Analysis of Web Archives

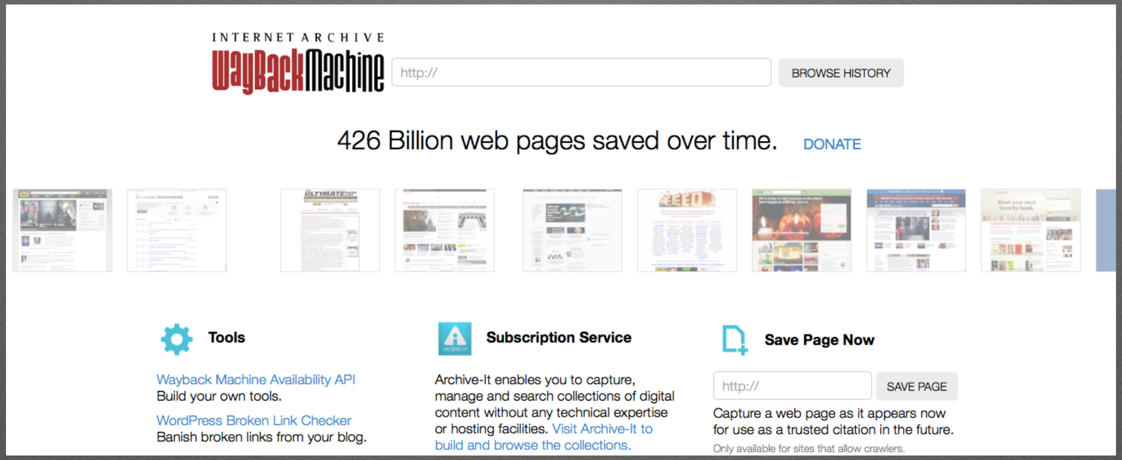
Vinay Goel
Senior Data Engineer

INTERNET ARCHIVE

Internet Archive

- Established in 1996
- 501(c)(3) non profit organization
- 20+ PB (compressed) of publicly accessible archival material
- Technology partner to libraries, museums, universities, research and memory institutions
- Currently archiving books, text, film, video, audio, images, software, educational content and the Internet

IA Web Archive



426 Billion web pages saved over time.

Tools

Wayback Machine Availability API
Build your own tools.

WordPress Broken Link Checker
Banish broken links from your blog.

Subscription Service

Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. Visit Archive-It to build and browse the collections.

Save Page Now

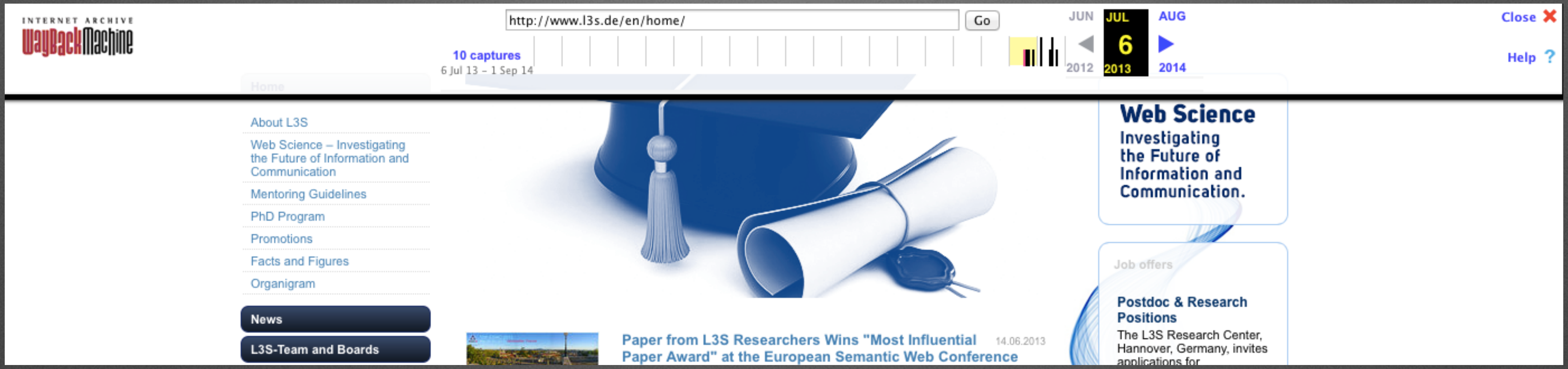
http://

Capture a web page as it appears now for use as a trusted citation in the future.

Only available for sites that allow crawlers.

- Began in 1996
- 426+ Billion publicly accessible web instances
- Operate web wide, survey, end of life, selective and resource specific web harvests
- Develop freely available, open source, web archiving and access tools

Access Web Archive Data



- About L3S
- Web Science – Investigating the Future of Information and Communication
- Mentoring Guidelines
- PhD Program
- Promotions
- Facts and Figures
- Organigram

News

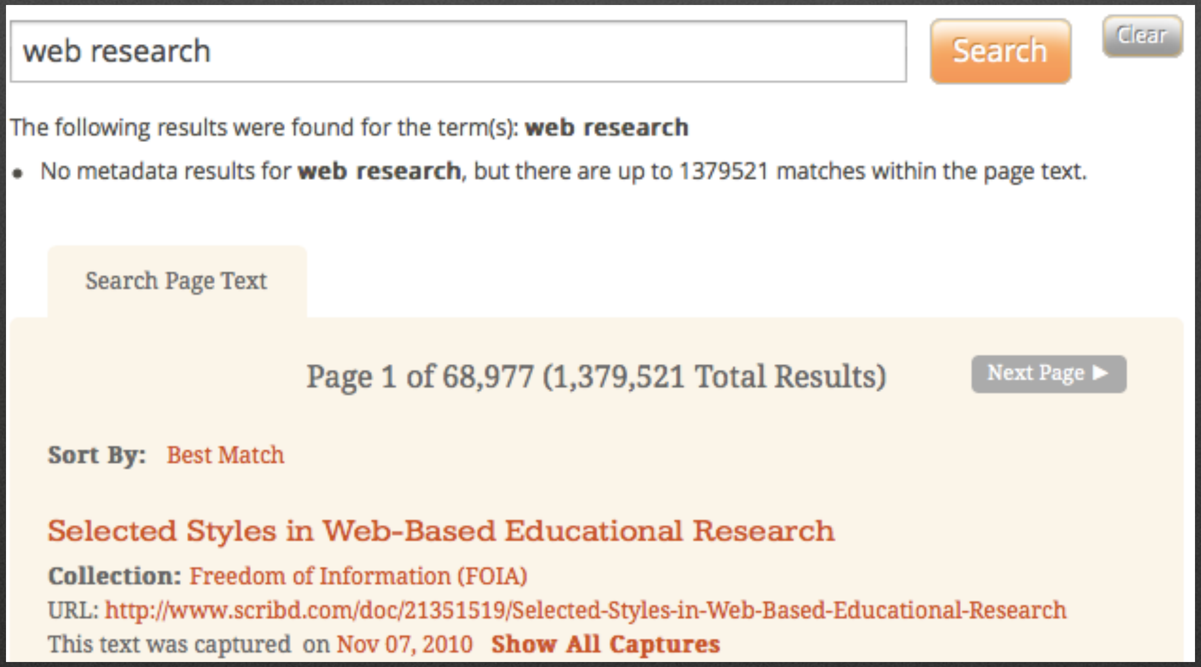
L3S-Team and Boards

Web Science
Investigating the Future of Information and Communication.

Paper from L3S Researchers Wins "Most Influential Paper Award" at the European Semantic Web Conference 14.06.2013

Job offers

Postdoc & Research Positions
The L3S Research Center, Hannover, Germany, invites applications for



The following results were found for the term(s): **web research**

- No metadata results for **web research**, but there are up to 1379521 matches within the page text.

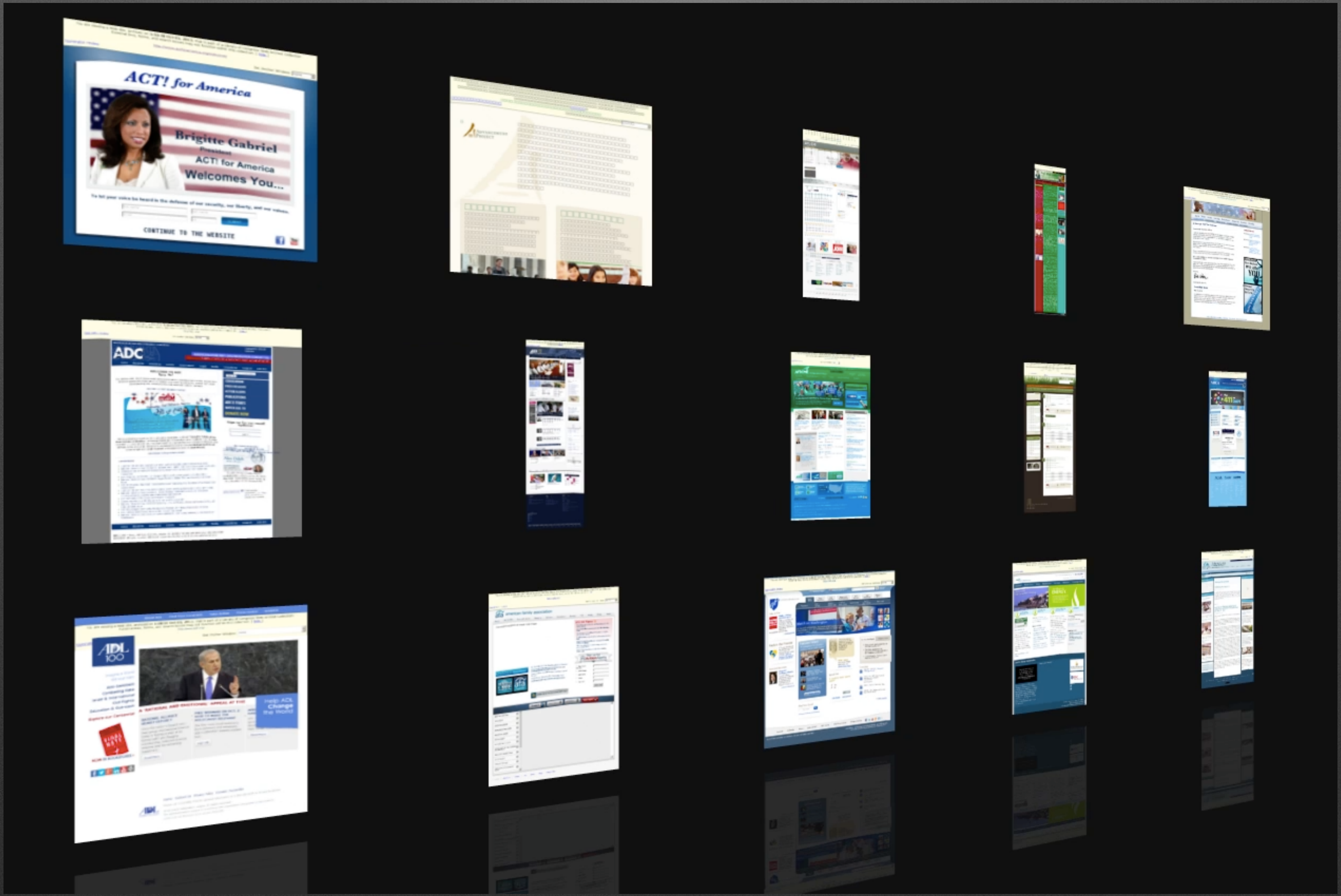
Page 1 of 68,977 (1,379,521 Total Results)

