Building Full Text Indexes of Web Content using Open Source Tools

Erik Hetzner
erik.hetzner@ucop.edu
UC Curation Center, California Digital Library

30 June 2012
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

- 49 public archives
- 19 partners
- 3684 web sites
- 489,898,652 URLs (×2)
- 25.5 TB (×2)
**Vital statistics**

- 49 public archives
- 19 partners
- 3684 web sites
- 489,898,652 URLs (×2)
- 25.5 TB (×2)
Vital statistics

- 49 public archives
- 19 partners
- 3684 web sites
- 489,898,652 URLs ($\times 2$)
- 25.5 TB ($\times 2$)
Vital statistics

- 49 public archives
- 19 partners
- 3684 web sites
- 489,898,652 URLs (×2)
- 25.5 TB (×2)
Vital statistics

- 49 public archives
- 19 partners
- 3684 web sites
- 489,898,652 URLs ($\times 2$)
- 25.5 TB ($\times 2$)
How we organize things

- Each curator creates projects
- Each project contains sites
- Each site contains jobs
Actually existing web archive search

Why do we always see this?

AP: 50 DIE IN MISSISSIPPI

Storm death toll rising
- Mississippi official tells AP 50 dead in one county
- Louisiana governor: "We believe we lost lives"
- Tens of thousands to be homeless for months
- 1.3 million without power
- Oil rig breaks loose, hits Mobile River bridge

This image is
URL Lookup

http://archive.org

http://archive.org has been crawled 9,061 times going all the way back to January 26, 1997. A crawl can be a duplicate of the last one. It happens about 25% of the time across 420,000,000 websites. FAQ
NutchWAX

- Web Archiving eXtensions for Nutch.
- Nutch is an open source web crawler, with search.
- Web Archiving eXtensions written by Internet Archive.
Actually existing web archive search

WAS

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.
Actually existing web archive search

Archive-IT

search the text within the archived pages. Or for more search options, use the Advanced Search options below.

Advanced Search

Contains any of:

etoile de dakar

Contains all of:

Exact phrase:

Not containing:

From the Host:
ex. www.archive-it.org

Results per host:

collecting organization, collection, site, specific URL or to search the text of archived webpages.

etoile de dakar

The following results were found for the term(s): etoile de dakar
• No metadata results for etoile de dakar, but there are up to 251 matches within the page text.

Etoile de Dakar

Collection: 2010 Lifestyles/Fads By Chief Umtuch Middle School
URL: http://allmusic.com/artist/etoile-de-dakar-p43383
This text was captured on Nov 10, 2010

Show All Captures

Etoile de Dakar Artist/Group Album Song Classical Work » New Releases » Editors’ Choice » Top Pages... Credits Charts & Awards Etoile de Dakar Years Active 1910 20 30 40 50 60 70 80 90 2000 Genres... Artist ID P 43383 Corrections to this Entry Biography by Craig Harris Etoile De Dakar was one of the... such western artists as Peter Gabriel, Paul Simon and David Byrne. The roots of the Etoile De Dakar were planted in 1960 when Ibra Kasse, owner of the Miami Club in
Actually existing web archive search

Portuguese Web Archive

etoile de dakar

between: 01/01/1996 and: 31/12/2011

Results 1 to 10 from 3,455

Did you mean: toilet de dakar

Man and the Biosphere Programme
20 May, 2009 - other dates
Direction de l'Environnement et des Etablissements Classés 23, Rue Calmette, Dakar – SENEGAL Tél ...

http://www.dakar.unesco.org/documents/mab_finalprt.htm

THEME 1: Zonage et amélioration du fonctionnement des réserv...
20 May, 2009 - other dates
– Centre de Coopération internationale en Recherche Agronomique pour le Développement BP 6189 Dakar Etoile ...

Erik Hetzner erik.hetzner@ucop.edu (CDL)
Library of Congress Web Archive

Your search **+etoile +de +dakar** - did not match any records.

Suggestions:
- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.

Go Back
Actually existing web archive search

Google

Etoile de Dakar - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Etoile_de_Dakar
Etoile de Dakar were a leading music group of Senegal in the 1970s. The group was formed in 1979 by Youssou N'Dour and members of the Star Band one of ...

Etoile de Dakar - Music Biography, Credits and Discography : AllMusic
www.allmusic.com/artist/etoile-de-dakar-mn0000208468
Find Etoile de Dakar bio, songs, credits, awards related and video information on AllMusic - Etoile De Dakar was one of the most influential bands to come out of ...

Etoile de Dakar - Thiely - YouTube
www.youtube.com/watch?v=KJ3lHlo2mm0
Jan 27, 2010 - 5 min - Uploaded by senegale
Etoile de Dakar - Thiely. senegale ... Xart.wmvby citoyendelmundo70298 views; Youssou et le Super ...
Scale

- IA collections > 2PB
- WAS collections > 50TB
Temporal search is not easy

[ michael jackson death ]
Resources

- Google’s 2011 revenue: $38 bn.
- UC’s 2011/12 revenue: $22 bn.
Deduplication

- Reduce redundant storage by storing pointers back to identical, previously captured content.
- ... but how to index this?
- Couldn’t figure how to make NutchWAX do this.
Curator-supplied metadata

- Our curators supply metadata (primarily tags) about the sites they capture
- This metadata should be indexed
- Curators should be able to modify this metadata at any time
Why a new indexing system?

