

# *Web Archive Analysis*

Vinay Goel  
Senior Data Engineer  
Internet Archive  
[vinay@archive.org](mailto:vinay@archive.org)



# Access Archive-It Data

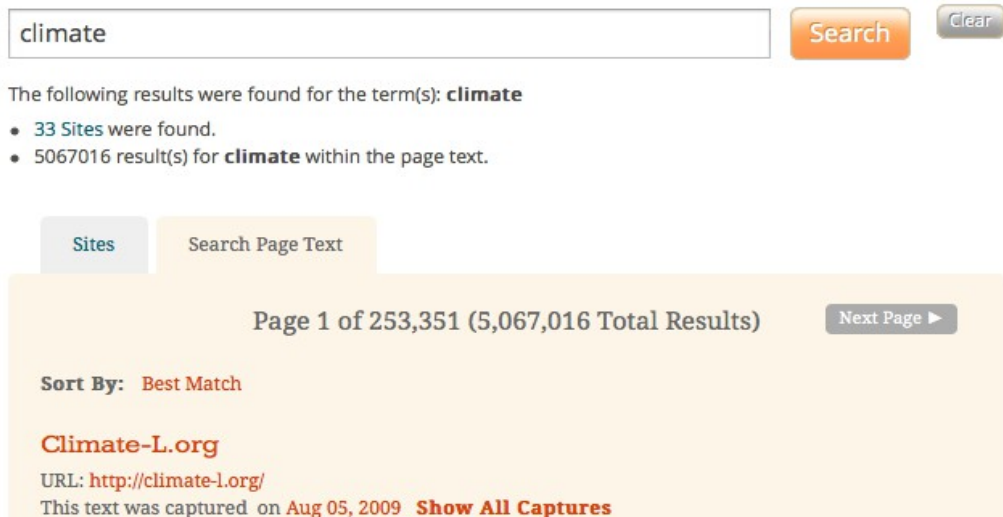
## Wayback Machine

You are viewing an archived web page, collected at the request of [Stanford University, Social Sciences Resource Group](#) using [Archive-It](#). This page was captured on 2:16:16 May 11, 2012, and is part of the [Climate change and environmental policy](#) collection. The information on this web page may be out of date. See [All versions](#) of this archived page.



The screenshot shows the top portion of a website. On the left is the World Bank logo with the text "THE WORLD BANK" and "Working for a World Free of Poverty". To the right is a search bar with a magnifying glass icon, the word "Search", and a "GO" button. Below the logo is a navigation bar with a "Home" button. A prominent red banner across the page contains the text "Climate Change" and a small right-pointing arrow.

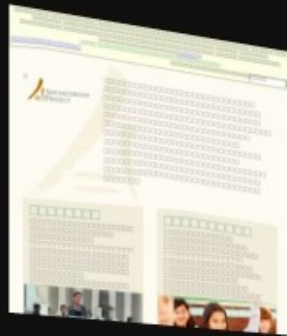
## Archive-It Search



The screenshot shows the search interface of the Archive-It project. A search box contains the word "climate". To its right are "Search" and "Clear" buttons. Below the search box, a message states: "The following results were found for the term(s): **climate**". This is followed by a bulleted list: "33 Sites were found." and "5067016 result(s) for **climate** within the page text." Below the list are two tabs: "Sites" (which is selected) and "Search Page Text". At the bottom, it displays "Page 1 of 253,351 (5,067,016 Total Results)" and a "Next Page" button. The "Sort By" is set to "Best Match". A result for "Climate-L.org" is shown with the URL "http://climate-l.org/" and the capture date "Aug 05, 2009", with a "Show All Captures" link.

# ***Download Archive-It Data***

- Download crawl data
  - Use command-line tools / browser
- Download collection, seed and document level metadata
  - Use Archive-It Web Application / OAI-PMH XML feed
- About 25% of all partners download their collection data
- The rest are encouraged to do the same
- Reasons to download crawl data:
  - Satisfy specific preservation requirements
  - Provide access to restricted collections
  - Custom reporting, wayback and search
  - Custom research & analysis



## *Enable Research & Analysis*



# *Data: WARC*

- Data written by web crawlers
- Web ARchive Container File Format (ISO standard)
- Each file contains a series of concatenated records
  - Full HTTP request/response records
  - Metadata records (links, crawler path, encoding etc.)
  - Records to store duplicate detection events
  - Records to support segmentation and conversion
- Usually 1 Gigabyte in size

