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.



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

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

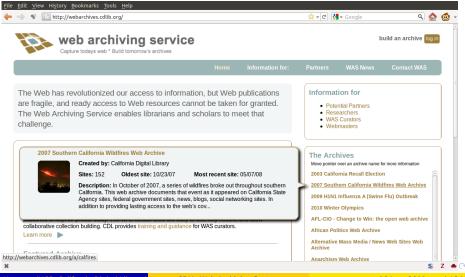
- 35 public archives
- 16 partners
- 2724 web sites
- 289,272,095 URLs (×2)
- 16.1 TB (×2)

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

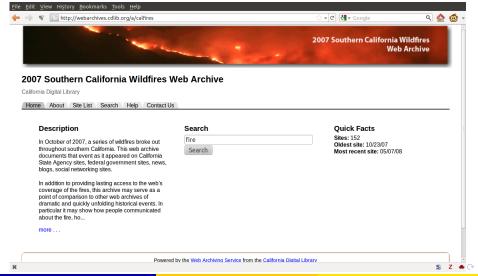
- 35 public archives
- 16 partners
- 2724 web sites
- 289,272,095 URLs (×2)
- 16.1 TB (×2)



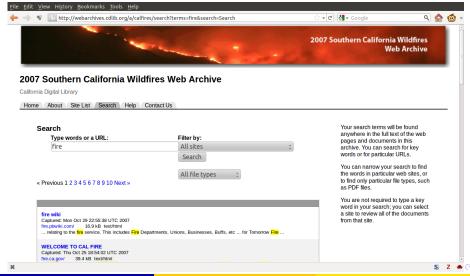
The Web Archiving Service



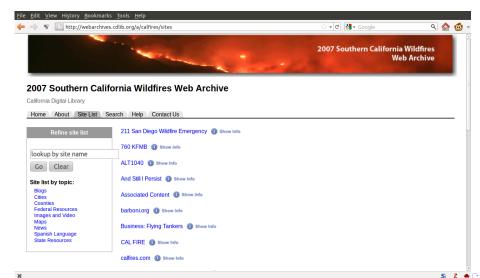
Archive



Search



Site list



Archived page



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.

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



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



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

- 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