Sort By: Best Match

Selected Styles in Web-Based Educational Research

Collection: Freedom of Information (FOIA)
URL: http://www.scribd.com/doc/21351519/Selected-Styles-in-Web-Based-Educational-Research
This text was captured on Nov 07, 2010 **Show All Captures**

Analyze Web Archive Data



Analysis Tools

- Arbitrary analysis of archived data
- Scales up and down
- Tools
    - Apache Hadoop (distributed storage and processing)
    - Apache Pig (batch processing with a data flow language)
    - Apache Hive (batch processing with a SQL like language)
    - Apache Giraph (batch graph processing)
    - Apache Mahout (scalable machine learning)

Data

- Crawler logs

HERITRIX

- Crawled data
- Crawled data derivatives

WaybackMachine

  - Wayback Index
  - Text Search Index

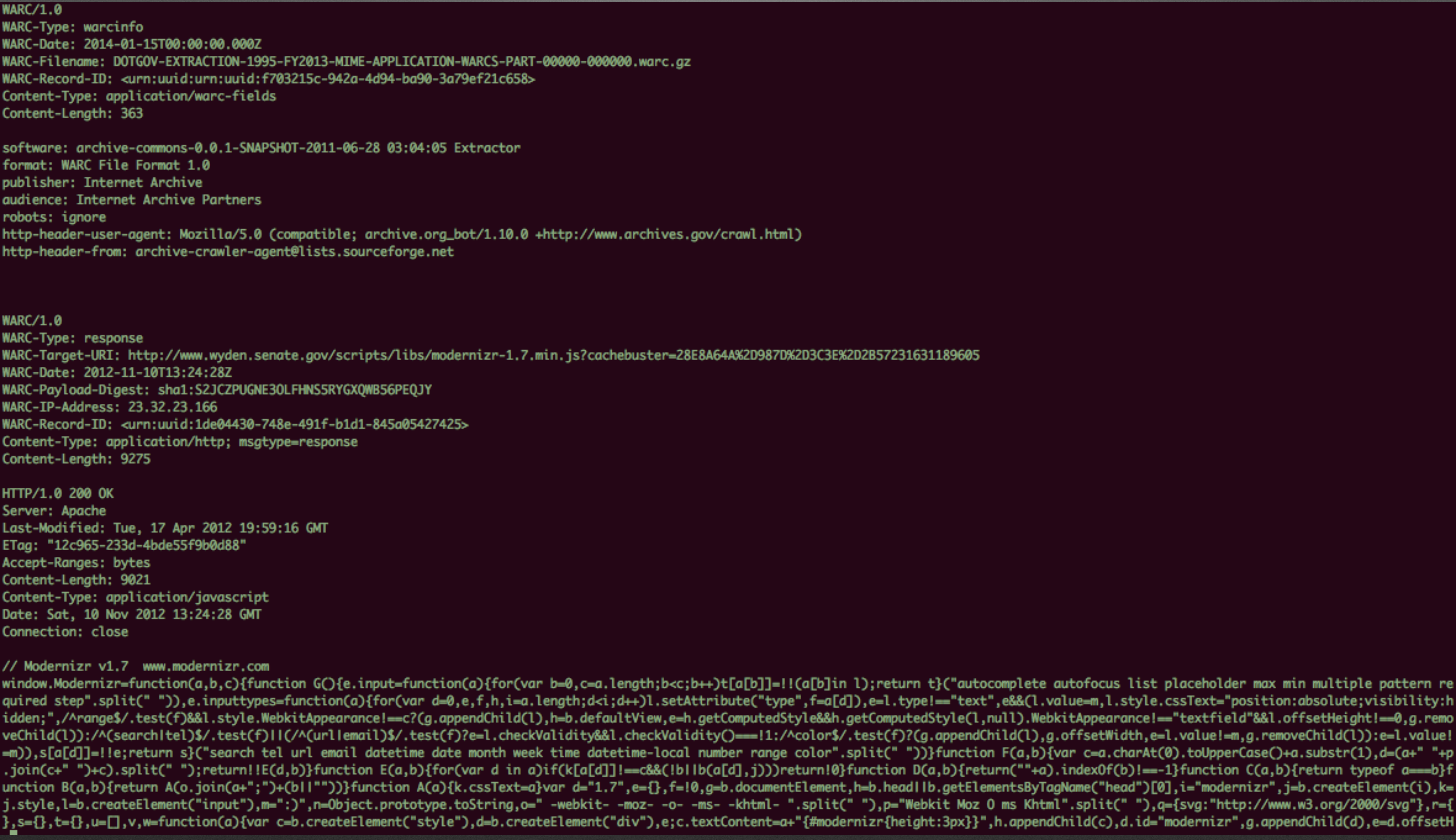
Lucene

  - WAT

Crawled data (ARC / WARC)

- Data written by web crawlers
- Before 2008, written into ARC files
- From 2008, IA began writing into WARC files
    - data donations from 3rd parties still include ARC files
- WARC file format (ISO standard) is a revision of the ARC file format
- Each (W)ARC file contains a series of concatenated records
    - Full HTTP request/response records
    - WARC files also contain metadata records, and records to store duplication events, and to support segmentation and conversion

WARC

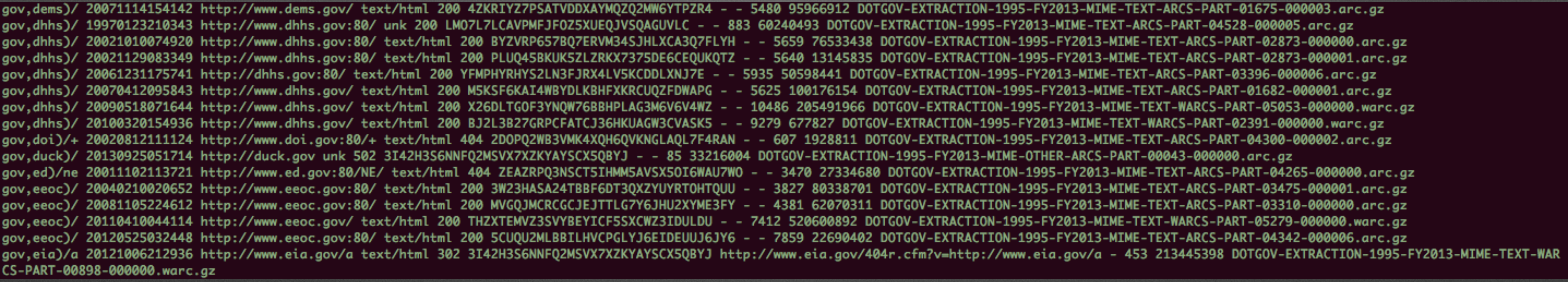


```
WARC/1.0
WARC-Type: warcinfo
WARC-Date: 2014-01-15T00:00:00.000Z
WARC-Filename: DOTGOV-EXTRACTION-1995-FY2013-MIME-APPLICATION-WARCS-PART-00000-000000.warc.gz
WARC-Record-ID: <urn:uuid:urn:uuid:f703215c-942a-4d94-ba90-3a79ef21c658>
Content-Type: application/warc-fields
Content-Length: 363

software: archive-commons-0.0.1-SNAPSHOT-2011-06-28 03:04:05 Extractor
format: WARC File Format 1.0
publisher: Internet Archive
audience: Internet Archive Partners
robots: ignore
http-header-user-agent: Mozilla/5.0 (compatible; archive.org_bot/1.10.0 +http://www.archives.gov/crawl.html)
http-header-from: archive-crawler-agent@lists.sourceforge.net



WARC/1.0
WARC-Type: response
WARC-Target-URI: http://www.wyden.senate.gov/scripts/libs/modernizr-1.7.min.js?cachebuster=28E8A64A%2D987D%2D3C3E%2D2B57231631189605
WARC-Date: 2012-11-10T13:24:28Z
WARC-Payload-Digest: sha1:S2JCZPUGNE3OLFHNS5RYGXQWB56PEQJY
WARC-IP-Address: 23.32.23.166
WARC-Record-ID: <urn:uuid:1de04430-748e-491f-b1d1-845a05427425>
Content-Type: application/http; msgtype=response
Content-Length: 9275

HTTP/1.0 200 OK
Server: Apache
Last-Modified: Tue, 17 Apr 2012 19:59:16 GMT
ETag: "12c965-233d-4bde55f9b0d88"
Accept-Ranges: bytes
Content-Length: 9021
Content-Type: application/javascript
Date: Sat, 10 Nov 2012 13:24:28 GMT
Connection: close

// Modernizr v1.7  www.modernizr.com
window.Modernizr=function(a,b,c){function G(){e.input=function(a){for(var b=0,c=a.length;b<c;b++)t[a[b]]=!!(a[b]in l);return t}("autocomplete autofocus list placeholder max min multiple pattern re
quired step".split(" ")),e.inputtypes=function(a){for(var d=0,e,f,h,i=a.length;d<i;d++)l.setAttribute("type",f=a[d]),e=l.type!=="text",e&&(l.value=m,l.style.cssText="position:absolute;visibility:h
idden;",/^range$/.test(f)&&l.style.WebkitAppearance!==c?(g.appendChild(l),h=b.defaultView,e=h.getComputedStyle&&h.getComputedStyle(l,null).WebkitAppearance!=="textfield"&&l.offsetHeight!==0,g.remo
veChild(l)):/^(search|tel)$/.test(f)||(/^(url|email)$/.test(f)?e=l.checkValidity&&l.checkValidity()===!1:/^color$/.test(f)?(g.appendChild(l),g.offsetWidth,e=l.value!=m,g.removeChild(l)):e=l.value!
=m)),s[a[d]]=!!e;return s}("search tel url email datetime date month week time datetime-local number range color".split(" "))}function F(a,b){var c=a.charAt(0).toUpperCase()+a.substr(1),d=(a+" "+p
.join(c+" ")+c).split(" ");return!!E(d,b)}function E(a,b){for(var d in a)if(k[a[d]]!==c&&(!b||b(a[d],j)))return!0}function D(a,b){return(""+a).indexOf(b)!==-1}function C(a,b){return typeof a===b}f
unction B(a,b){return A(o.join(a+";")+(b||""))}function A(a){k.cssText=a}var d="1.7",e={},f=!0,g=b.documentElement,h=b.head||b.getElementsByTagName("head")[0],i="modernizr",j=b.createElement(i),k=
j.style,l=b.createElement("input"),m=":)",n=Object.prototype.toString,o=" -webkit- -moz- -o- -ms- -khtml- ".split(" "),p="Webkit Moz O ms Khtml".split(" "),q={svg:"http://www.w3.org/2000/svg"},r={
},s={},t={},u=[],v,w=function(a){var c=b.createElement("style"),d=b.createElement("div"),e;c.textContent=a+"{#modernizr{height:3px}}",h.appendChild(c),d.id="modernizr",g.appendChild(d),e=d.offsetH
```