# *Derived Data: CDX*

- Index for Wayback Machine
- Space delimited text file
- Contains only essential fields needed by Wayback
  - URL, Timestamp, Content Digest
  - MIME type, HTTP Status Code
  - WARC filename and file offset of record
  - Optional: redirect URL, meta tags, size

# Wayback Machine



Clinical and Translational Science Awards Web Archive (National Institutes of Health)



Enter Web Address:  All

Searched for <http://casemed.case.edu/ctsc/>  
[Look up URL](#) in general Internet Archive web collection

11 Results [RSS](#) [Metadata](#)  
[Proxy Mode Help](#)

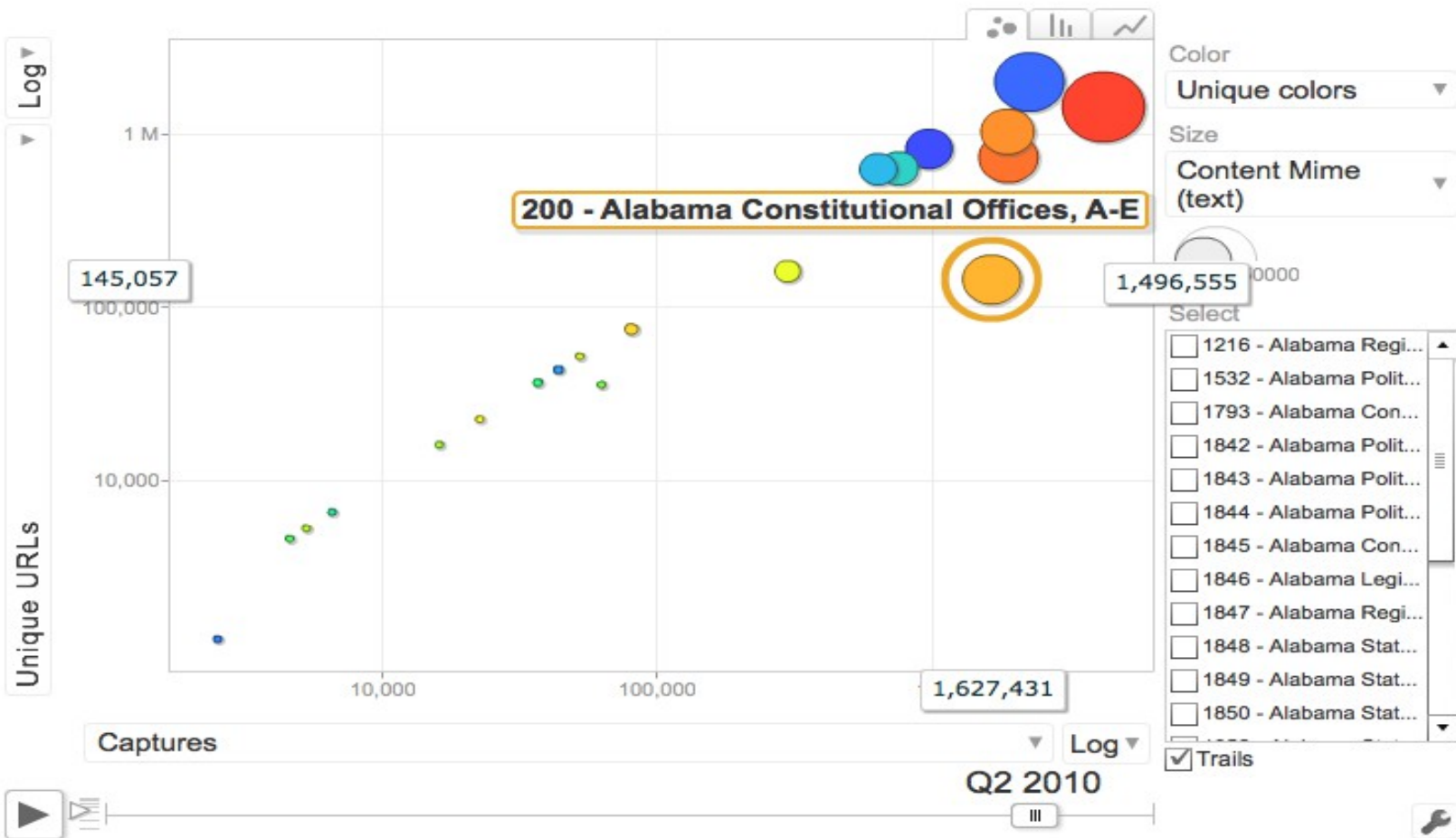
\* denotes when page was updated

## Found 11 Captures between Oct 14, 2010 - Oct 22, 2012

2010	2011	2012
3 pages	4 pages	4 pages
<a href="#">Oct 14, 2010</a> *	<a href="#">Jan 22, 2011</a> *	<a href="#">Jan 22, 2012</a> *
<a href="#">Oct 14, 2010</a> *	<a href="#">Apr 22, 2011</a> *	<a href="#">Apr 22, 2012</a> *
<a href="#">Oct 22, 2010</a> *	<a href="#">Jul 22, 2011</a> *	<a href="#">Jul 22, 2012</a> *
	<a href="#">Oct 22, 2011</a> *	<a href="#">Oct 22, 2012</a> *

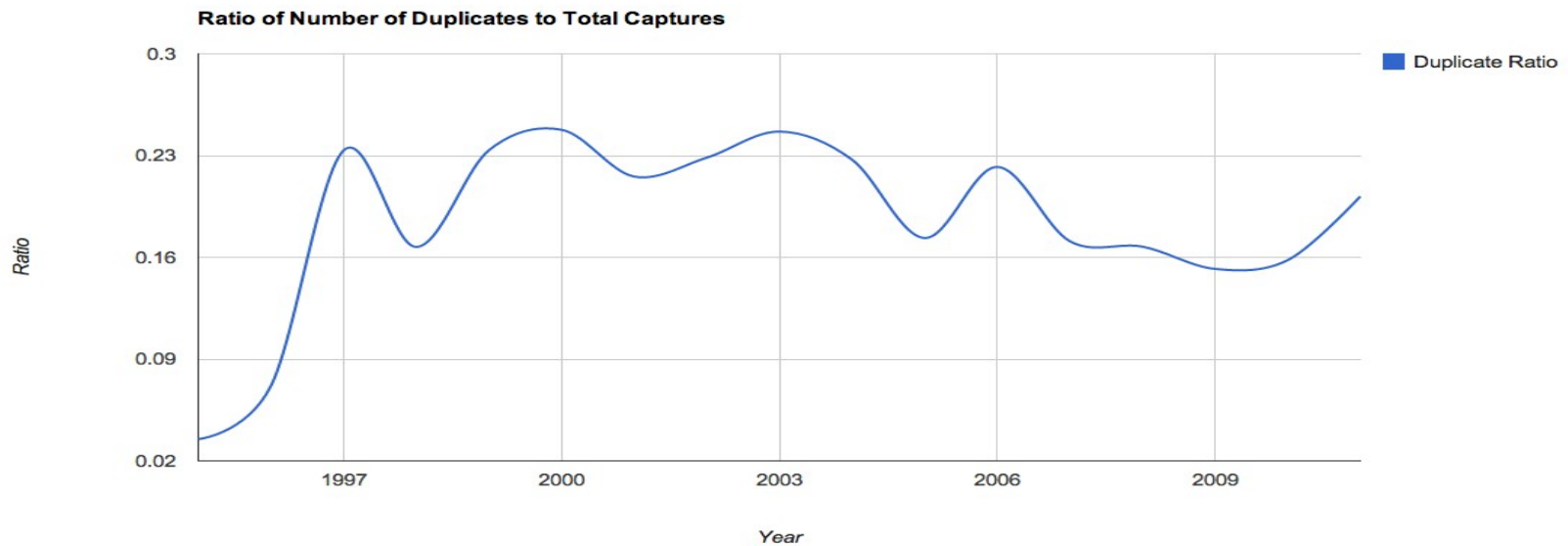
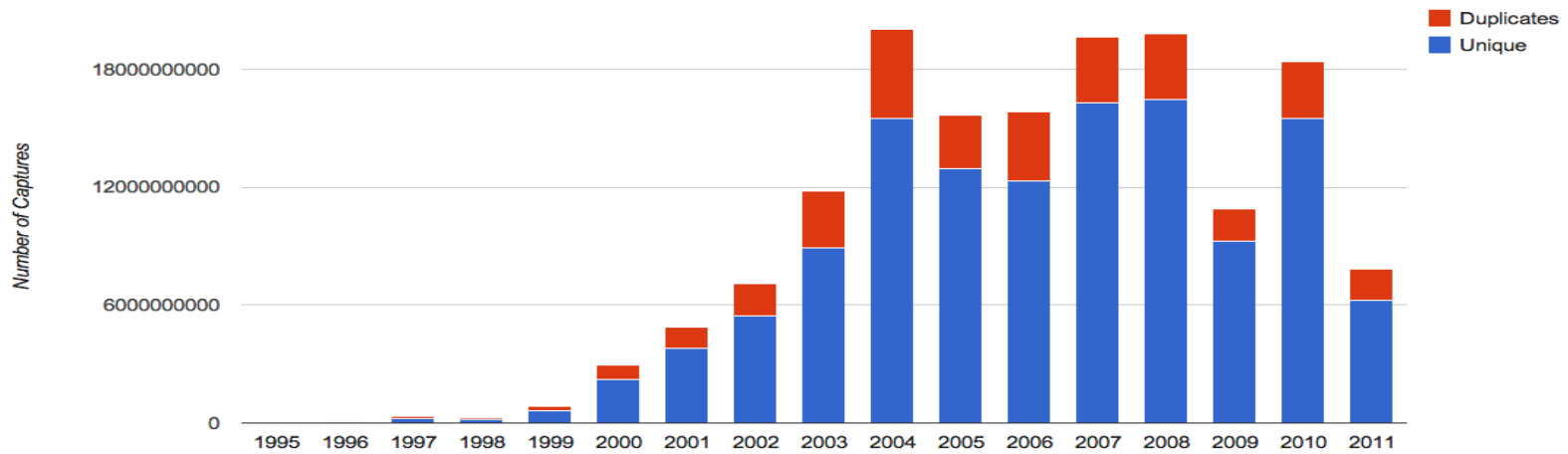
# Track growth of content

