9-23-2013

On the Change in Archivability of Websites Over Time

Mat Kelly
Old Dominion University

Justin F. Brunelle

Michele C. Weigle
Old Dominion University, mweigle@odu.edu

Michael L. Nelson
Old Dominion University, mnelson@odu.edu

On the Change in Archivability of Websites Over Time

Mat Kelly, Justin F. Brunelle, Michele C. Weigle, and Michael L. Nelson

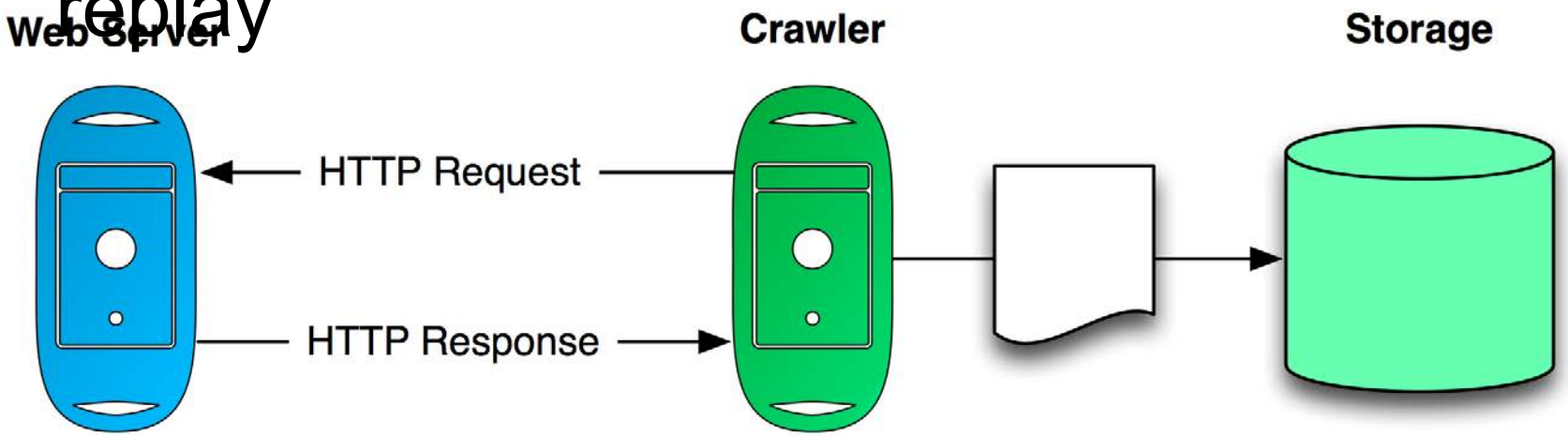
Old Dominion University

{mkelly,jbrunelle,mweigle,mln}@cs.odu.edu

TPDL 2013, Valletta, Malta,
September 23, 2013

Preserving Web Pages

- Identify page content needed for re-render
- Save page contents
  - HTML, Images, CSS, JavaScript
- Rewrite inter-resource references for replay



From M. Klein, J. F. Brunelle, WADL 2013

Ease of Archiving

- **HTML** – text-based, references other resources
- Images
- CSS
- JavaScript

EXAMPLE

```
<html><body>
   <img src="meatball-logo.png" />
</body></html>
```

Ease of Archiving

- HTML
- **Images** – binary data, no embedded URIs
- CSS
- JavaScript

EXAMPLE

Ease of Archiving

- HTML
- Images
- **CSS** – text-based, references other resources
- JavaScript

EXAMPLE

```
body {
    margin: 5px;
    color: black;
    background-image: url('outerSpace.png');
}
```

Ease of Archiving

- HTML
- Images
- CSS
- **JavaScript** – text-based, references other resources, URIs (not necessarily known until runtime)

EXAMPLE

```
$.ajax({
    url: “meatball-logo” +“.”+“png”; //build logo URI at runtime
});
```

Ease of Archiving

- HTML
- Images
- CSS
- **JavaScript** – text-based, references other resources, URIs (not necessarily known until runtime)

EXAMPLE

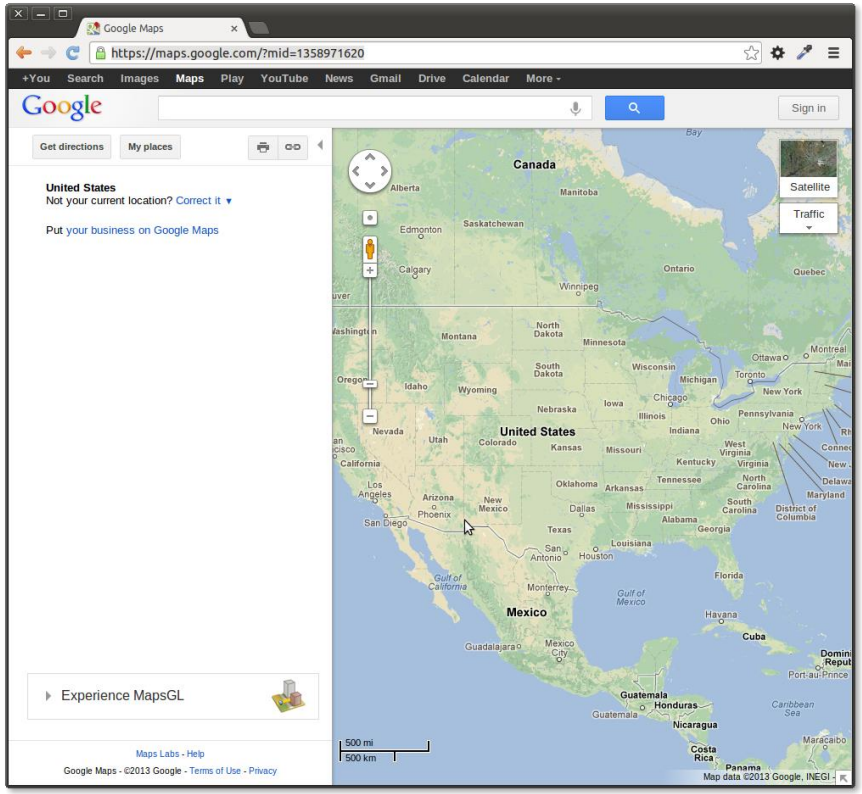
```
$.ajax({
    url: “wormlogo” +“.”+“png”; //build logo URI at runtime
});
```

BIG Problem

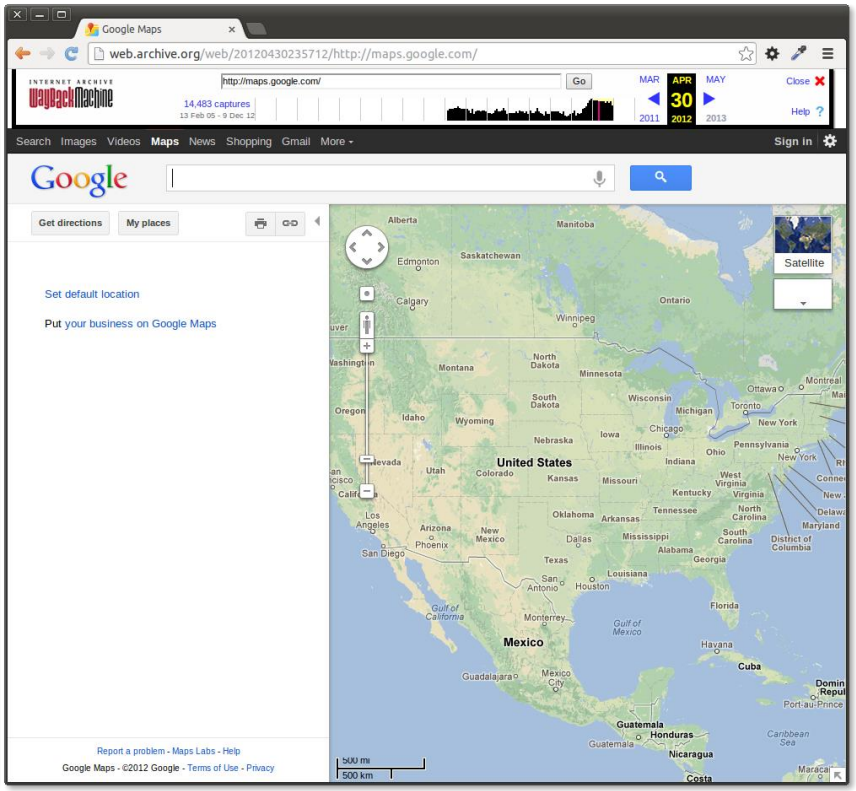
- JavaScript meant for browser
  - Capable of JS Execution
- Crawlers only recently became capable of exec
- Crawlers getting smarter at extracting URIs
  - Nowhere near perfect

