979

FOLLOWING 38

FOLLOWERS 9,437

Tweets　Tweets & replies　Photos & videos

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

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

8

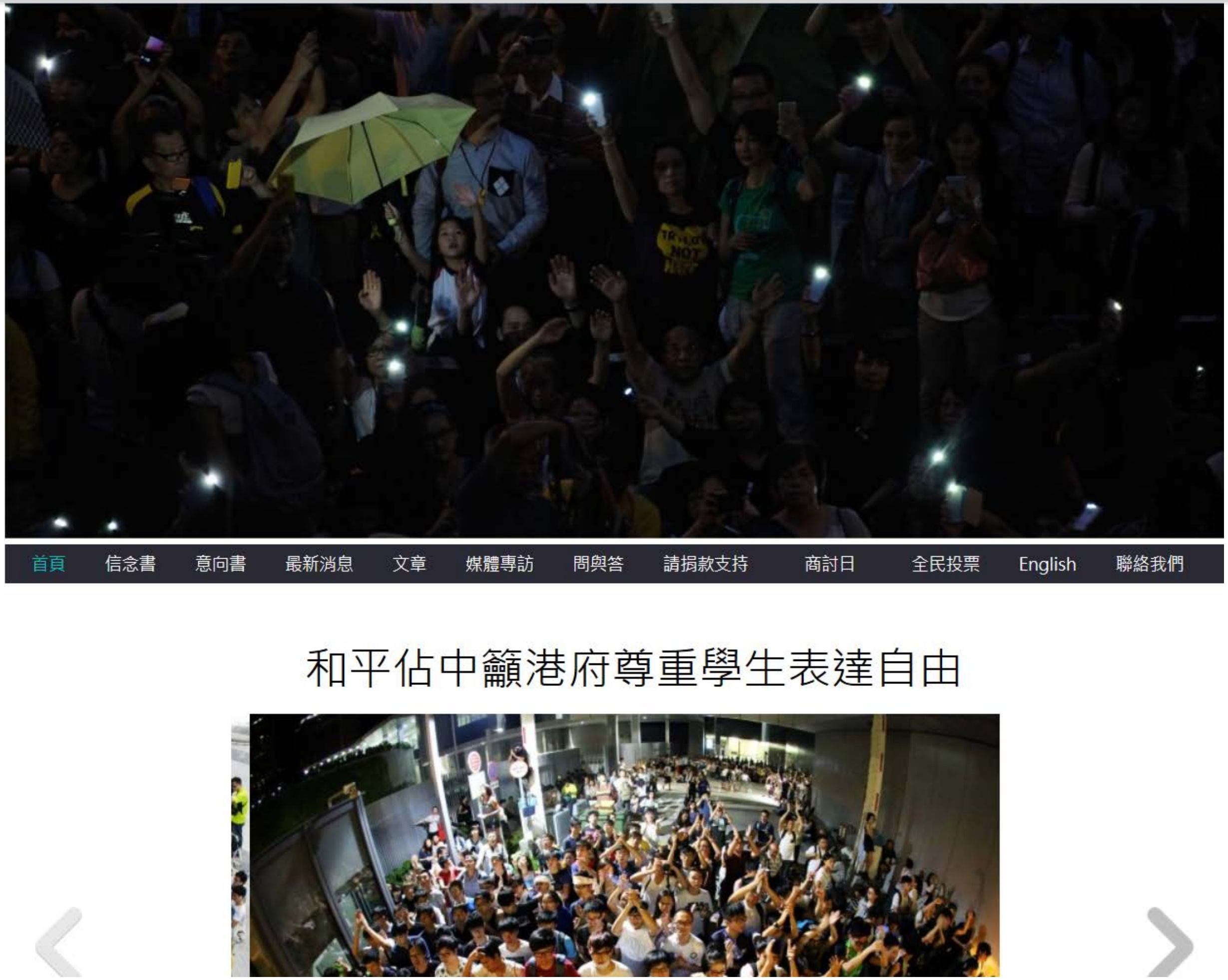