## Alabama State Archives





# Find rate of duplication



# ***Derived Data: Parsed Text***

- Input to build text indexes for Search
- Text is extracted from WARC files
- HTML boilerplate is stripped out
- Also contains metadata for each record
  - URL, Timestamp, Content Digest
  - MIME type, HTTP status code
  - Record Length
- Stored in Hadoop Sequence Files

# *Derived Data: WAT*

- Metadata format
- Essential metadata for many types of analyses
- Avoids barriers to data exchange: copyright, privacy
- Less data than WARC, more than CDX
- Extensible
- WAT records are WARC metadata records

# ***Derived Data: WAT (current version)***

- Contains all metadata needed for CDX
- 1:1 mapping to records in the corresponding WARC
- Contains for every HTML page in the WARC,
  - Title text
  - Description text
  - Meta keywords
  - Embeds and outgoing links with alt / anchor text

# *Analysis Toolkit*

- Enable arbitrary analysis of web archives
- Easy and quick to learn
- Scales up and down
- Tools
  - Hadoop (HDFS, MapReduce)
  - Apache Giraph
  - Apache Hive
  - Apache Mahout
  - Apache Pig

# Analysis Workflow

APACHE  
**HBASE**



Target Resource "Analysis"  
Live Snapshot Generation

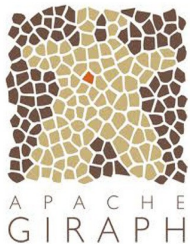
WAT Generation  
"Browser" Log Analysis  
"Crawl" Log Analysis  
Filtering APIs/Feeds



Seeding Crawler Frontier/s &  
Alternate Capture  
Mechanisms



Web Graph Generation  
In-link Analyses/Ranking  
Anchor, Description, Full Text  
Indexing & Mining