Wayback Index (CDX)

- Index for the Wayback Machine
- Generated by parsing crawled (W)ARC data
- Plain text file with one line per captured resource
- Each line contains only essential metadata required by the Wayback software
    - URL, Timestamp, Content Digest
    - MIME Type, HTTP Status Code, Size
    - Meta tags, Redirect URL (when applicable)
    - (W)ARC filename and file offset of record

CDX

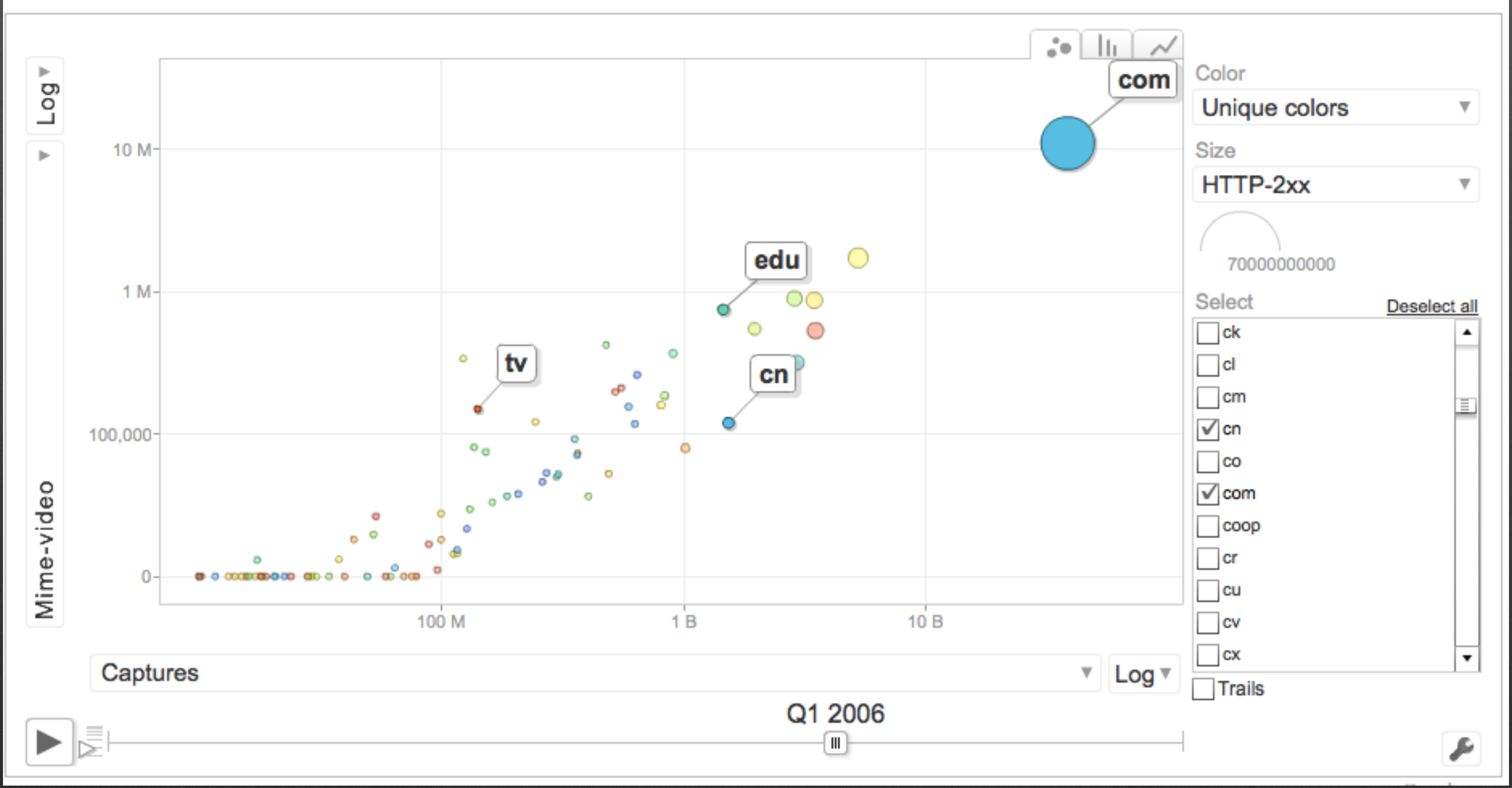


```
gov,dems)/ 20071114154142 http://www.dems.gov/ text/html 200 4ZKRIYZ7PSATVDDXAYMQZQ2MW6YTPZR4 - - 5480 95966912 DOTGOV-EXTRACTION-1995-FY2013-MIME-TEXT-ARCS-PART-01675-000003.arc.gz
gov,dhhs)/ 19970123210343 http://www.dhhs.gov:80/ unk 200 LMO7L7LCAVPMFJFOZ5XUEQJVSQAGUVLC - - 883 60240493 DOTGOV-EXTRACTION-1995-FY2013-MIME-TEXT-ARCS-PART-04528-000005.arc.gz
gov,dhhs)/ 20021010074920 http://www.dhhs.gov:80/ text/html 200 BYZVRP657BQ7ERVM34SJHLXCA3Q7FLYH - - 5659 76533438 DOTGOV-EXTRACTION-1995-FY2013-MIME-TEXT-ARCS-PART-02873-000000.arc.gz
gov,dhhs)/ 20021129083349 http://www.dhhs.gov:80/ text/html 200 PLUQ45BKUK5ZLZRKX7375DE6CEQUKQTZ - - 5640 13145835 DOTGOV-EXTRACTION-1995-FY2013-MIME-TEXT-ARCS-PART-02873-000001.arc.gz
gov,dhhs)/ 20061231175741 http://dhhs.gov:80/ text/html 200 YFMPHYRHYS2LN3FJRX4LV5KCDDLXNJ7E - - 5935 50598441 DOTGOV-EXTRACTION-1995-FY2013-MIME-TEXT-ARCS-PART-03396-000006.arc.gz
gov,dhhs)/ 20070412095843 http://www.dhhs.gov/ text/html 200 M5KSF6KAI4WBYDLKBHFXKRCUQZFDWAPG - - 5625 100176154 DOTGOV-EXTRACTION-1995-FY2013-MIME-TEXT-ARCS-PART-01682-000001.arc.gz
gov,dhhs)/ 20090518071644 http://www.dhhs.gov/ text/html 200 X26DLTGOF3YNQW76BBHPLAG3M6V6V4WZ - - 10486 205491966 DOTGOV-EXTRACTION-1995-FY2013-MIME-TEXT-WARCS-PART-05053-000000.warc.gz
gov,dhhs)/ 20100320154936 http://www.dhhs.gov/ text/html 200 BJ2L3B27GRPCFATCJ36HKUAGW3CVASK5 - - 9279 677827 DOTGOV-EXTRACTION-1995-FY2013-MIME-TEXT-WARCS-PART-02391-000000.warc.gz
gov,doi)/+ 20020812111124 http://www.doi.gov:80/+ text/html 404 2DOPQ2WB3VMK4XQH6QVKNGLAQL7F4RAN - - 607 1928811 DOTGOV-EXTRACTION-1995-FY2013-MIME-TEXT-ARCS-PART-04300-000002.arc.gz
gov,duck)/ 20130925051714 http://duck.gov unk 502 3I42H3S6NNFQ2MSVX7XZKYAYSCX5QBYJ - - 85 33216004 DOTGOV-EXTRACTION-1995-FY2013-MIME-OTHER-ARCS-PART-00043-000000.arc.gz
gov,ed)/ne 20011102113721 http://www.ed.gov:80/NE/ text/html 404 ZEAZRPQ3NSCT5IHMM5AVSX5OI6WAU7WO - - 3470 27334680 DOTGOV-EXTRACTION-1995-FY2013-MIME-TEXT-ARCS-PART-04265-000000.arc.gz
gov,eeoc)/ 20040210020652 http://www.eeoc.gov:80/ text/html 200 3W23HASA24TBBF6DT3QXZYUYRTOHTQUU - - 3827 80338701 DOTGOV-EXTRACTION-1995-FY2013-MIME-TEXT-ARCS-PART-03475-000001.arc.gz
gov,eeoc)/ 20081105224612 http://www.eeoc.gov:80/ text/html 200 MVGQJMCRCGCJEJTTLG7Y6JHU2XYME3FY - - 4381 62070311 DOTGOV-EXTRACTION-1995-FY2013-MIME-TEXT-ARCS-PART-03310-000000.arc.gz
gov,eeoc)/ 20110410044114 http://www.eeoc.gov/ text/html 200 THZXTEMVZ3SVYBEYICF5SXCWZ3IDULDU - - 7412 520600892 DOTGOV-EXTRACTION-1995-FY2013-MIME-TEXT-WARCS-PART-05279-000000.warc.gz
gov,eeoc)/ 20120525032448 http://www.eeoc.gov:80/ text/html 200 5CUQU2MLBBILHVCPGLYJ6EIDEUUJ6JY6 - - 7859 22690402 DOTGOV-EXTRACTION-1995-FY2013-MIME-TEXT-ARCS-PART-04342-000006.arc.gz
gov,eia)/a 20121006212936 http://www.eia.gov/a text/html 302 3I42H3S6NNFQ2MSVX7XZKYAYSCX5QBYJ http://www.eia.gov/404r.cfm?v=http://www.eia.gov/a - 453 213445398 DOTGOV-EXTRACTION-1995-FY2013-MIME-TEXT-WAR
CS-PART-00898-000000.warc.gz
```