學民思潮 Scholarism @Scholarismhk · 3h

愛字頭辱罵黃之鋒審錄 fb.me/1ndhdaNuA

https://twitter.com/scholarismhk/

Exhibit C: Occupy Central

Note: English page



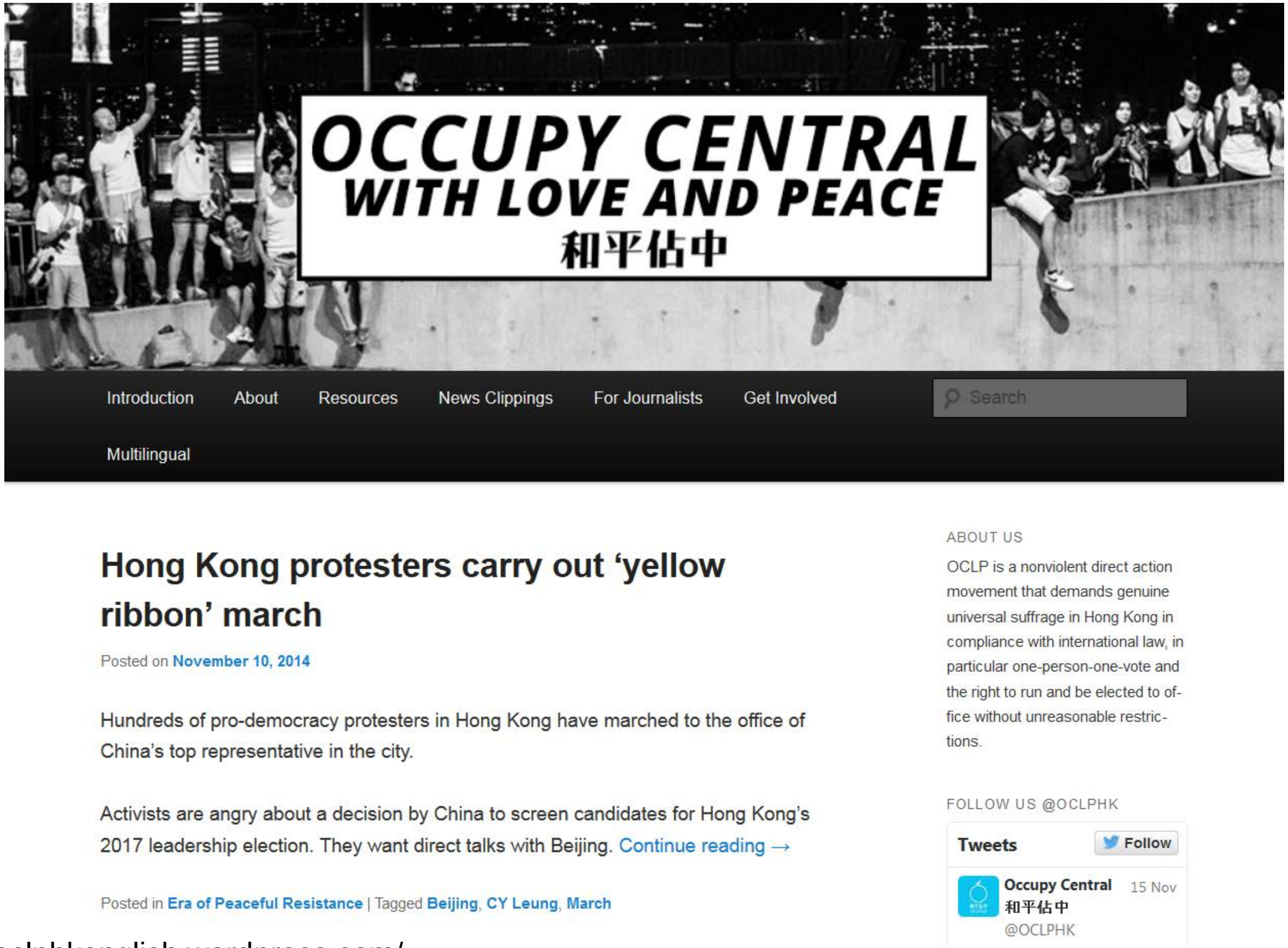
首頁 信念書 意向書 最新消息 文章 媒體專訪 問與答 請捐款支持 商討日 全民投票 English 聯絡我們