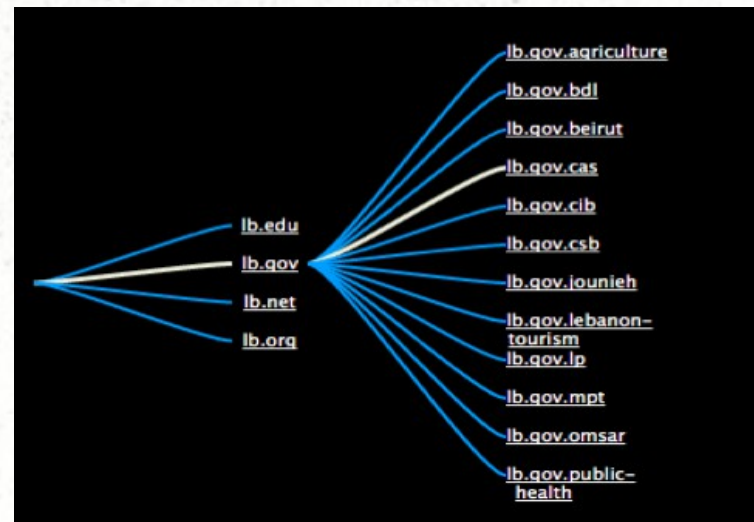
Embed & Out-link Analyses



*Lucene*

# Crawl Log Analysis (Hive/Giraph)

- Crawl Log Warehouse
- Distribution of HTTP status codes, MIME types
- Find timeout errors, duplicate content, crawler traps, robots exclusions
- Trace path of the crawler



# Combine with text metadata to fill in missing fields

## Youtube Videos

<< < > >> Rows per page: 20 Jump to page: 1

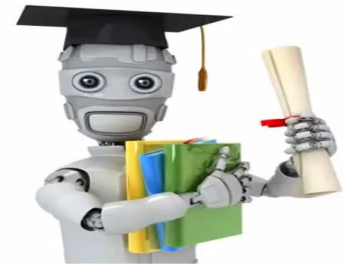
Page 1 of 6 -- displaying rows 1-20 of 109

Title ▲/▼	Size (bytes) ▲/▼	Date ▲/▼	Discovered from Seed ▲/▼	Referring Page (Live Web) ▲/▼	URL of video
Funding for Williston Park Road Improvements - YouTube	24627453	20121221	http://ackerman.house.gov/	<a href="http://www.youtube.com">www.youtube.com</a>	<a href="#">Play (from Archive)</a> <a href="#">Download (from Archive)</a>
Supporting State Department Authorization Bill - YouTube	5494322	20130101	http://ackerman.house.gov/	<a href="http://www.youtube.com">www.youtube.com</a>	<a href="#">Play (from Archive)</a> <a href="#">Download (from Archive)</a>
Credit Ratings and the First Amendment - YouTube	16804391	20130101	http://ackerman.house.gov/	<a href="http://www.youtube.com">www.youtube.com</a>	<a href="#">Play (from Archive)</a> <a href="#">Download (from Archive)</a>
Broader Focus on Afghanistan - YouTube	3807340	20121221	http://ackerman.house.gov/	<a href="http://www.youtube.com">www.youtube.com</a>	<a href="#">Play (from Archive)</a> <a href="#">Download (from Archive)</a>



# *Text Analysis (Pig/Mahout)*

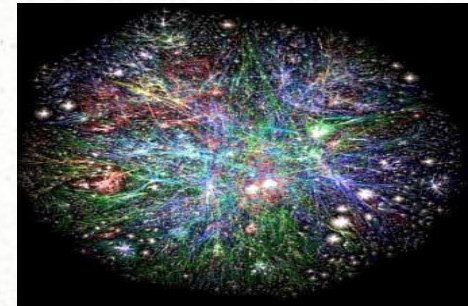
- Text extracted from WARC / Parsed Text / WAT files
- Use curated collections to train Classifiers
- Cluster documents in collections
- Topic Modeling
  - Discover topics
  - Study how topics evolve over time
- Compare how a page describes itself (meta text) vs. how other pages linking to it describe it (anchor text)



Machine Learning

# Link Analysis (Pig/Giraph)

- Links extracted from crawl logs / WARC metadata records / Parsed Text / WAT files
- Indegree and Outdegree information
- Inter-host and Intra-host link information
- Study how linking behavior changes over time
- Rank resources by PageRank
  - Identify important resources
  - Prioritize crawling of missing resources
- Find possible spam pages by running biased PageRank algorithms



# *Questions?*

- E-mail: [vinay@archive.org](mailto:vinay@archive.org)
- Code Repository
  - <https://github.com/internetarchive/>
  - <https://github.com/vinaygoel/archive-analysis/>