Identifying Missing Resources

- Sometimes missing resources are subtle



From Live Web

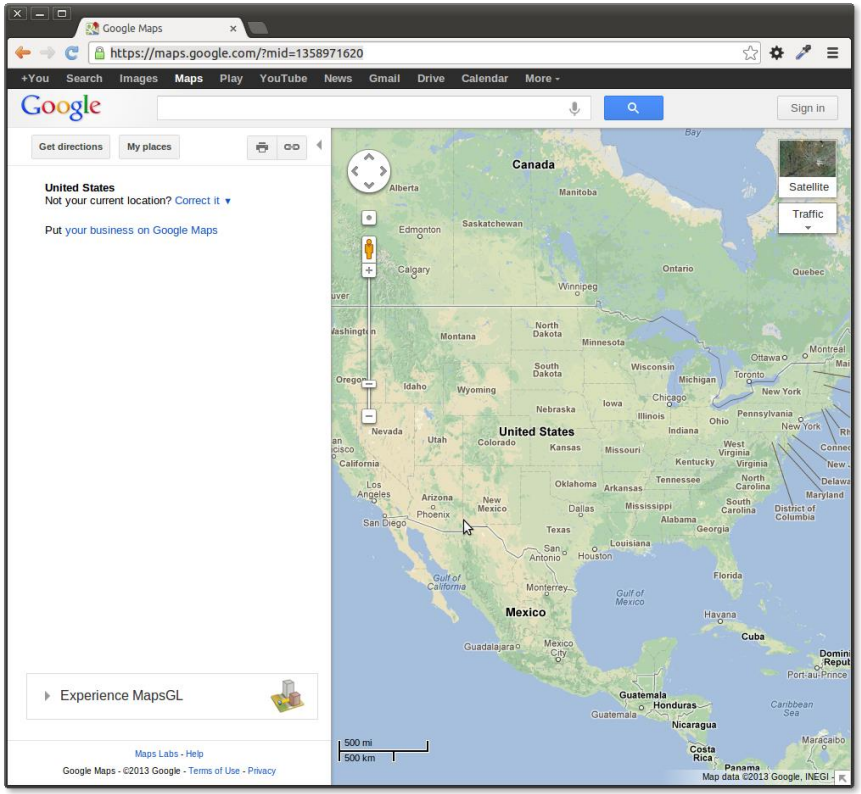


From Internet Archive Apr. 30, 2012

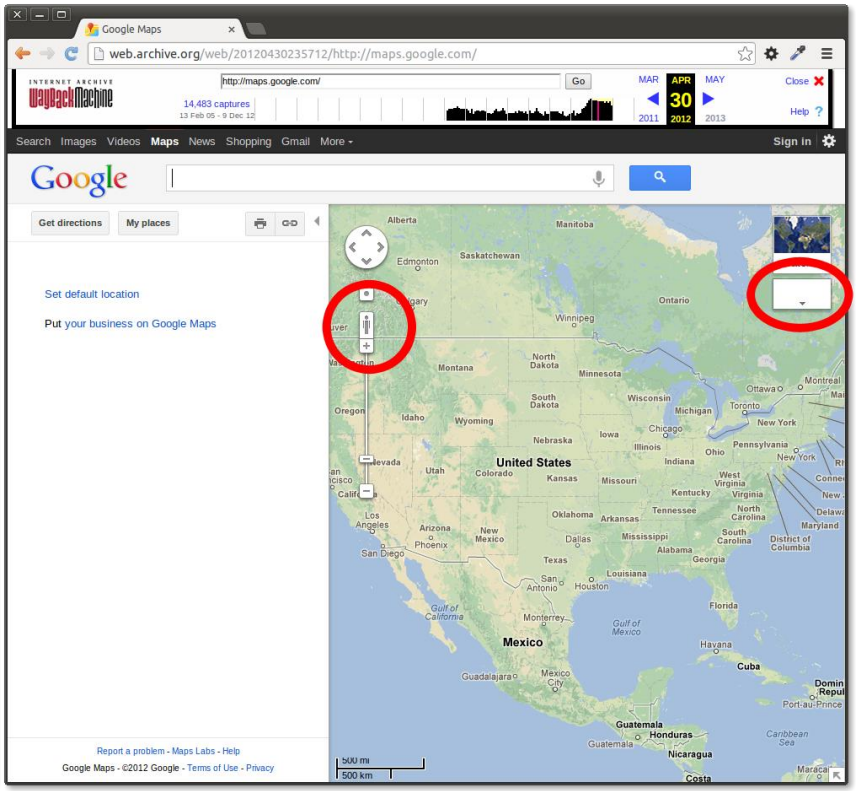
http://web.archive.org/web/20120430235712/http://maps.google.com

