Title
Web Archiving to Excavate Hidden Collections

Permalink
https://escholarship.org/uc/item/7t8771q0

Author
Tsang, Daniel C

Publication Date
2014-12-02
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 — UCS

PRESENCE
YOUR INSTITUTIONAL MEMORY
AS YOUR SITES
EVOLVE & 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
Internet Archive

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…

“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…”

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
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
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.

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).

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.

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.

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
- Assess 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?

http://www.scmp.com/topics/occupy-central
Exhibit B: Scholarism

https://twitter.com/scholarismhk/
Exhibit C: Occupy Central

Note: English page

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

http://oclpkenglish.wordpress.com/
Exhibit E: SocREC

Facebook site: https://www.facebook.com/socrec
A cartoon timeline: Harry's View on Occupy Central | South China Morning Post
Exhibit G: 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

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?

http://therealnewshk.wordpress.com/about/
Thank you very much!

I welcome emails at:

dtsang@uci.edu