CDX Analysis

- Store generated CDX data in Hadoop (HDFS)
- Create Hive table
- Partition the data by partner, collection, crawl instance
    - reduce I/O and query times
- Run queries using HiveQL (a SQL like language)

CDX Analysis: Growth of Content



Color
Unique colors ▼

Size
HTTP-2xx ▼

70000000000

Select
Deselect all

- □ ck
- □ cl
- □ cm
- ☑ cn
- □ co
- ☑ com
- □ coop
- □ cr
- □ cu
- □ cv
- □ cx

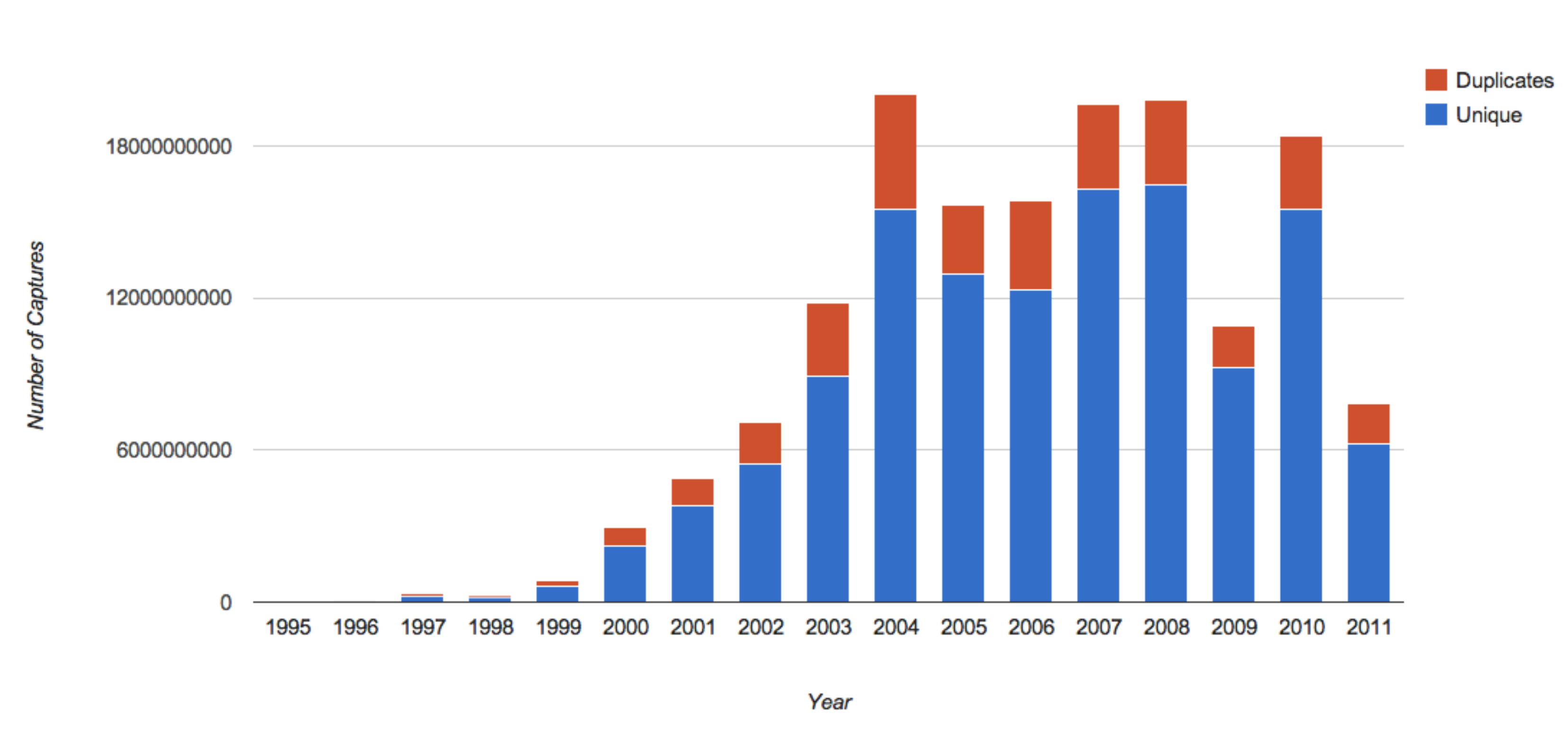
□ Trails

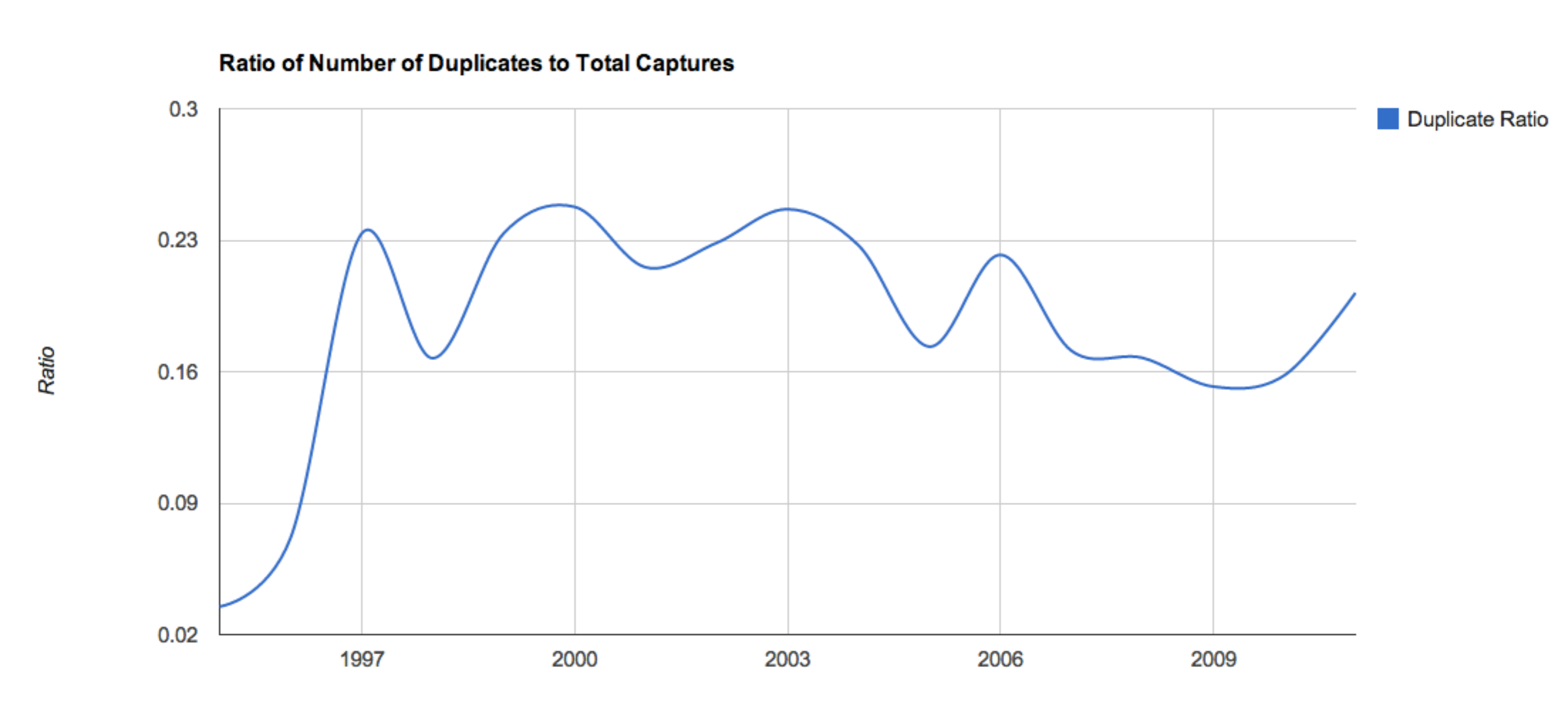
Mime-video / Log ▼

Captures ▼ / Log ▼

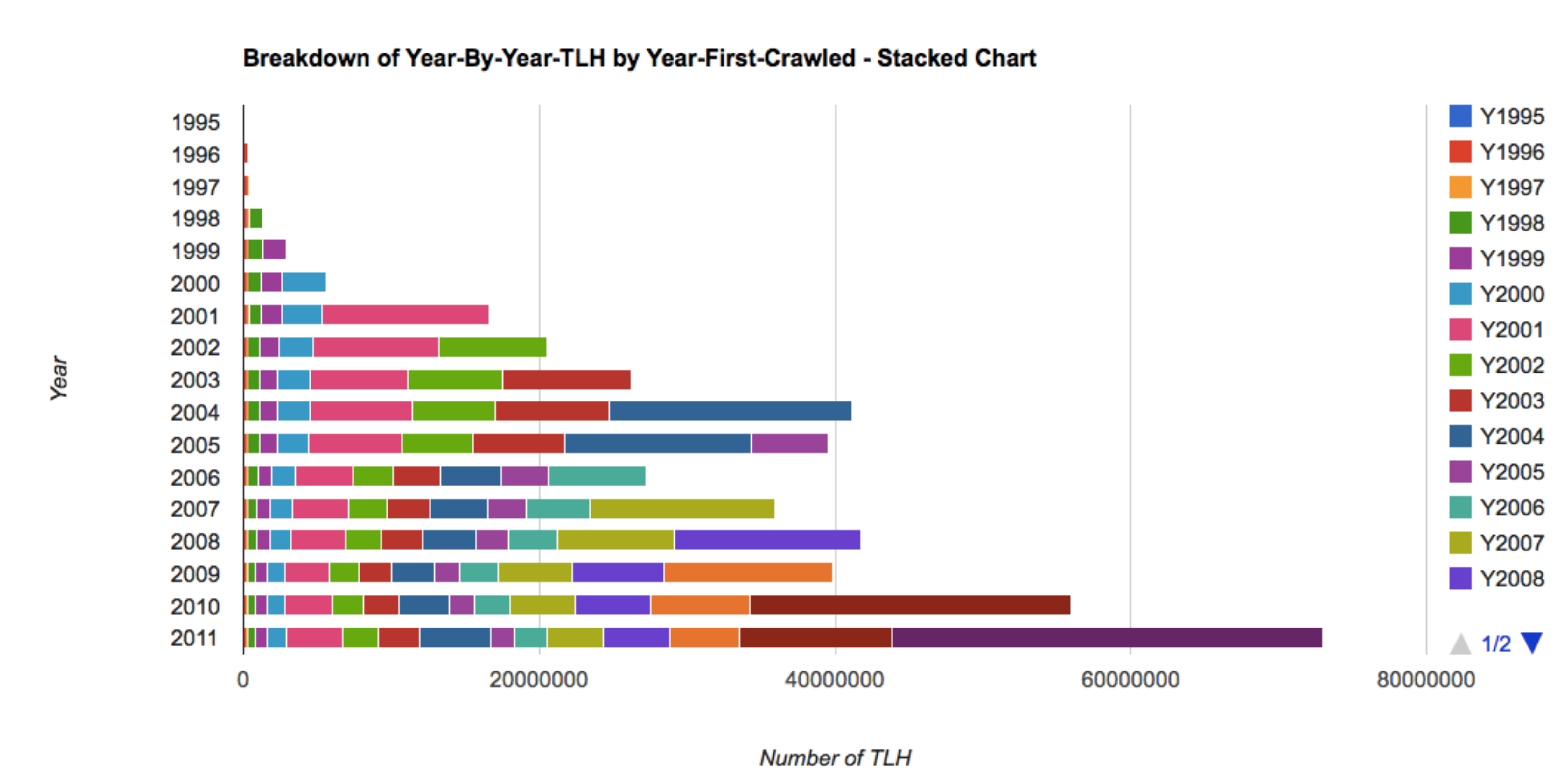
Q1 2006

CDX Analysis: Rate of Duplication





CDX Analysis: Breakdown by Year First Crawled