Identifying Missing Resources

- Archived version not interactive & resources are missing



From Live Web

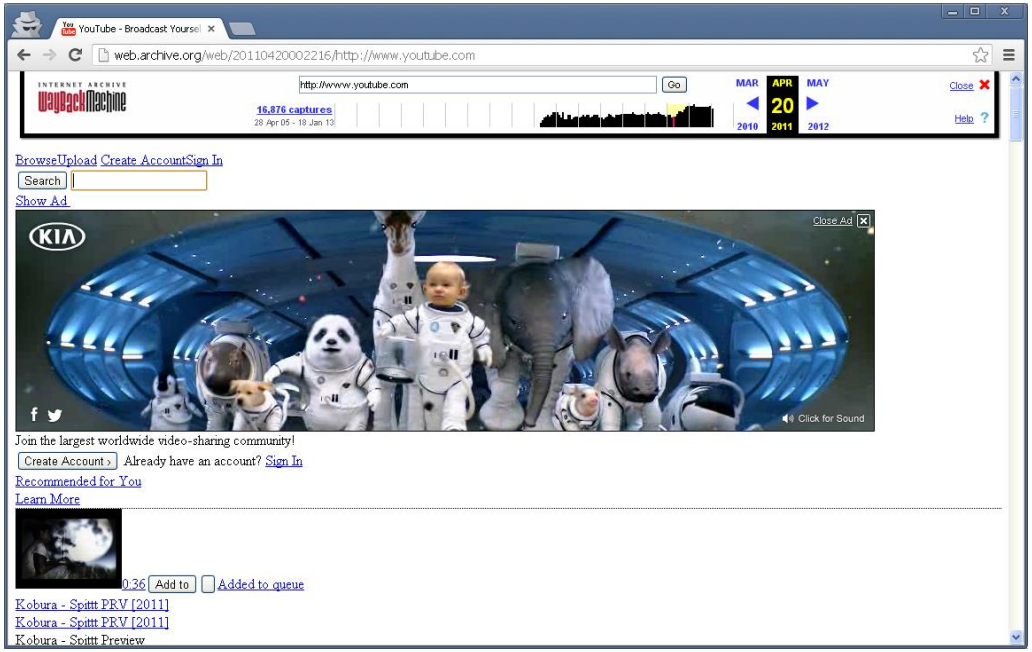


From Internet Archive Apr. 30, 2012

http://web.archive.org/web/20120430235712/http://maps.google.com

Identifying Missing Resources

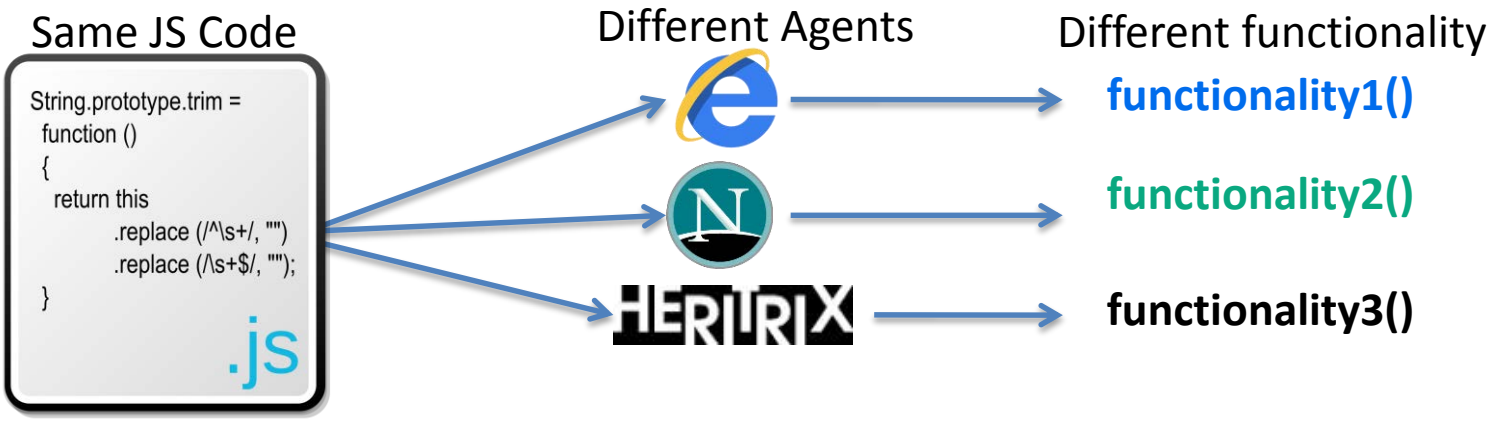
- Other times, a failed AJAX call prevents other resources from loading



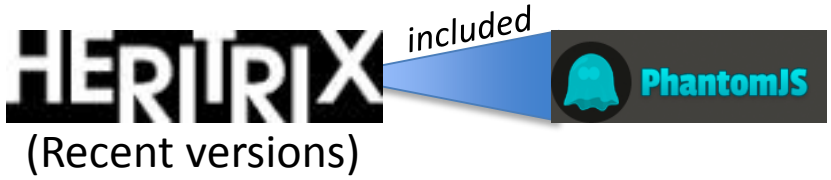
YouTube (2011) in archives with failed Ajax call

http://web.archive.org/web/20110420002216/http://www.youtube.com/

Why JavaScript makes it difficult



- Archival crawlers don't interpret DOM, are made for capture and thus faster



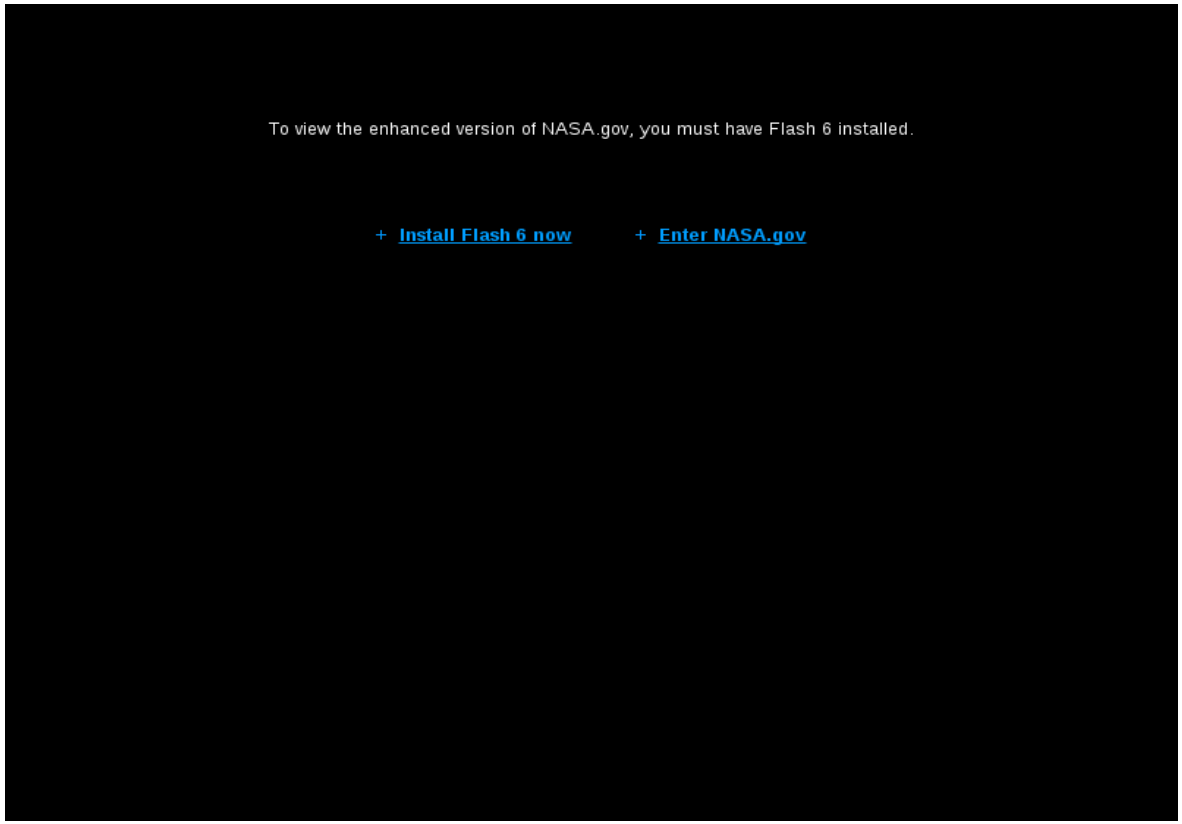
- Interprets JavaScript
- Enabled pages w/ JS to be archived better!

JavaScript and Accessibility

- Content displaying should not be dependent on script execution

- U.S. Government
  - required to comply with accessibility standards
  - Content should be available to crawler w/o JS
  - Therefore, government sites are better preserved
    - Right?

