VT Web Archiving

Anthony Rinaldi and Dev Mehta
CS 4624
Clients: Mohamed Magdy and Tarek Kanan
Blacksburg, VA
3/5/2014

Heritirx

- Java based Web Crawler
- Web Archiving
- Web Browser or Command Line

Wayback Machine

- Front End for Heritrix-based crawls
- WARC Files
- Archive Search Engine
- http://archive.org/web/

INTERNET ARCHIVE
WayBackMachine

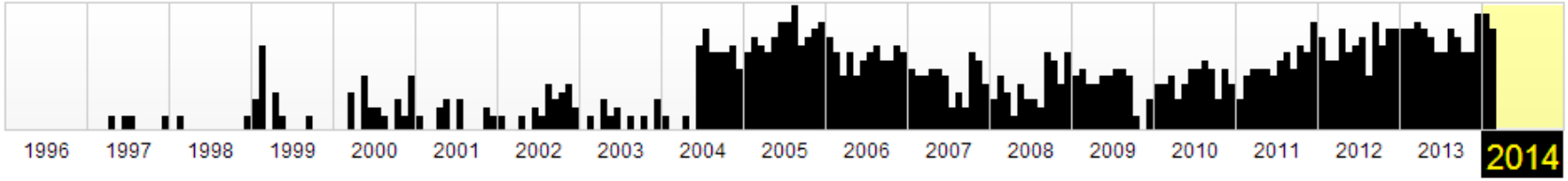
http://www.vt.edu

BROWSE HISTORY

http://www.vt.edu

Saved **2,147 times** between April 1, 1997 and February 24, 2014.

PLEASE DONATE TODAY. Your generosity preserves knowledge for future generations. Thank you.



JAN

		1	2	3	4	
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

FEB

						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	

MAR

						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

APR

		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30			

Advanced search
Language tools

Google Search I'm Feeling Lucky

What did the world search for this year? Remember the moments of 2013

Solr

- Search Engine
- Feed WARC Files
- Used to index web pages archived by Heritrix

Drupal

- Front end to Solr
- Web publishing system
- Content management

Status

Complete required reading on Heritrix and test installations - 17 February

Install and initial setup of Heritrix on main server - 28 February

Set-up Wayback machine in java to read warc files from heritrix. - Early April

Set-up Solr to apply search filters to warc files. - End April

Sources

- https://webarchive.jira.com/wiki/display/Heritrix/Heritrix
- http://archive.org/web/
- https://lucene.apache.org/solr/
- https://drupal.org/

Questions?