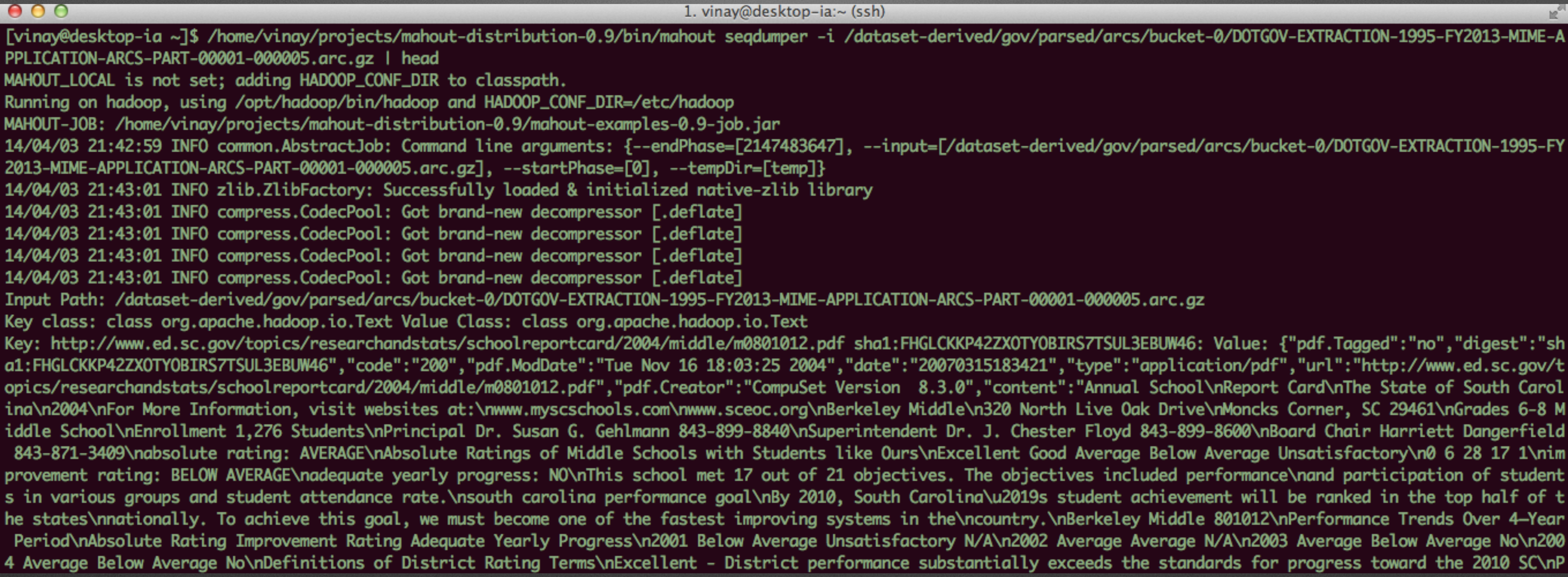
Log Warehouse

- Similar Hive set up for Crawler logs
- Distribution of Domains, HTTP Status codes, MIME types
- Enable crawler engineer to find timeout errors, duplicate content, crawler traps, robots exclusions, etc.

Text Search Index

- Use the Parsed Text files: input to build text indexes for Search
- Generated by running a Hadoop MapReduce Job that parses (W)ARC files
- HTML boilerplate is stripped out
- Also contains metadata
    - URL, Timestamp, Content Digest, Record Length
    - MIME Type, HTTP Status Code
    - Title, description and meta keywords
    - Links with anchor text
- Stored in Hadoop Sequence Files