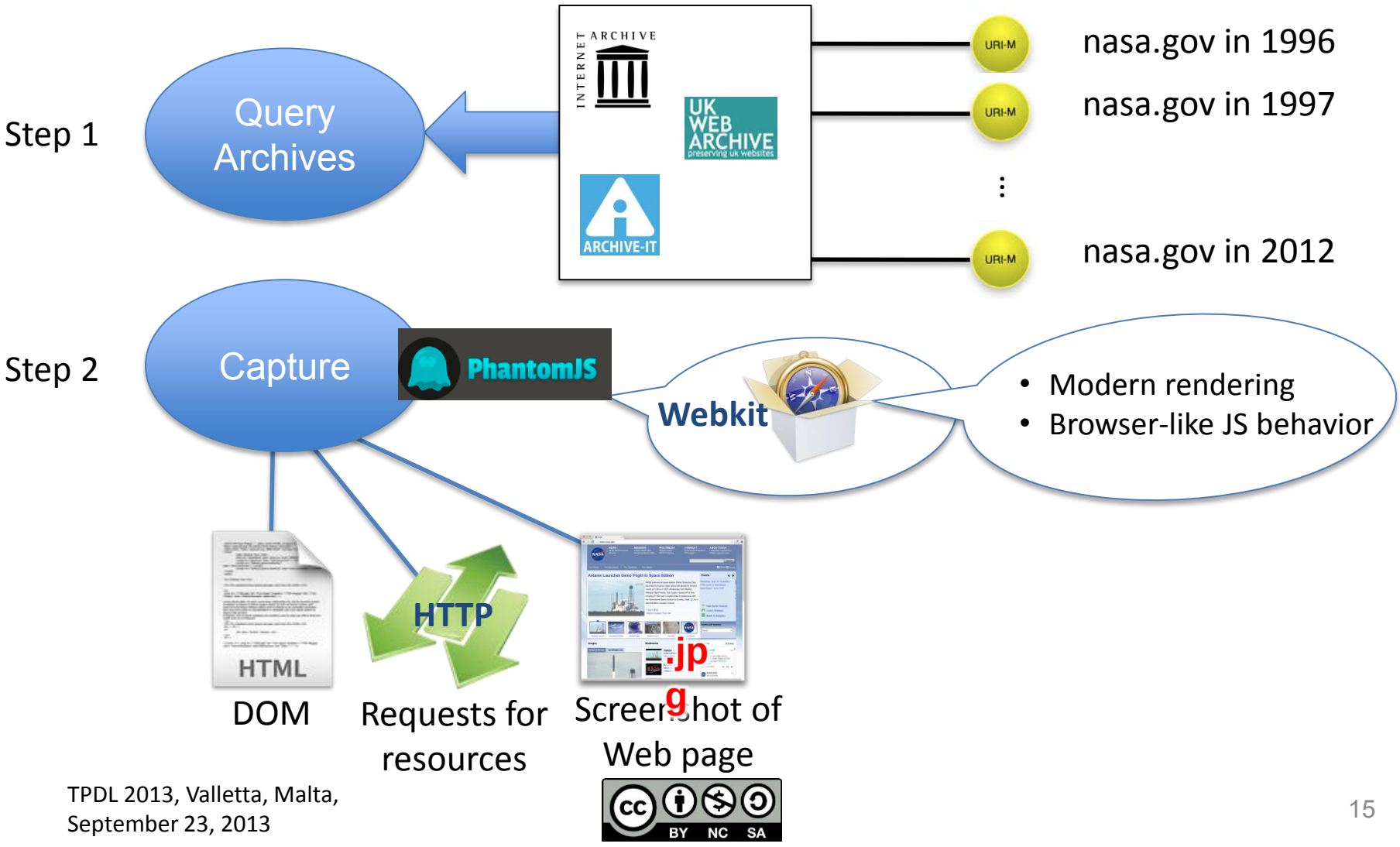
If not Accessible then not Archivable



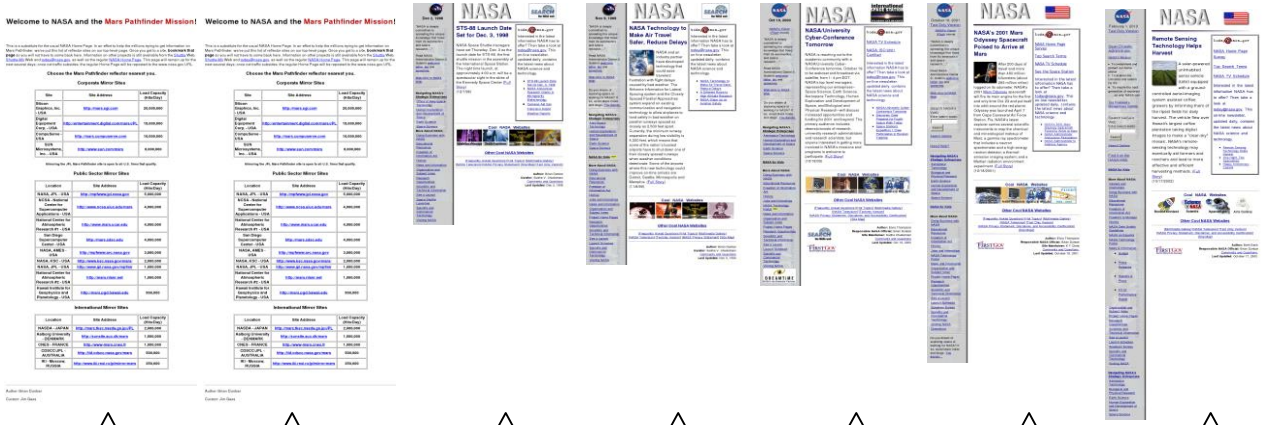
NASA.gov in Internet Archive, 2004

http://web.archive.org/web/20041014205942/http://www.nasa.gov

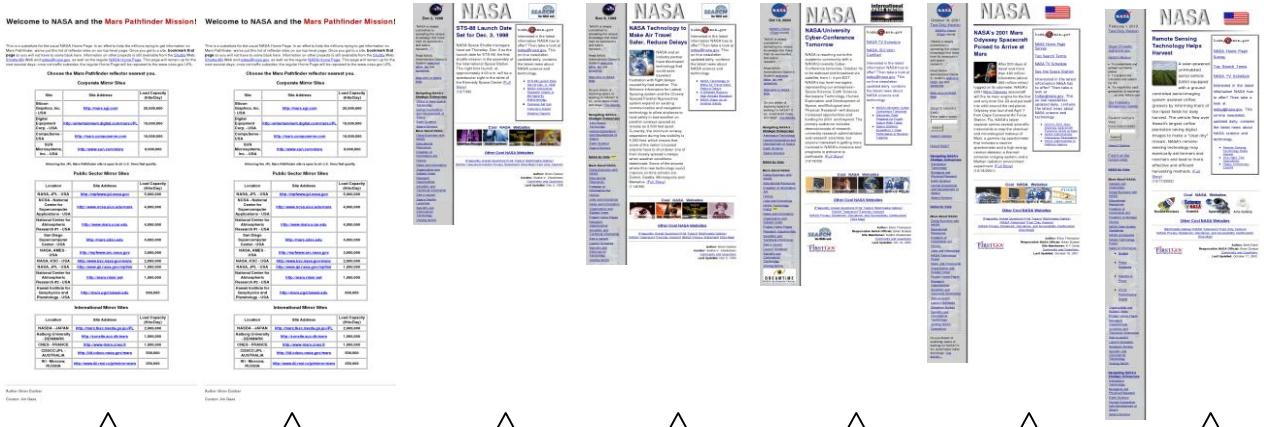
Finding out where NASA went wrong

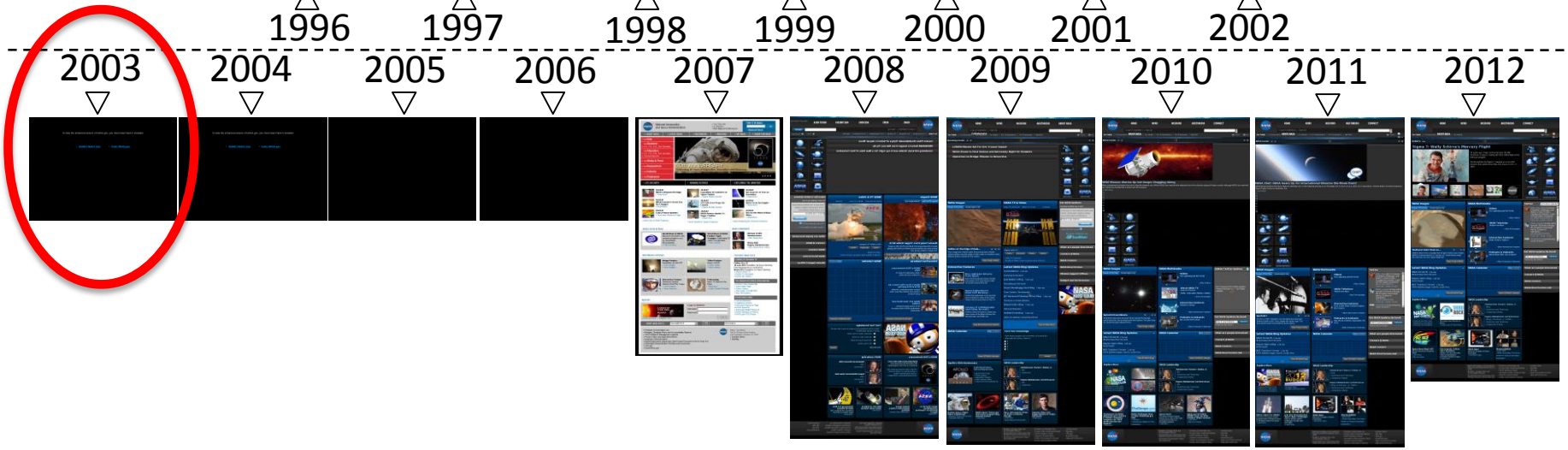


NASA.gov over time



NASA.gov over time





Does this apply to popular sites?

- Alexa's Top 10 websites

Alexa Rank	Web Site Name
1	Facebook.com
2	Google.com
3	YouTube.com
4	Yahoo.com
5	Baidu.com
6	Wikipedia.org
7	Live.org
8	Amazon.com
9	QQ.com
10	Twitter.com

Does this apply to popular sites?

- Alexa's Top 10 websites
- No Mementos
  - robots.txt exclusion prevents crawl

Alexa Rank	Web Site Name
1	Facebook.com
2	Google.com
3	YouTube.com
4	Yahoo.com
5	Baidu.com
6	Wikipedia.org
7	Live.org
8	Amazon.com
9	QQ.com
10	Twitter.com

Does this apply to popular sites?

