CDL’s Web Archiving System

Erik Hetzner

UC3, California Digital Library

16 June 2011
- We don’t decide **what** to collect.
- We don’t decide **when** to collect it.
- We build tools to allow curators to make those decisions.
Vital statistics

- 35 public archives
- 16 partners
- 2724 web sites
- 289,272,095 URLs (×2)
- 16.1 TB (×2)
Vital statistics

- 35 public archives
- 16 partners
- 2724 web sites
- 289,272,095 URLs \((\times 2)\)
- 16.1 TB \((\times 2)\)
Vital statistics

- 35 public archives
- 16 partners
- 2724 web sites
- 289,272,095 URLs (×2)
- 16.1 TB (×2)
Vital statistics

- 35 public archives
- 16 partners
- 2724 web sites
- 289,272,095 URLs (×2)
- 16.1 TB (×2)
Vital statistics

- 35 public archives
- 16 partners
- 2724 web sites
- 289,272,095 URLs (×2)
- 16.1 TB (×2)
The Web has revolutionized our access to information, but Web publications are fragile, and ready access to Web resources cannot be taken for granted. The Web Archiving Service enables librarians and scholars to meet that challenge.
Introduction

Archive

2007 Southern California Wildfires Web Archive
California Digital Library

Description

In October of 2007, a series of wildfires broke out throughout southern California. This web archive documents that event as it appeared on California State Agency sites, federal government sites, news, blogs, social networking sites.

In addition to providing lasting access to the web’s coverage of the fires, this archive may serve as a point of comparison to other web archives of dramatic and quickly unfolding historical events. In particular it may show how people communicated about the fire, ho...

more...

Search

- fire

Quick Facts
- Sites: 152
- Oldest site: 10/23/07
- Most recent site: 05/07/08

Powered by the Web Archiving Service from the California Digital Library
2007 Southern California Wildfires Web Archive

California Digital Library

Search

Type words or a URL:

fire

Filter by:

All sites

All file types

Your search terms will be found anywhere in the full text of the web pages and documents in this archive. You can search for keywords or for particular URLs.

You can narrow your search to find the words in particular web sites, or to find only particular file types, such as PDF files.

You are not required to type a key word in your search; you can select a site to review all of the documents from that site.
Site list

2007 Southern California Wildfires Web Archive
California Digital Library

Refine site list

lookup by site name

Go Clear

Site list by topic:
- Blogs
- Cities
- Counties
- Federal Resources
- Images and Video
- Maps
- News
- Spanish Language
- State Resources

211 San Diego Wildfire Emergency  Show Info
760 KFMB  Show Info
ALT1040  Show Info
And Still I Persist  Show Info
Associated Content  Show Info
barboni.org  Show Info
Business: Flying Tankers  Show Info
CAL FIRE  Show Info
calfires.com  Show Info
Collection focus (unofficial)

- Middle East political sites (Stanford)
- Social movements (Tamiment, NYU)
- California government sites (UC)
Tools

- Heritrix 1.14.x
- Open-source Wayback
- Nutchwax (moving to Solr)
- CDL’s legacy Digital Preservation Repository
- ... and a lot of UI code
- ... ARC management, indexing scripts, etc.
Web archiving is easy*, but there are some difficulties.
Uneven coverage

We only crawl what our curators select.
Human selection

High precision; low recall.
We are not Internet Archive scale: but we are big enough that it takes a long time to do anything.
Collection mismatch

Our crawls are organized into ‘collections’. Everybody else has ‘one big archive’.
We are customer-driven: we need to convince customers that collaboration is good for them.
What’s on our plate

- Deduplication
  - ... requires a new index (Solr)
- Moving to our new Merritt repository
- Implementing Memento
What’s on our plate

- Deduplication
- ... requires a new index (Solr)
- Moving to our new Merritt repository
- Implementing Memento
What’s on our plate

- Deduplication
- ... requires a new index (Solr)
- Moving to our new Merritt repository
- Implementing Memento
Possibilities

What’s on our plate

- Deduplication
- ... requires a new index (Solr)
- Moving to our new Merritt repository
- Implementing Memento
Evaluating community needs

What do we have that *you* need?
What do you have that *we* need?
Collaboration with researchers

The hard, fun problems are not necessarily the ones that we need to be solved. But maybe we can work it out.
Temporal search

How can we rank (and display) results across time?
Standards for sharing, or providing computational access to, metadata or full content.
Flash and HTML5 throw a monkeywrench in the web.
Cross-archive collections

There is no reason why our curators should only be using ‘our’ crawls. How can we build collections that span archives?
CDL’s Web Archiving Service

- We build tools; curators build collections.
- We are ready to be part of a global web archive infrastructure.
- What next?

Thanks for having me, and thanks for listening.

erik.hetzner@ucop.edu