

# 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 (×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)

# Vital statistics

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

# The Web Archiving Service

The screenshot shows a web browser window displaying the Web Archiving Service website. The browser's address bar shows the URL <http://webarchives.cdlib.org/>. The website header includes the logo, the text "web archiving service", and the tagline "Capture today's web \* Build tomorrow's archives". A navigation bar contains links for Home, Information for, Partners, WAS News, and Contact WAS. A "build an archive" button with a "log in" link is also present.

The main content area features a large introductory paragraph: "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."

An "Information for" section lists the following roles:

- Potential Partners
- Researchers
- WAS Curators
- Webmasters

The "The Archives" section lists several archived collections:

- 2003 California Recall Election
- [2007 Southern California Wildfires Web Archive](#)
- 2009 H1N1 Influenza A (Swine Flu) Outbreak
- 2010 Winter Olympics
- AFL-CIO - Change to Win: the open web archive
- African Politics Web Archive
- Alternative Mass Media / News Web Sites Web Archive
- Anarchism Web Archive

A detailed view of the "2007 Southern California Wildfires Web Archive" is shown, including a thumbnail image of a wildfire. The details are as follows:

- Created by:** California Digital Library
- Sites:** 152
- Oldest site:** 10/23/07
- Most recent site:** 05/07/08
- 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 cov...

Below the description, there is a link to "Learn more" and a URL: <http://webarchives.cdlib.org/a/calfires>.



# Archive

File Edit View History Bookmarks Tools Help

http://webarchives.cdlib.org/a/calfires

Google

## 2007 Southern California Wildfires Web Archive

California Digital Library

Home About Site List Search Help Contact Us

### 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

Search

### 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](#)

## Search

File Edit View History Bookmarks Tools Help

http://webarchives.cdlib.org/a/calfires/search?terms=fire&search=Search

Google



## 2007 Southern California Wildfires Web Archive

California Digital Library

Home About Site List Search Help Contact Us

### Search

Type words or a URL:

Filter by:

« Previous 1 2 3 4 5 6 7 8 9 10 Next »

**fire wiki**  
 Captured: Mon Oct 29 22:55:38 UTC 2007  
[fire.pbwiki.com/](http://fire.pbwiki.com/) 16.9 kB text/html  
 ... relating to the **fire** service. This includes **Fire** Departments, Unions, Businesses, Buffs, etc ... for Tomorrow **Fire** ...

**WELCOME TO CAL FIRE**  
 Captured: Thu Oct 25 18:54:02 UTC 2007  
[fire.ca.gov/](http://fire.ca.gov/) 39.4 kB text/html

Your search terms will be found anywhere in the full text of the web pages and documents in this archive. You can search for key words 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.

S Z

## Site list

File Edit View History Bookmarks Tools Help

http://webarchives.cdlib.org/a/calfires/sites

2007 Southern California Wildfires Web Archive

## 2007 Southern California Wildfires Web Archive

California Digital Library

Home About Site List Search Help Contact Us

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](#)

## Archived page

File Edit View History Bookmarks Tools Help

← → ↻ ↺ http://webarchives.cdlib.org/sw1697010/http://www.760kfm.com/ ☆ ☰ Google 🔍 🏠 🍌




## 2007 Southern California Wildfires Web Archive

**Title:** 760 KFMB AM - San Diego, CA - Talk Radio  
**Archival URL:** http://webarchives.cdlib.org/sw1697010/http://www.760kfm.com/  
**Date captured:** 10/31/07 04:48 PM  
[About this archive](#)

Dates: 10/23/07 11:51 PM

[show document](#) [show details](#) [help](#)

This document is an archived copy for study and research. The original may be available at <http://www.760kfm.com/>



**SAN DIEGO, CALIFORNIA**  
JUNE 16, 2011


**LISTEN** **TRAFFIC** **COASTAL**  
Partly Cloudy  
HIGH: 69 LOW: 56


🔊 🚦 🌤️

**HOME** **LISTEN LIVE** **NEWS** **PERSONALITIES** **PROGRAM SCHEDULE** **FEATURES** **RICK'S REWARDS** **KFMB INFO**

**KEYWORD SEARCH**

**760 KFMB LINE UP**

 **RICK ROBERTS**  
5am-10am  
[email](#) | [www](#)

 **ON-AIR NOW** 📻  
**BILL O'REILLY**  
10am-12pm

**RELIEF FUND**  
KFMB Stations and the Salvation Army are now accepting donations for those displaced by the fire. [DONATE ONLINE >](#)

- DONATE TO THE WILDFIRE RELIEF FUND
- Rick's Rewards
- I Want My Country Back Message Board
- 760 AM On Demand
- Rick Roberts' Newsletter
- Advertisers' Directory

**Get 1 Free Magazine**  
SAN DIEGO'S neighborhood guide to **FREE Offers**

## 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.

# Scale

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'.

# Politics

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

# 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

Standards for sharing, or providing computational access to, metadata or full content.

# The changing web

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`