Alexa Rank	Web Site Name	Sampled Mementos
1	Facebook.com	No memento robots.txt exclusion
2	Google.com	15 mementos 1998 to 2012
3	YouTube.com	7 mementos 2006 to 2012
4	Yahoo.com	16 mementos 1997 to 2012
5	Baidu.com	No memento robots.txt exclusion
6	Wikipedia.org	12 mementos 2001 to 2012
7	Live.org	15 mementos 1999 to 2012
8	Amazon.com	14 mementos 1999 to 2012
9	QQ.com	15 mementos 1998 to 2012
10	Twitter.com	No memento robots.txt exclusion

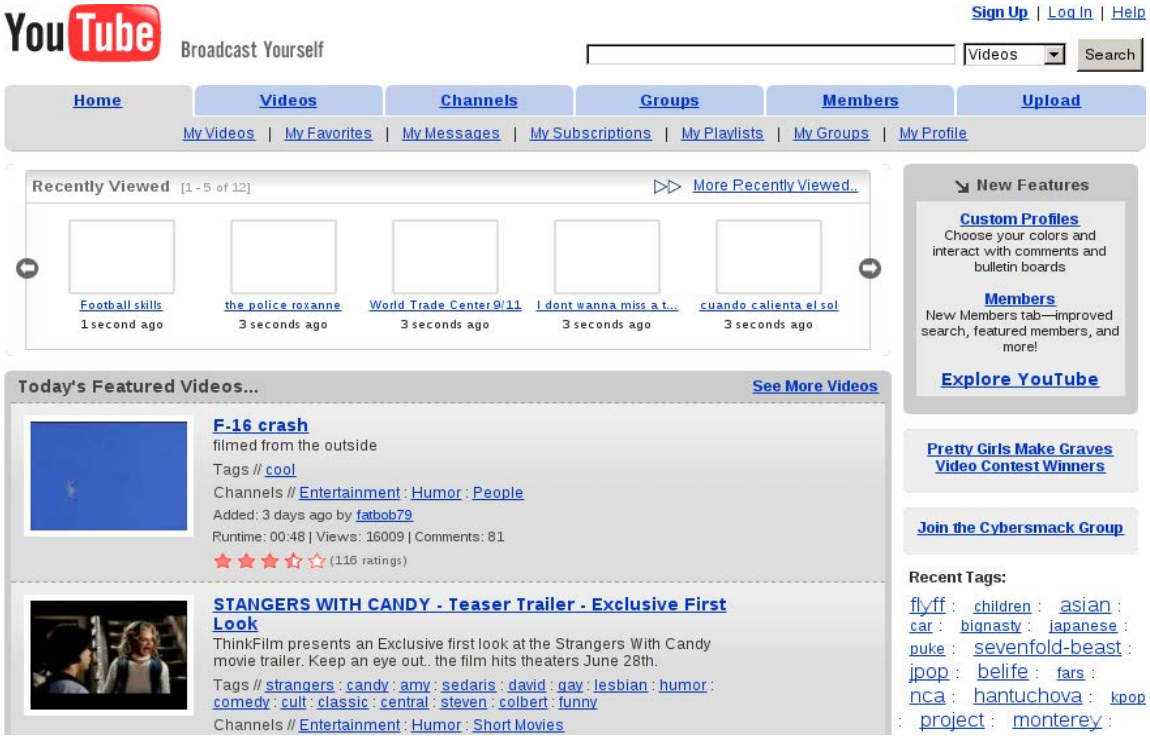
all thumbnails at: http://www.cs.odu.edu/~mkelly/semester/2013_spring/20130127alexatop10/

Case Study with Ajax emphasis: YouTube

Alexa Rank	Web Site Name	Sampled Mementos
1	Facebook.com	No memento robots.txt exclusion
2	Google.com	15 mementos 1998 to 2012
3	YouTube.com	7 mementos 2006 to 2012
4	Yahoo.com	16 mementos 1997 to 2012
5	Baidu.com	No memento robots.txt exclusion
6	Wikipedia.org	12 mementos 2001 to 2012
7	Live.org	15 mementos 1999 to 2012
8	Amazon.com	14 mementos 1999 to 2012
9	QQ.com	15 mementos 1998 to 2012
10	Twitter.com	No memento robots.txt exclusion

Ajax's Effects on the Archivability of YouTube

- Recently Viewed content missing

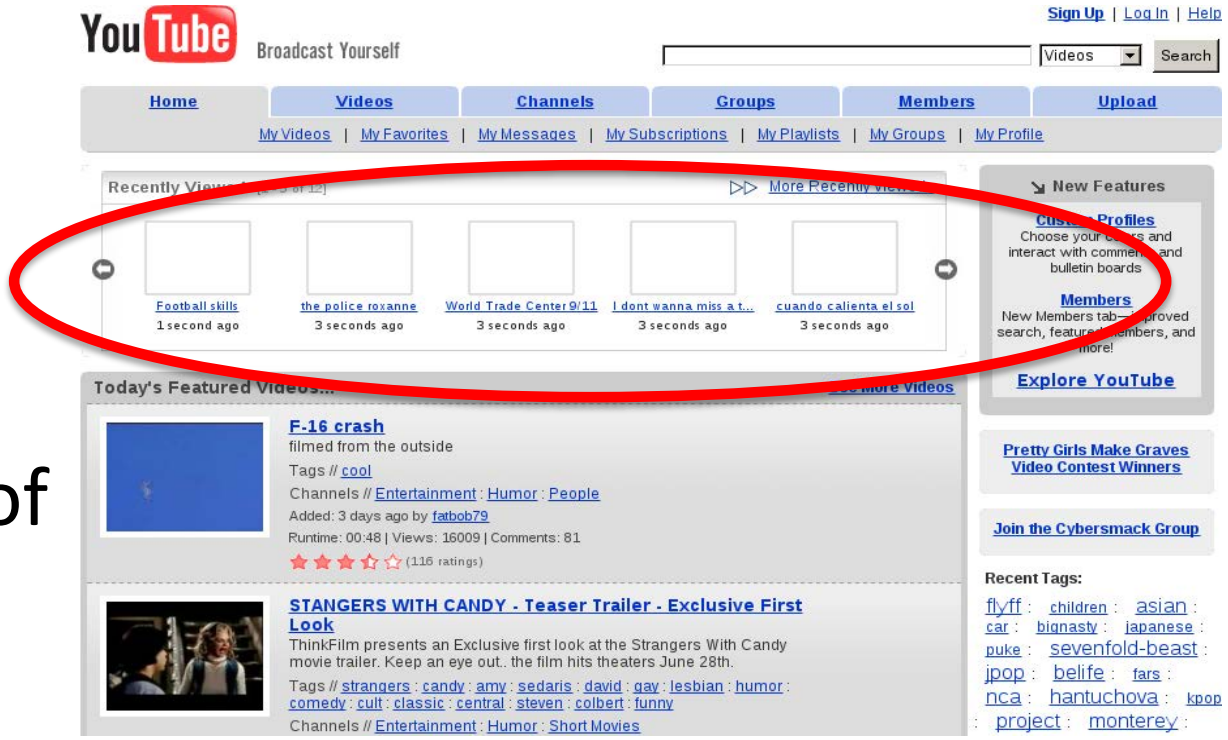


2006 YouTube.com from Internet Archive

http://web.archive.org/web/20060427213420/http://youtube.com

Ajax's Effects on the Archivability of YouTube

- Recently Viewed content missing
- How much of this is because of JavaScript?