和平佔中籲港府尊重學生表達自由

http://oclp.hk/

Exhibit D: Occupy Central

Alternative site



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

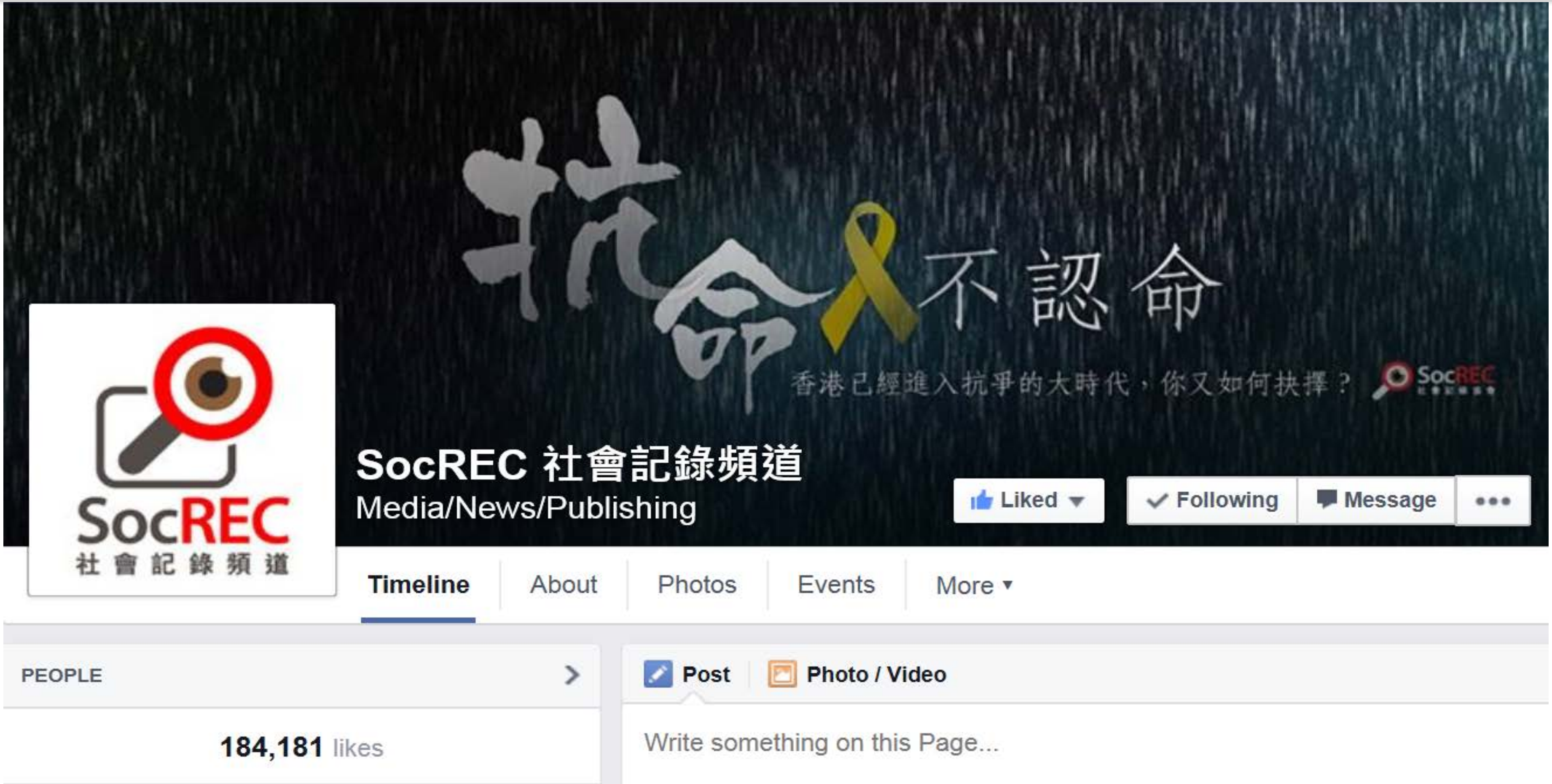
ABOUT US

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