Parsed Text

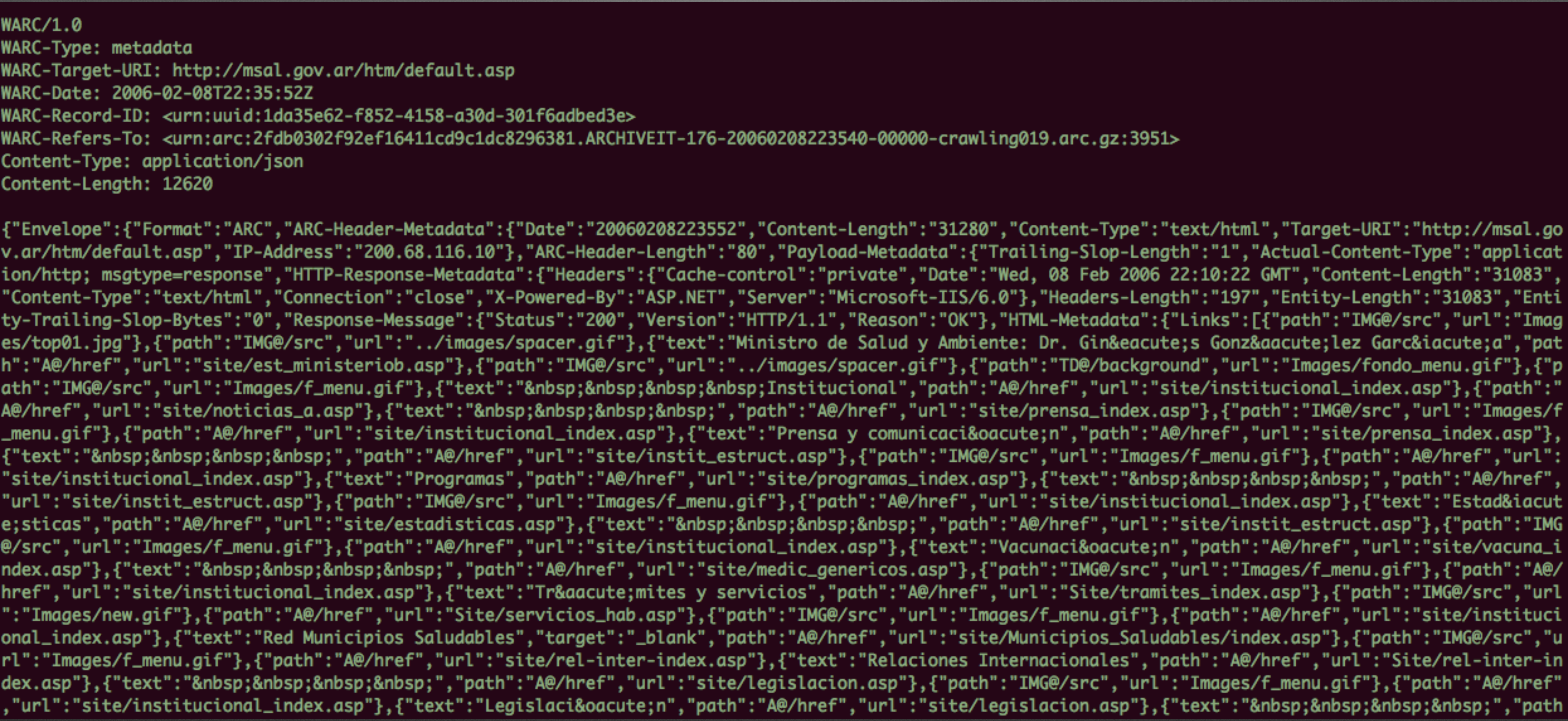


```
[vinay@desktop-ia ~]$ /home/vinay/projects/mahout-distribution-0.9/bin/mahout seqdumper -i /dataset-derived/gov/parsed/arcs/bucket-0/DOTGOV-EXTRACTION-1995-FY2013-MIME-A
PPLICATION-ARCS-PART-00001-000005.arc.gz | head
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using /opt/hadoop/bin/hadoop and HADOOP_CONF_DIR=/etc/hadoop
MAHOUT-JOB: /home/vinay/projects/mahout-distribution-0.9/mahout-examples-0.9-job.jar
14/04/03 21:42:59 INFO common.AbstractJob: Command line arguments: {--endPhase=[2147483647], --input=[/dataset-derived/gov/parsed/arcs/bucket-0/DOTGOV-EXTRACTION-1995-FY
2013-MIME-APPLICATION-ARCS-PART-00001-000005.arc.gz], --startPhase=[0], --tempDir=[temp]}
14/04/03 21:43:01 INFO zlib.ZlibFactory: Successfully loaded & initialized native-zlib library
14/04/03 21:43:01 INFO compress.CodecPool: Got brand-new decompressor [.deflate]
14/04/03 21:43:01 INFO compress.CodecPool: Got brand-new decompressor [.deflate]
14/04/03 21:43:01 INFO compress.CodecPool: Got brand-new decompressor [.deflate]
14/04/03 21:43:01 INFO compress.CodecPool: Got brand-new decompressor [.deflate]
Input Path: /dataset-derived/gov/parsed/arcs/bucket-0/DOTGOV-EXTRACTION-1995-FY2013-MIME-APPLICATION-ARCS-PART-00001-000005.arc.gz
Key class: class org.apache.hadoop.io.Text Value Class: class org.apache.hadoop.io.Text
Key: http://www.ed.sc.gov/topics/researchandstats/schoolreportcard/2004/middle/m0801012.pdf sha1:FHGLCKKP42ZXOTYOBIRS7TSUL3EBUW46: Value: {"pdf.Tagged":"no","digest":"sh
a1:FHGLCKKP42ZXOTYOBIRS7TSUL3EBUW46","code":"200","pdf.ModDate":"Tue Nov 16 18:03:25 2004","date":"20070315183421","type":"application/pdf","url":"http://www.ed.sc.gov/t
opics/researchandstats/schoolreportcard/2004/middle/m0801012.pdf","pdf.Creator":"CompuSet Version  8.3.0","content":"Annual School\nReport Card\nThe State of South Carol
ina\n2004\nFor More Information, visit websites at:\nwww.myscschools.com\nwww.sceoc.org\nBerkeley Middle\n320 North Live Oak Drive\nMoncks Corner, SC 29461\nGrades 6-8 M
iddle School\nEnrollment 1,276 Students\nPrincipal Dr. Susan G. Gehlmann 843-899-8840\nSuperintendent Dr. J. Chester Floyd 843-899-8600\nBoard Chair Harriett Dangerfield
 843-871-3409\nabsolute rating: AVERAGE\nAbsolute Ratings of Middle Schools with Students like Ours\nExcellent Good Average Below Average Unsatisfactory\n0 6 28 17 1\nim
provement rating: BELOW AVERAGE\nadequate yearly progress: NO\nThis school met 17 out of 21 objectives. The objectives included performance\nand participation of student
s in various groups and student attendance rate.\nsouth carolina performance goal\nBy 2010, South Carolina\u2019s student achievement will be ranked in the top half of t
he states\nnationally. To achieve this goal, we must become one of the fastest improving systems in the\ncountry.\nBerkeley Middle 801012\nPerformance Trends Over 4-Year
 Period\nAbsolute Rating Improvement Rating Adequate Yearly Progress\n2001 Below Average Unsatisfactory N/A\n2002 Average Average N/A\n2003 Average Below Average No\n200
4 Average Below Average No\nDefinitions of District Rating Terms\nExcellent - District performance substantially exceeds the standards for progress toward the 2010 SC\nP
```

WAT

- Extensible metadata format
- Essential metadata for many types of analyses
- Avoids barriers to data exchange: copyright, privacy
- Less data than (W)ARC, more than CDX
- WAT records are WARC metadata records
- Contains for every HTML page in the (W)ARC,
    - Title, description and meta keywords
    - Embeds and outgoing links with alt/anchor text

WAT



```
WARC/1.0
WARC-Type: metadata
WARC-Target-URI: http://msal.gov.ar/htm/default.asp
WARC-Date: 2006-02-08T22:35:52Z
WARC-Record-ID: <urn:uuid:1da35e62-f852-4158-a30d-301f6adbed3e>
WARC-Refers-To: <urn:arc:2fdb0302f92ef16411cd9c1dc8296381.ARCHIVEIT-176-20060208223540-00000-crawling019.arc.gz:3951>
Content-Type: application/json
Content-Length: 12620
```

```
{"Envelope":{"Format":"ARC","ARC-Header-Metadata":{"Date":"20060208223552","Content-Length":"31280","Content-Type":"text/html","Target-URI":"http://msal.go
v.ar/htm/default.asp","IP-Address":"200.68.116.10"},"ARC-Header-Length":"80","Payload-Metadata":{"Trailing-Slop-Length":"1","Actual-Content-Type":"applicat
ion/http; msgtype=response","HTTP-Response-Metadata":{"Headers":{"Cache-control":"private","Date":"Wed, 08 Feb 2006 22:10:22 GMT","Content-Length":"31083",
"Content-Type":"text/html","Connection":"close","X-Powered-By":"ASP.NET","Server":"Microsoft-IIS/6.0"},"Headers-Length":"197","Entity-Length":"31083","Enti
ty-Trailing-Slop-Bytes":"0","Response-Message":{"Status":"200","Version":"HTTP/1.1","Reason":"OK"},"HTML-Metadata":{"Links":[{"path":"IMG@/src","url":"Imag
es/top01.jpg"},{"path":"IMG@/src","url":"../images/spacer.gif"},{"text":"Ministro de Salud y Ambiente: Dr. Gin&eacute;s Gonz&aacute;lez Garc&iacute;a","pat
h":"A@/href","url":"site/est_ministeriob.asp"},{"path":"IMG@/src","url":"../images/spacer.gif"},{"path":"TD@/background","url":"Images/fondo_menu.gif"},{"p
ath":"IMG@/src","url":"Images/f_menu.gif"},{"text":"&nbsp;&nbsp;&nbsp;&nbsp;Institucional","path":"A@/href","url":"site/institucional_index.asp"},{"path":"
A@/href","url":"site/noticias_a.asp"},{"text":"&nbsp;&nbsp;&nbsp;&nbsp;","path":"A@/href","url":"site/prensa_index.asp"},{"path":"IMG@/src","url":"Images/f
_menu.gif"},{"path":"A@/href","url":"site/institucional_index.asp"},{"text":"Prensa y comunicaci&oacute;n","path":"A@/href","url":"site/prensa_index.asp"},
{"text":"&nbsp;&nbsp;&nbsp;&nbsp;","path":"A@/href","url":"site/instit_estruct.asp"},{"path":"IMG@/src","url":"Images/f_menu.gif"},{"path":"A@/href","url":
"site/institucional_index.asp"},{"text":"Programas","path":"A@/href","url":"site/programas_index.asp"},{"text":"&nbsp;&nbsp;&nbsp;&nbsp;","path":"A@/href",
"url":"site/instit_estruct.asp"},{"path":"IMG@/src","url":"Images/f_menu.gif"},{"path":"A@/href","url":"site/institucional_index.asp"},{"text":"Estad&iacut
e;sticas","path":"A@/href","url":"site/estadisticas.asp"},{"text":"&nbsp;&nbsp;&nbsp;&nbsp;","path":"A@/href","url":"site/instit_estruct.asp"},{"path":"IMG
@/src","url":"Images/f_menu.gif"},{"path":"A@/href","url":"site/institucional_index.asp"},{"text":"Vacunaci&oacute;n","path":"A@/href","url":"site/vacuna_i
ndex.asp"},{"text":"&nbsp;&nbsp;&nbsp;&nbsp;","path":"A@/href","url":"site/medic_genericos.asp"},{"path":"IMG@/src","url":"Images/f_menu.gif"},{"path":"A@/
href","url":"site/institucional_index.asp"},{"text":"Tr&aacute;mites y servicios","path":"A@/href","url":"Site/tramites_index.asp"},{"path":"IMG@/src","url
":"Images/new.gif"},{"path":"A@/href","url":"Site/servicios_hab.asp"},{"path":"IMG@/src","url":"Images/f_menu.gif"},{"path":"A@/href","url":"site/instituci
onal_index.asp"},{"text":"Red Municipios Saludables","target":"_blank","path":"A@/href","url":"site/Municipios_Saludables/index.asp"},{"path":"IMG@/src","u
rl":"Images/f_menu.gif"},{"path":"A@/href","url":"site/rel-inter-index.asp"},{"text":"Relaciones Internacionales","path":"A@/href","url":"Site/rel-inter-in
dex.asp"},{"text":"&nbsp;&nbsp;&nbsp;&nbsp;","path":"A@/href","url":"site/legislacion.asp"},{"path":"IMG@/src","url":"Images/f_menu.gif"},{"path":"A@/href"
,"url":"site/institucional_index.asp"},{"text":"Legislaci&oacute;n","path":"A@/href","url":"site/legislacion.asp"},{"text":"&nbsp;&nbsp;&nbsp;&nbsp;","path
```

Text Analysis

- Text extracted from (W)ARC / Parsed Text / WAT
- Use Pig
    - extract text from records of interest
    - tokenize, remove stop words, stemming
    - generate top terms by TF-IDF
    - prepare text for input to Mahout to generate vectorized documents (Topic Modeling, Classification, Clustering etc.)