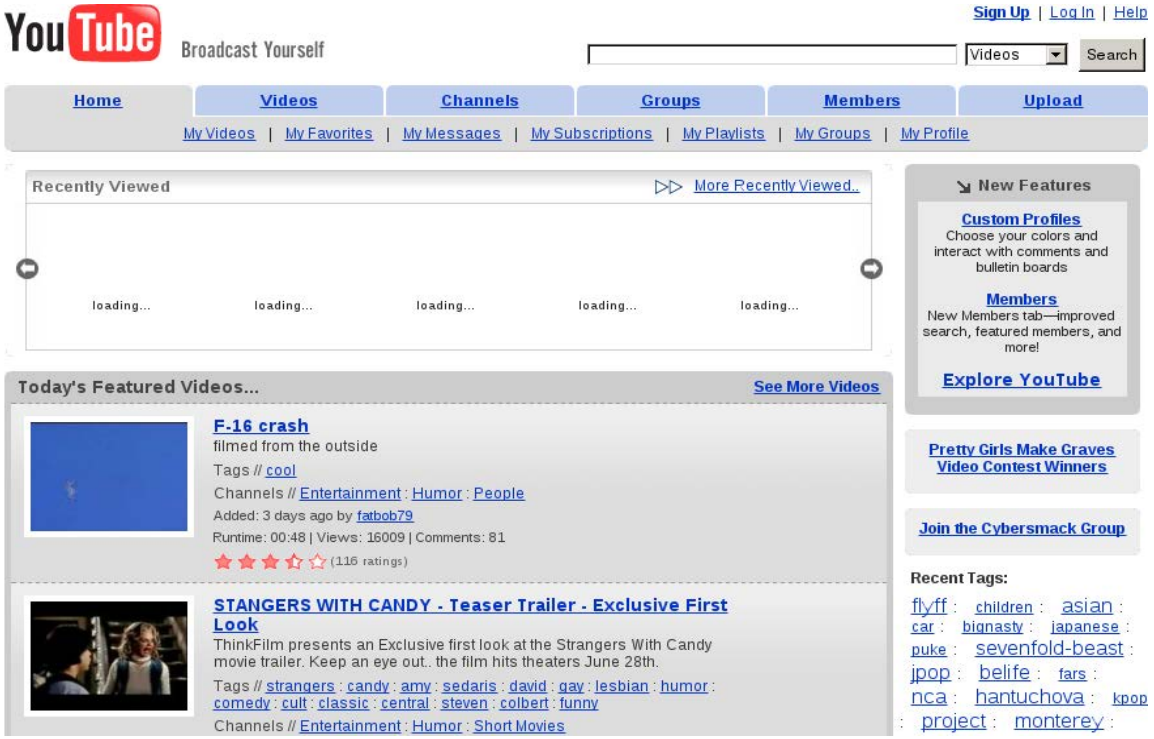


2006 YouTube.com from Internet Archive

http://web.archive.org/web/20060427213420/http://youtube.com

YouTube as captured without JavaScript capability

- Titles fetched with JavaScript
- Primary content (preview) still missing



2006 YouTube.com from Internet Archive

http://web.archive.org/web/20060427213420/http://youtube.com

Dependent Loading of Resources

GET
http://web.archive.org/web/20121208145112cs_/http://s.ytimg.com/yt/cssbin/www-core-vfl_OJqFG.css 404 (Not Found) www.youtube.com:15

GET
http://web.archive.org/web/20121208145115js_/http://s.ytimg.com/yt/jsbin/www-core-vfl8PDcRe.js 404 (Not Found) www.youtube.com:45

Uncaught TypeError: Object #<Object> has no method 'setConfig' www.youtube.com:56

Uncaught TypeError: Cannot read property 'home' of undefined www.youtube.com:76

Uncaught TypeError: Cannot read property 'ajax' of undefined www.youtube.com:86

Uncaught TypeError: Object #<Object> has no method 'setConfig' www.youtube.com:101

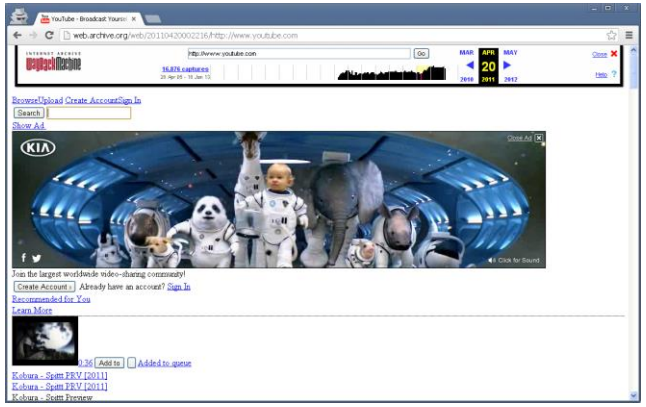
Uncaught ReferenceError: _gel is not defined www.youtube.com:1784

Uncaught TypeError: Object #<Object> has no method 'setConfig' www.youtube.com:1929

Uncaught TypeError: Cannot read property 'home' of undefined www.youtube.com:524

GET
http://web.archive.org/web/20130101024721im_/http://i2.ytimg.com/vi/1f7neSzDqvc/default.jpg 404 (Not Found)

YouTube 2011



http://web.archive.org/web/20110420002216/http://www.youtube.com/

Dependent Loading of Resources

GET http://web.archive.org/web/20121208145112cs_/http://s.ytimg.com/yt/cssbin/www-core-vfl_OJqFG.css 404 (Not Found) www.youtube.com:15

GET http://web.archive.org/web/20121208145115js_/http://s.ytimg.com/yt/jsbin/www-core-vfl8PDcRe.js 404 (Not Found) www.youtube.com:45

Uncaught TypeError: Object #<Object> has no method 'setConfig' www.youtube.com:56

Uncaught TypeError: Cannot read property 'home' of undefined www.youtube.com:76

Uncaught TypeError: Cannot read property 'ajax' of undefined www.youtube.com:86

Uncaught TypeError: Object #<Object> has no method 'setConfig' www.youtube.com:101

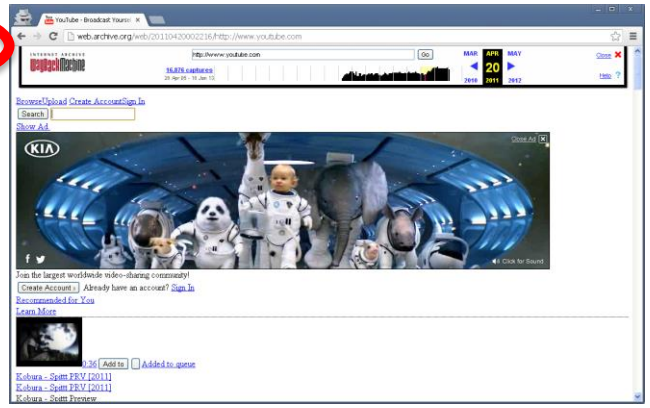
Uncaught ReferenceError: _gel is not defined www.youtube.com:1784

Uncaught TypeError: Object #<Object> has no method 'setConfig' www.youtube.com:1929

Uncaught TypeError: Cannot read property 'home' of undefined www.youtube.com:524

GET http://web.archive.org/web/20130101024721im_/http://i2.ytimg.com/vi/1f7neSzDqvc/default.jpg 404 (Not Found)

YouTube 2011



http://web.archive.org/web/20110420002216/http://www.youtube.com/

Missing CSS

Missing JS

Missing JS-dependent resources

Dependent Loading of Resources

GET
http://web.archive.org/web/20121208145112cs_/http://s.ytimg.com/yt/cssbin/www-core-vfl_OJqFG.css 404 (Not Found) www.youtube.com:15

GET
http://web.archive.org/web/20121208145115js_/http://s.ytimg.com/yt/jsbin/www-core-vfl8PDcRe.js 404 (Not Found) www.youtube.com:45

Uncaught TypeError: Object #<Object> has no method 'setConfig' www.youtube.com:56

Uncaught TypeError: Cannot read property 'home' of undefined www.youtube.com:76

Uncaught TypeError: Cannot read property 'ajax' of undefined www.youtube.com:86

Uncaught TypeError: Object #<Object> has no method 'setConfig' www.youtube.com:101

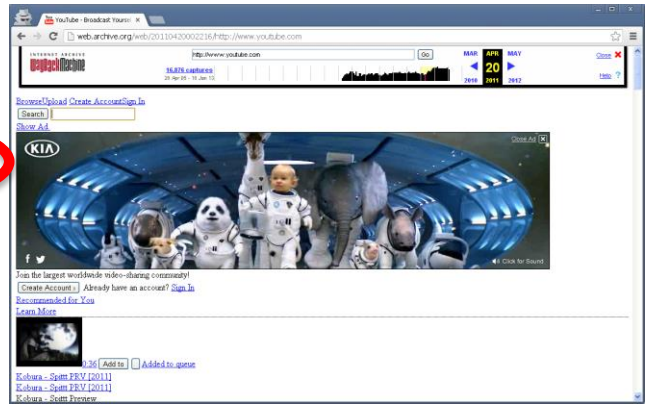
Uncaught ReferenceError: _gel is not defined www.youtube.com:1784

Uncaught TypeError: Object #<Object> has no method 'setConfig' www.youtube.com:1929

Uncaught TypeError: Cannot read property 'home' of undefined www.youtube.com:524

GET
http://web.archive.org/web/20130101024721im_/http://i2.ytimg.com/vi/1f7neSzDqvc/default.jpg 404 (Not Found)