http://oclphkenglish.wordpress.com/

Exhibit E: SocREC



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

SocREC 社會記錄頻道

Media/News/Publishing

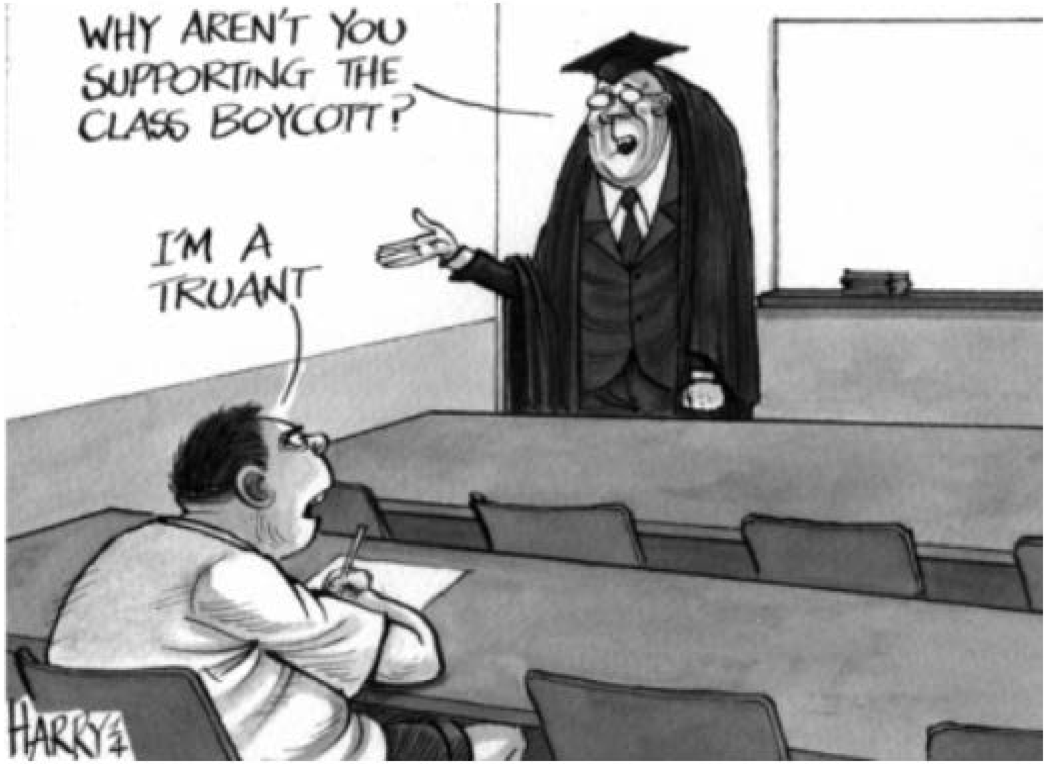
184,181 likes

Write something on this Page...