NutchWAX

... and besides, Nutch is aging.
Nutch now focused on crawling, not search.
Our usage of NutchWAX was very slow.
Why a new indexing system?

Temporal web

- ... furthermore, web archive indexing is different.
- We capture the same URLs, again and again.
- It would be nice to build a web search system that takes time into account.
- We began writing a new indexing system
- We want to write as little as possible (see resources, above)
- So we stitched together FOSS tools
Scala

- Written in the Scala language
- To interact with Pig, Solr, etc.
We mostly need to parse HTML, but PDFs are very important to our users.
Not to mention Office
Apache software project
Wraps parsers for different file types in a uniform interface.
Parses most common file types.
Use the same code to parse different types.
Tika difficulties

- Some files are slow to parse.
- Some files blow up your memory.
- Some file parses never return.
Tika solutions

- Don’t parse files that are too big (e.g. > 2 MB)
- Fork and monitor process from the outside (Hadoop comes in handy)
- Preparse everything
What is Pig?

- Platform for data analysis from Apache.
- Based on Hadoop.
  - fault tolerant
  - distributed processing
- Can be used for ad-hoc analysis, without writing Java code.
- Embraced by the Internet Archive.
Why solr?

- Why not?
- Widely used.
- Takes the ‘kitchen sink’ approach to features.
- Hathitrust work seems to show that it can scale up to our needs.
Solr difficulties

- Cannot modify documents
- Solution: use stored fields, merge
- Need fast check for deduplicated content
- Solution: fetch document IDs, lookup in Bloom Filter
Thrift

- To communicate between our WAS-specific Ruby code and Scala
Hadoop File System (HDFS)

- To store parsed JSON files.
digest : MQXNCI7KA3YBSJUZVHGXY3X2KBS56444
url :
http://www.googlebooksettlement.com/help/bin/answer.py?answer=1
arcname :
CDL-20120530062015-00000-tanager.ucop.edu-00306642.arc.gz
date :
2012-05-30T06:37:03Z
digest : MQXNCI7KA3YBSJUZVHGXY3X2KBS56444
url : http://www.googlebooksettlement.com/help/bin/answer.py?answer=1
arcname : CDL-20120530062015-00001-tanger.ucop.edu-00306642.arc.gz
date : 2012-05-30T06:20:50Z
Merged

digest : MQXNCI7KA3YBSJUZVHGXY3X2KBS56444
url : http://www.googlebooksettlement.com/help/bin/answer.py?answer=134644
arcname : CDL-20120530062015-00000-tanager.ucop.edu-00306642.arc.gz, CDL-20120530062015-00001-tanager.ucop.edu-00306642.arc.gz
date :
2012-05-30T06:37:03Z
2012-05-30T06:20:50Z
So far

- about 200 m. unique documents
- 4 solr shards
- 2 TBs of index
### Search

**department of motor vehicles**

**Title:** Department of Motor Vehicles (Business Partner Automation Program)

**Captured:** 06/25/11 07:15 PM


**Abstract:** Department of Motor Vehicles (Business Partner Automation Program) ... Department of Motor Vehicles (Business Partner Automation Program)

---

**Title:** Consumers - Automobile Dealers - California Dept. of Justice - Office

**Captured:** 06/22/11 04:28 PM

**URL:** [aq.ca.gov/consumers/general/automobile_dealers.php](http://aq.ca.gov/consumers/general/automobile_dealers.php)

**Abstract:** 1888 The Department of Motor Vehicles licenses and regulates new and used motor vehicle ... contact the Department of Motor Vehicles, Division of Investigations ... motor vehicle dealers. If you have a contractual dispute (purchase

---

**Title:** What's New at that Motor Vehicle Board

**Captured:** 06/25/11 06:40 PM

**URL:** [www.rmvb.ca.gov/rss/whats-new_feed.xml](http://www.rmvb.ca.gov/rss/whats-new_feed.xml)

**Abstract:** What's New at that Motor Vehicle Board ... What's New at that Motor Vehicle Board What's New at that

---

### Refine Your Results

**web site:**
- California State Contr... [5,310]
- A new test [3,250]
- Department of Industri... [2,458]
- Bureau of State Audits [2,925]

**site topic:**
- Economy [21,994]
- Environment [18,541]
- Health [14,132]
- Social Services [11,506]

**media type:**
- Pdf [53,728]
- Html [48,169]
- Office [4,158]
- Compressed [68]
Better ranking

- We have not explored ranking very much
- We store a Rabin fingerprint for every URL and its outlinks
- Have done some basic work with Webgraph tools to calculate ranks
- http://webgraph.di.unimi.it/
Speed improvements

- Currently we index about 3k jobs per day
- A lot of the slowness is related to merging content
- Some of the slowness is probably Solr tuning
weari: A WEB ARchive Indexer

- Tika + HDFS + Pig + Solr = weari

http://bitbucket.org/cdl/weari

Thanks!

erik.hetzner@ucop.edu