YouTube 2011



http://web.archive.org/web/20110420002216/http://www.youtube.com/

Missing CSS

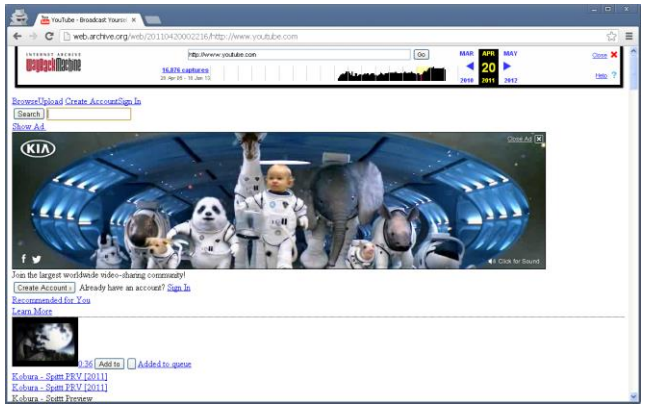
Missing JS

Missing JS-dependent resources

Dependent Loading of Resources

GET
http://web.archive.org/web/20121208145112cs_/http://s.ytimg.com/yt/cssbin/www-core-vfl_OJqFG.css 404 (Not Found) www.youtube.com:15
GET
http://web.archive.org/web/20121208145115js_/http://s.ytimg.com/yt/jsbin/www-core-vfl8PDcRe.js 404 (Not Found) www.youtube.com:45
Uncaught TypeError: Object #<Object> has no method 'setConfig' www.youtube.com:56
Uncaught TypeError: Cannot read property 'home' of undefined www.youtube.com:76
Uncaught TypeError: Cannot read property 'ajax' of undefined www.youtube.com:86
Uncaught TypeError: Object #<Object> has no method 'setConfig' www.youtube.com:101
Uncaught ReferenceError: _gel is not defined www.youtube.com:1784
Uncaught TypeError: Object #<Object> has no method 'setConfig' www.youtube.com:1929
Uncaught TypeError: Cannot read property 'home' of undefined www.youtube.com:524
GET
http://web.archive.org/web/20130101024721im_/http://i2.ytimg.com/vi/1f7neSzBqvc/default.jpg 404 (Not Found)

YouTube 2011

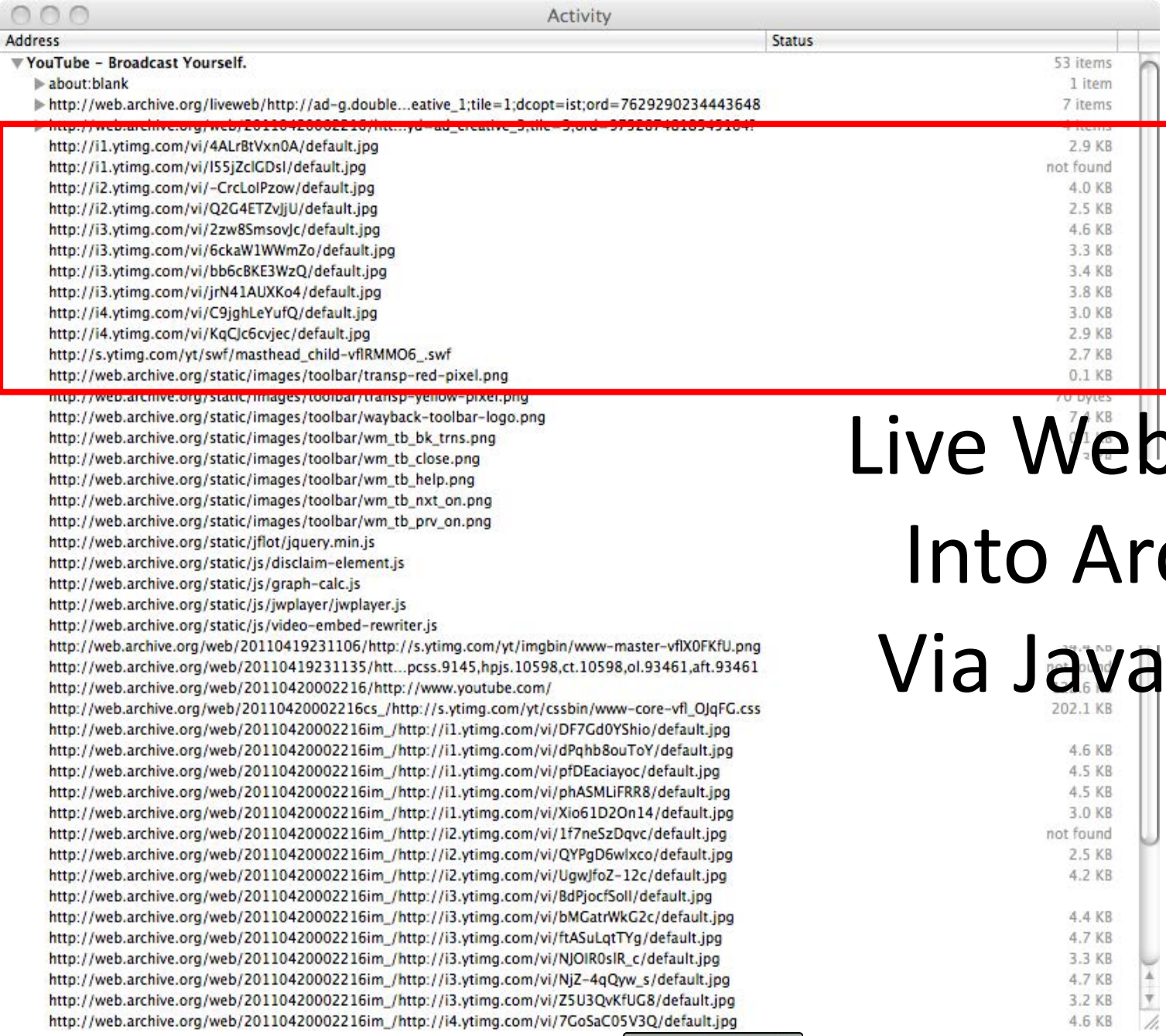


http://web.archive.org/web/20110420002216/http://www.youtube.com/

Missing CSS

Missing JS

Missing JS-dependent resources

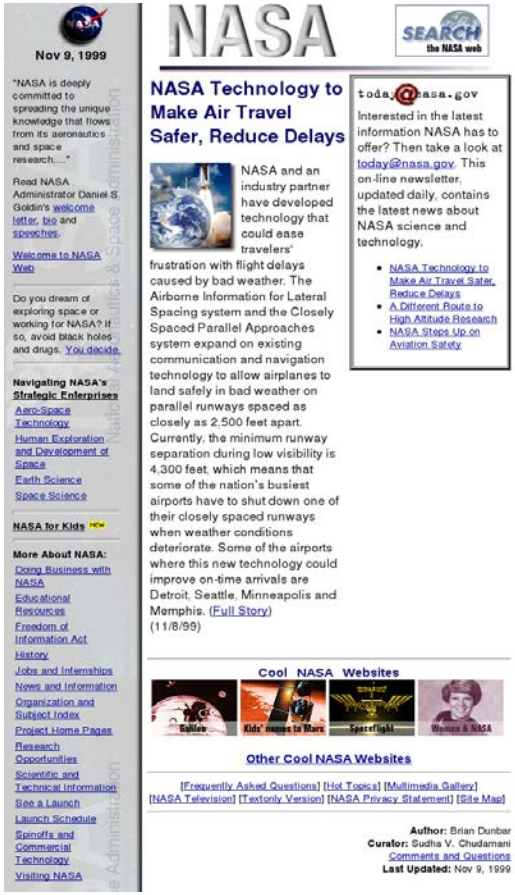


Live Web Leaks Into Archive Via Javascript

see: http://ws-dl.blogspot.com/2012/10/2012-10-10-zombies-in-archives.html

Stepping Back, NASA 1999

- Few embedded resources
- Little to no JavaScript
  - If some, not Ajax
- No resources dependent on JavaScript