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

Exhibit F: SocRec

Alternative capture for selective SCMP content

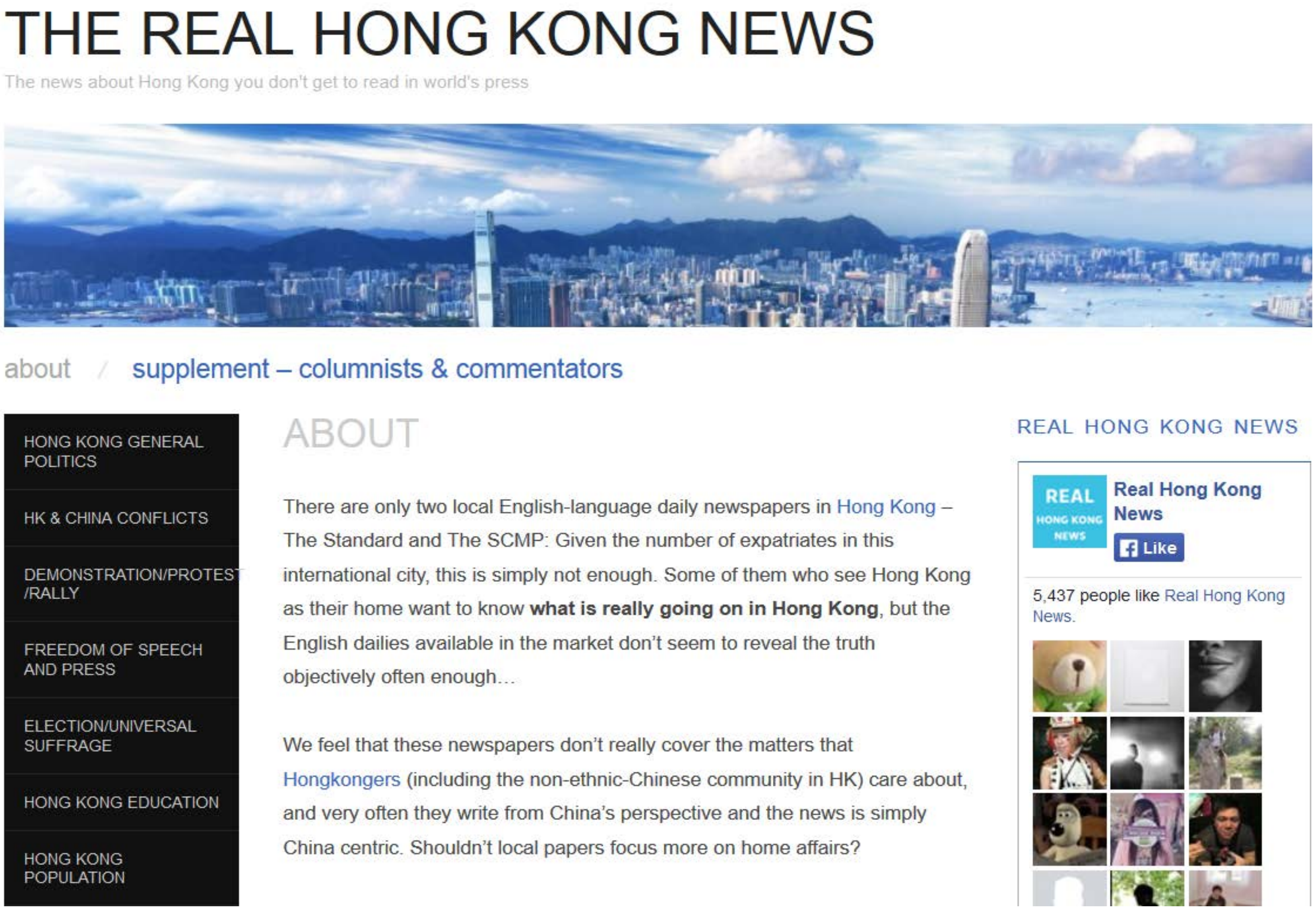


Monday, September 22

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

Exhibit G: Real Hong Kong News

Translated content site



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
REAL HONG KONG NEWS
REAL HONG KONG NEWS
Real Hong Kong News
Like

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?

5,437 people like Real Hong Kong News.

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

Thanks

Thank you very much!

I welcome emails at:

dtsang@uci.edu