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
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](https://webarchive.jira.com/wiki/display/Heritrix/Heritrix)
- [https://lucene.apache.org/solr/](https://lucene.apache.org/solr/)
- [https://drupal.org/](https://drupal.org/)
Questions?