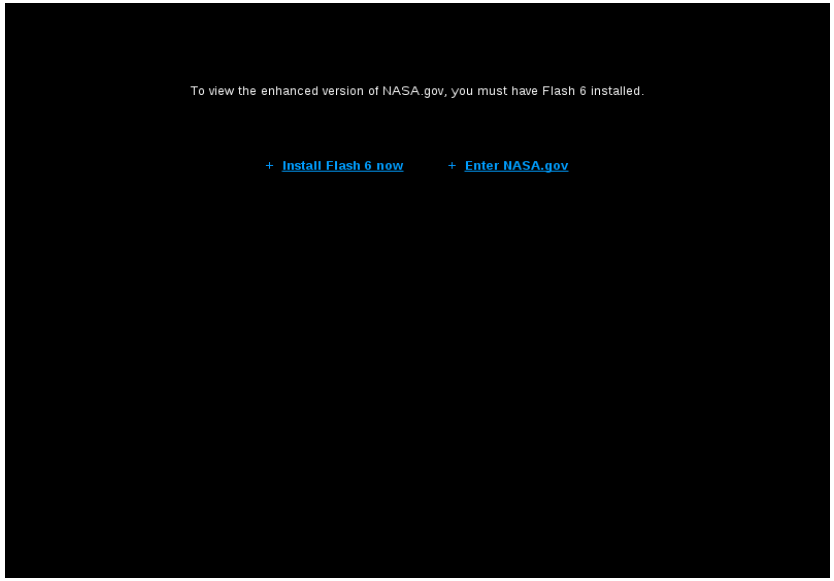


http://web.archive.org/web/19991109105105/http://www.nasa.gov

Stepping Back, NASA 2003

- Content requires JS
- Accessibility goes awry
- Archivability decreases

http://web.archive.org/web/20041014205942/http://www.nasa.gov

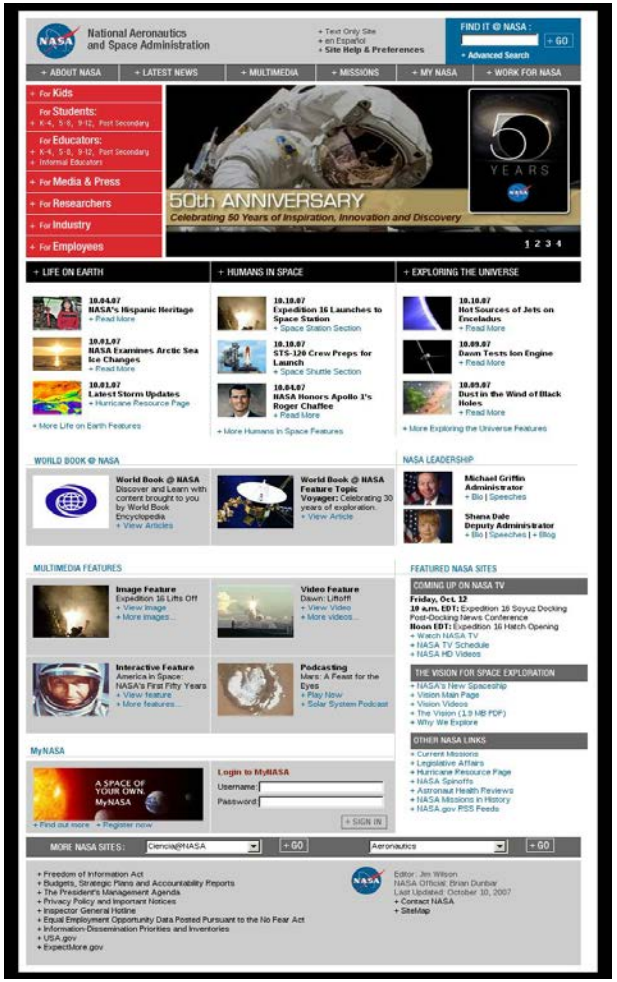



```
var fstr = '';
if(hasFlash(6)) {
        fstr+='<object id="screenreader.swf">...</object>';
        window.status = 'Flash 6 Detected...';
} else {
        fstr+='...To view the enhanced version of NASA.gov, you must
have Flash 6 installed....';
}
with(document) { open('text/html'); write(fstr); close(); }
```

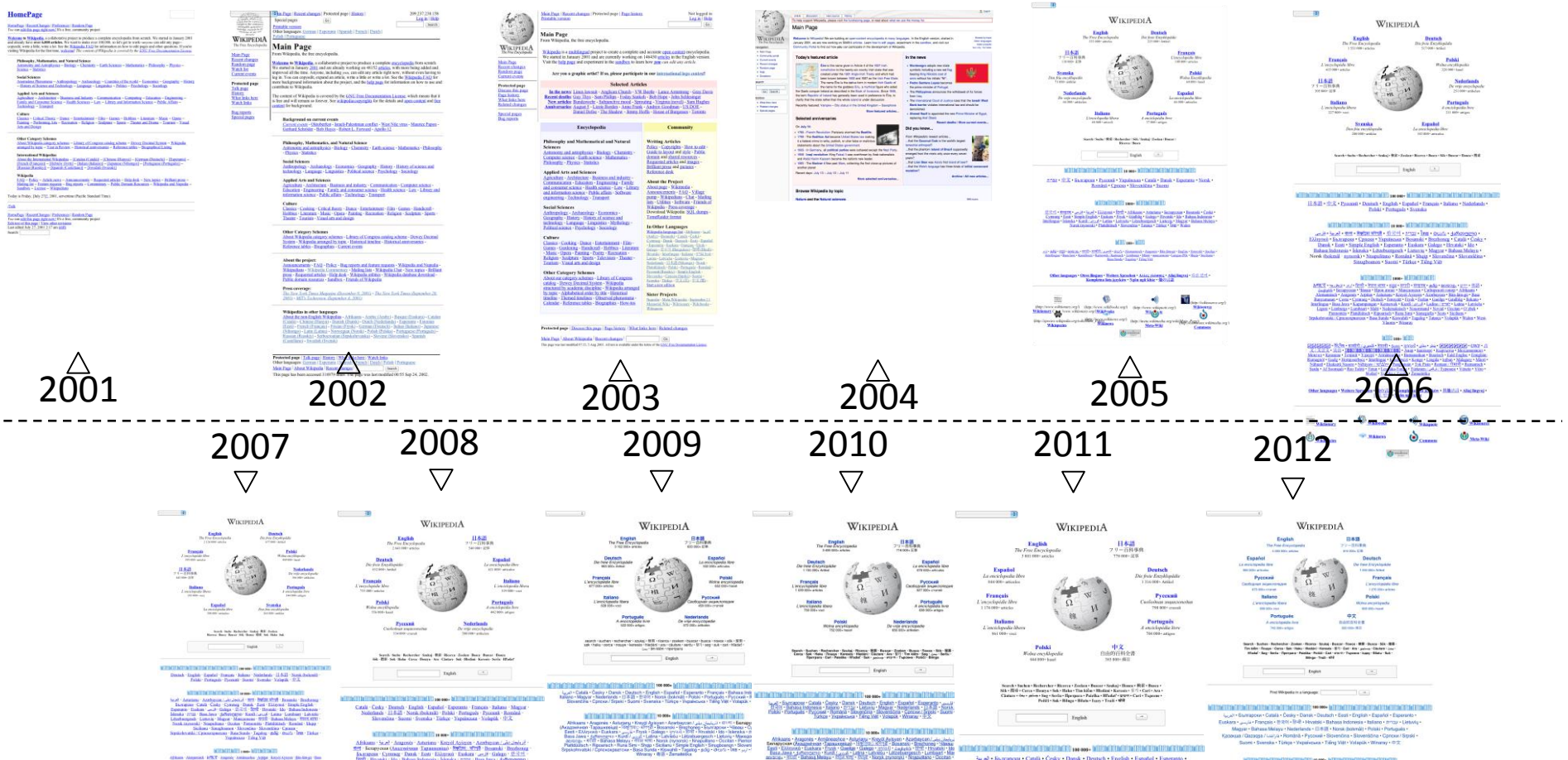
Stepping Back, NASA 2007

http://web.archive.org/web/20071011055607/http://www.nasa.gov

- JS check removed
- Content is Accessible
- Archivability Goes Up
- Accessibility→Archivability



Reinforcing Case: Wikipedia



Summary

- Javascript: good for interaction, bad for archiving
  - crawlers miss URIs at crawl time
  - rendering yesterday's pages causes them to reach into today's web
- Different trends of archivability over time:
  - YouTube: bad-->worse
  - NASA: good-->bad-->good
  - Wikipedia: good
  - see all: http://www.cs.odu.edu/~mkelly/semester/2013_spring/20130127alexatop10/
- Overall, archivability is getting worse
  - 24% increase in missing embedded resources from 2006-2010 due to Javascript