

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 ($\times 2$)
- 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 ($\times 2$)

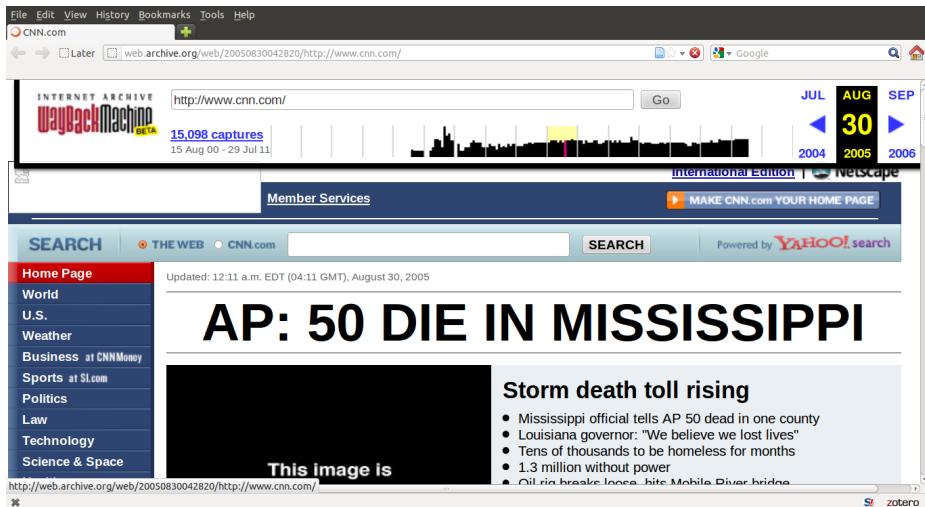
Vital statistics

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

How we organize thing

- Each curator creates projects
- Each project contains sites
- Each site contains jobs

Why do we always see this?



6 / 38

NutchWAX

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



Archive-IT

File Edit View History Bookmarks Tools Help

http://www.arc...w=Collections

www.archive-it.org/explore?q=etoile+de+dakar&page=1&show=Collections

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

Search

clear

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.

Search Page Text

Page 1 of 13 (251 Total Results)

Next Page ▶

Sort By: **Best Match**

Etoile de Dakar

Collection: 2010 Lifestyles/Fads By Chief Umtuch Middle School

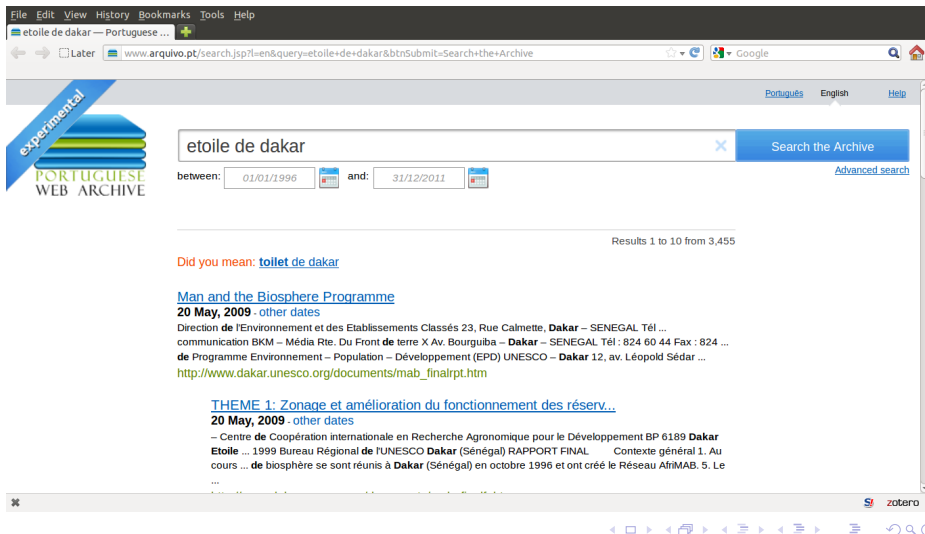
URL: <http://allmusic.com/artist/etoile-de-dakar-p43383>

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

zotero

Portuguese Web Archive



File Edit View History Bookmarks Tools Help

etoile de dakar — Portuguese ...