Link Analysis

- Links extracted from crawl logs / WARC metadata records / Parsed Text / WAT
- Use Pig
    - extract links from records of interest
    - generate host & domain graphs for a given period
    - find links in common between a pair of hosts/domains
    - extract embedded links and compare with CDX to find resources yet to be crawled

Archival Web Graph

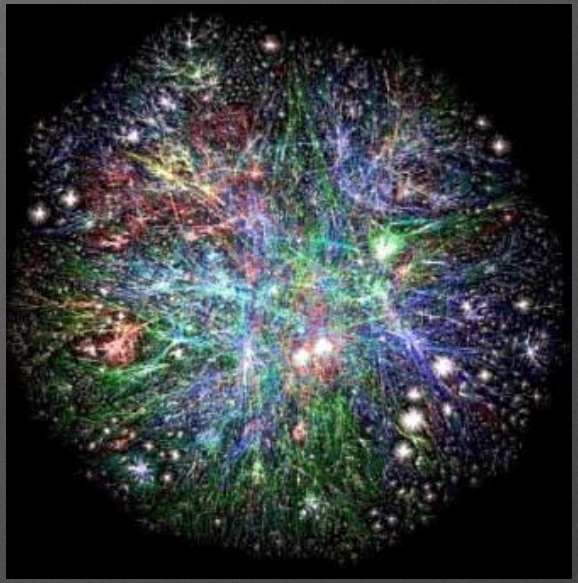
- Use Pig to generate an Archival Web Graph (ID-Map and ID-Graph)
- ID-Map: Mapping of integer (or fingerprint) ID to source and destination URLs
- ID-Graph: An adjacency list using the assigned IDs and timestamp info
- Compact representation of graph data

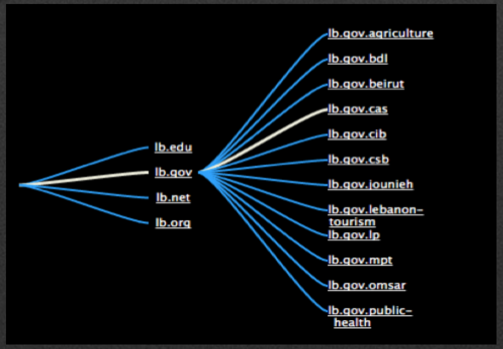
Link Analysis using Giraph

- Hadoop MapReduce not the best fit for iterative algorithms
    - each iteration is a MapReduce Job with the graph structure being read from and written to HDFS
- Use Giraph: open-source implementation of Google's Pregel
    - Vertex centric Bulk Synchronous Parallel (BSP) execution model
    - runs on Hadoop
    - computation executed in memory and proceeds as sequence of iterations called supersteps

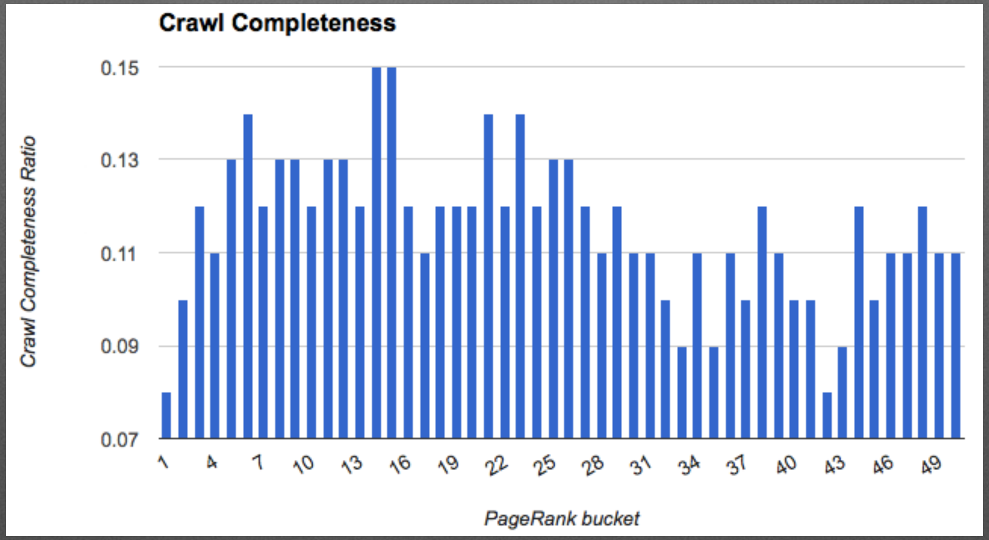
Link Analysis

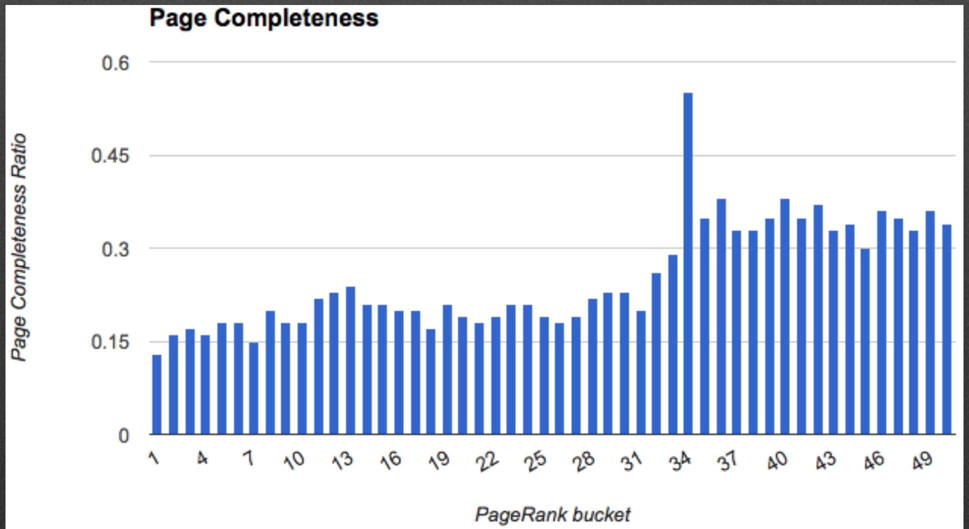
- Indegree and Outdegree distributions
- Inter-host and Intra-host link information
- Rank resources by PageRank
    - Identify important resources
    - Prioritize crawling of missing resources
- Find possible spam pages by running biased PageRank
- Trace path of crawler using graph generated from crawl logs
- Determine Crawl and Page Completeness



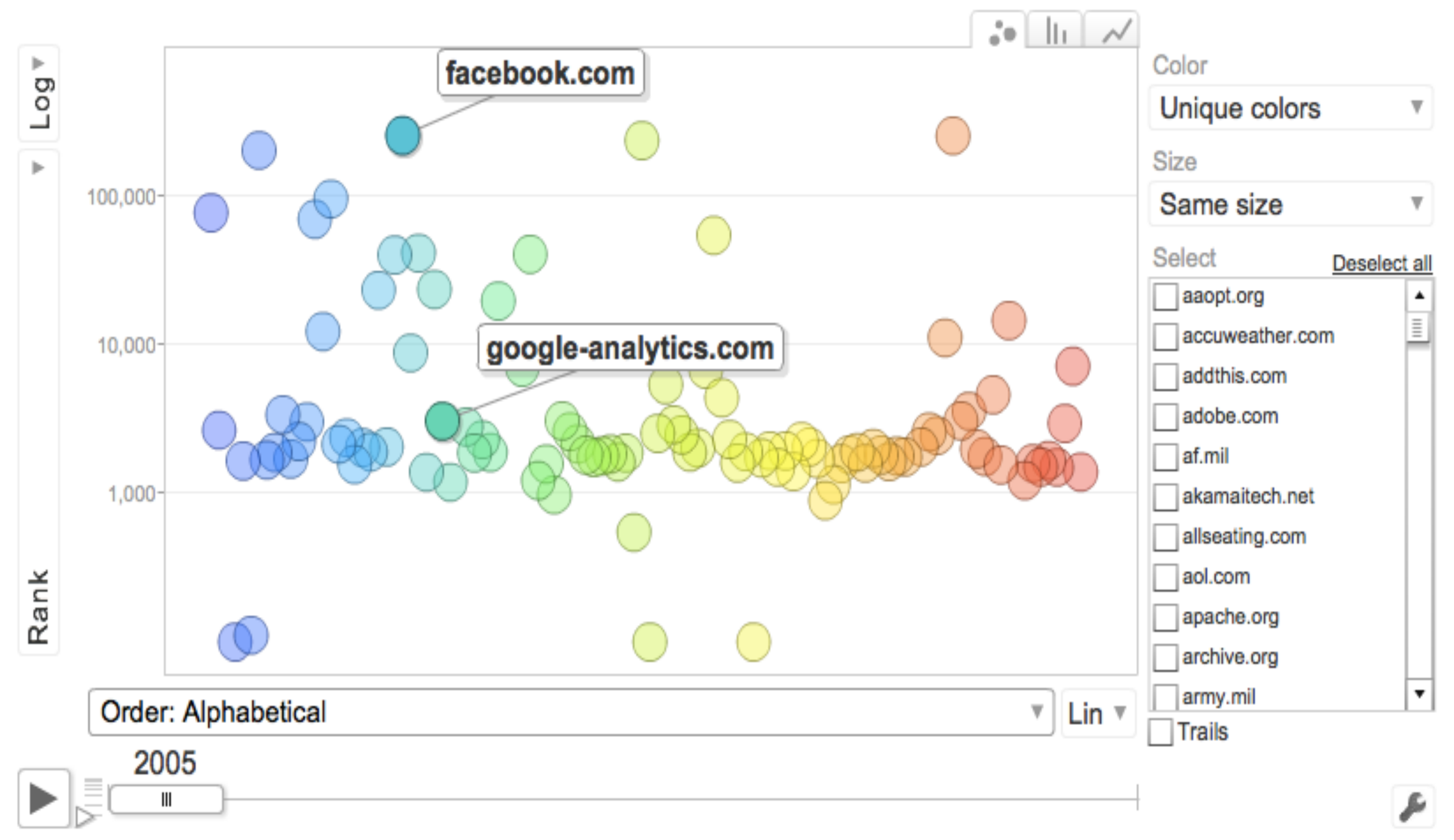


Link Analysis: Completeness





Link Analysis: PageRank over Time



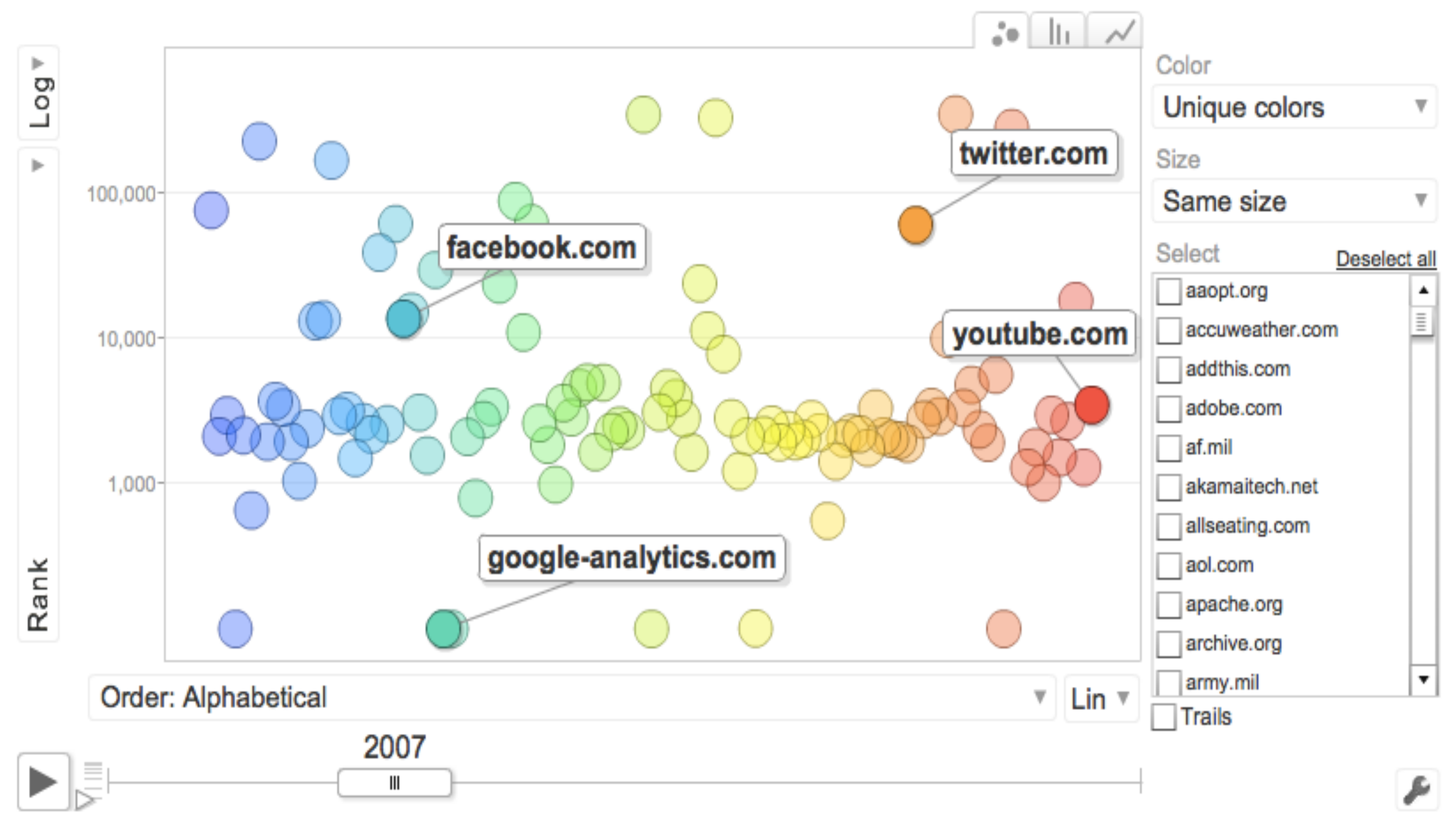
Color
Unique colors

Size
Same size

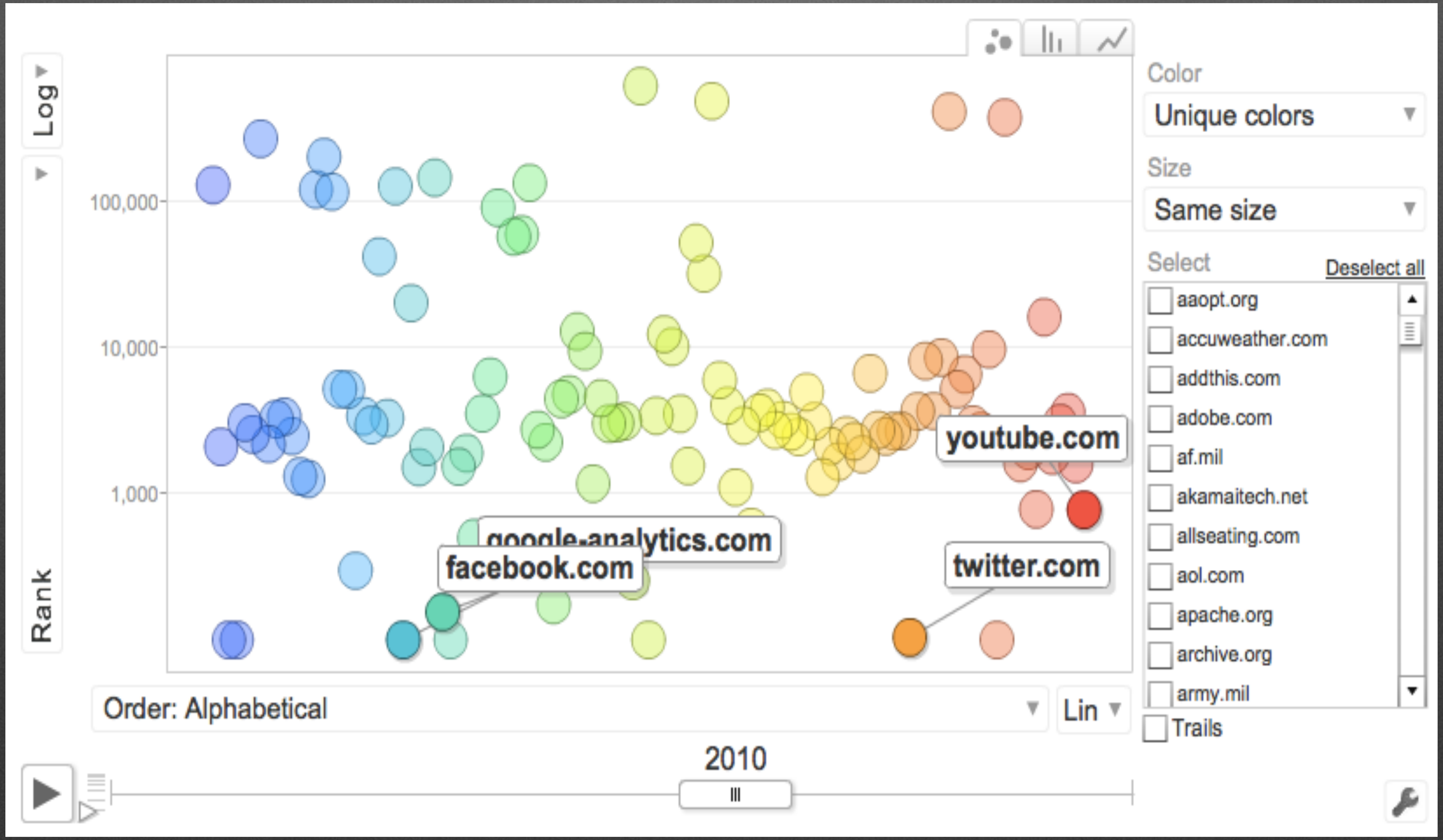
Order: Alphabetical

Lin

Link Analysis: PageRank over Time



Link Analysis: PageRank over Time



Color: Unique colors

Size: Same size

Order: Alphabetical

Lin

2010

Web Archive Analysis Workshop

- Self guided workshop
- Generative derivatives: CDX, Parsed Text, WAT
- Set up CDX Warehouse using Hive
- Extract text from WARCs / WAT / Parsed Text
- Extract links from WARCs / WAT / Parsed Text
- Generate Archival web graphs, host and domain graphs
- Text and Link Analysis Examples
- Data extraction tools to repackage subsets of data into new (W)ARC files
- https://webarchive.jira.com/wiki/display/Iresearch/Web+Archive+Analysis+Workshop