Later www.arquivo.pt/search.jsp?l=en&query=etoile+de+dakar&btnSubmit=Search+the+Archive

Português English Help

experimental

PORTUGUESE WEB ARCHIVE

etoile de dakar

Search the Archive

Advanced search

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 ...
communication BKM – Média Rte. Du Front de terre X Av. Bourguiba – **Dakar** – SENEGAL Tél : 824 60 44 Fax : 824 ...
de Programme Environnement – Population – Développement (EPD) UNESCO – **Dakar** 12, av. Léopold Sédar ...
http://www.dakar.unesco.org/documents/mab_finalrpt.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 ... 1999 Bureau Régional de l'UNESCO **Dakar** (Sénégal) RAPPORT FINAL Contexte général 1. Au
cours ... de biosphère se sont réunis à **Dakar** (Sénégal) en octobre 1996 et ont créé le Réseau AfrimAB. 5. Le

zotero

Library of Congress

File Edit View History Bookmarks Tools Help

Search Results (+etoile +de +d...) +

Later lcweb2.loc.gov/diglib/lcwa/searchAll?query=etoile+de+dakar&field=all&sort=titlesort Google

The Library of Congress >> More Online Collections

Library of Congress Web Archives *Minerva*

BROWSE | SEARCH | TECHNICAL INFORMATION

[LC Web Archives](#) >> [Search all](#) >> Search results

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](#)

zotero

Google

The screenshot shows a Google search results page for the query "etoile de dakar". The browser's address bar shows the URL: www.google.com/#hl=en&gs_nf=1&gs_mss=etoile&cp=10&gs_id=5p&xhr=t&q=etoile+de+dakar&pf=p&output=search&. The search bar contains the text "etoile de dakar". The search results are displayed under the heading "Search" with the text "About 1,400,000 results (0.26 seconds)".

The results are categorized by type:

- Web**:
 - [Étoile de Dakar - Wikipedia, the free encyclopedia](http://en.wikipedia.org/wiki/%C3%89toile_de_Dakar)
en.wikipedia.org/wiki/Étoile_de_Dakar
 - Étoile 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 ...
- Images**
- Maps**
- Videos**:
 - [Etoile de Dakar - Music Biography, Credits and Discography : AllMusic](http://www.allmusic.com/artist/etoile-de-dakar-mn000208468)
www.allmusic.com/artist/etoile-de-dakar-mn000208468
 - 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 ...
- News**
- Shopping**
- Blogs**
- More**

The video result for "Etoile de Dakar - Thiely - YouTube" is expanded, showing a thumbnail of the Senegalese flag and the video title "Etoile de Dakar - Thiely. senegagale Xarit.wmvby citoyendelmundo70298 views; Youssou et le Super ...". The video duration is 5:13.

The browser's status bar at the bottom shows "Oakland, CA" and "zotero".

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

NutchWAX

- ... and besides, Nutch is aging.
- Nutch now focused on crawling, not search.
- Our usage of NutchWAX was very slow.

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.

weari: a WEb ARchive Indexer

- 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.

Tika

- 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

```
{ "filename"          :  
  "CDL-20070613172954-00002-ingest1.arc.gz",  
  "digest"           : "DWHNMIQN3OZLG3ZW2PZQCTEU0AWCL5RJ",  
  "url"               : "http://medlineplus.gov/",  
  "date"              : 1181755806000,  
  "title"             : "MedlinePlus Health Information ...",  
  "length"            : 24655,  
  "content"           : "MedlinePlus Health Information ...",  
  "suppliedContentType" : { "top" : "text", "sub" : "html" },  
  "detectedContentType" : { "top" : "text", "sub" : "html" },  
  "outlinks"          : [ 623129493561446160, ... ] }
```

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.

Original

digest : MQXNCI7KA3YBSJUZVHGXY3X2KBS56444

url :

<http://www.googlebooksettlement.com/help/bin/answer.py?answer=>

arcname :

CDL-20120530062015-00000-tanager.ucop.edu-00306642.arc.gz

date :

2012-05-30T06:37:03Z

New

digest : MQXNCI7KA3YBSJUZVHGXY3X2KBS56444

url :

<http://www.googlebooksettlement.com/help/bin/answer.py?answer=>

arcname :

CDL-20120530062015-00001-tanager.ucop.edu-00306642.arc.gz

date :

2012-05-30T06:20:50Z

Merged

digest : MQXNCI7KA3YBSJUZVHGXY3X2KBS56444

url :

<http://www.googlebooksettlement.com/help/bin/answer.py?answer=1>

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

in Full text

Search

display: 10 | 25 | 50 | 100

brief records | titles only | URLs only

« Previous 1 2 3 4 5 6 7 8 9 ... 10926 10927 Next »

- Title:** Department of Motor Vehicles (Business Partner Automation Program)
788.7 kB
Captured: 06/25/11 07:15 PM
URL: www.oal.ca.gov/res/docs/pdf/disapp...I_decisions/2008/2008-0821-03S.pdf
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 - Offi...
5.5 kB
Captured: 06/22/11 04:28 PM
URL: ag.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
1.9 kB
Captured: 06/25/11 06:40 PM
URL: www.nmmb.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

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

Use quotes to search an exact phrase. Example: "attorney general". See search help for details.

Refine Your Results

web site:

California State Contr... [5,310]
A new test [3,250]
Department of Industri... [2,458]
Bureau of State Audits [2,325]
[see all >](#)

site topic:

Economy [21,994]
Environment [18,541]
Health [14,132]
Social Services [11,506]
[see all >](#)

media type:

Pdf [53,728]
Html [46,189]
Office [4,158]
Compressed [58]

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`