

9-25-2013

Who and What Links to the Internet Archive

Yasmin AlNoamany
Old Dominion University

Ahmed Alsum
Old Dominion University

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

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

Follow this and additional works at: https://digitalcommons.odu.edu/computerscience_presentations



Part of the [Archival Science Commons](#)

Recommended Citation

AlNoamany, Yasmin; Alsum, Ahmed; Weigle, Michele C.; and Nelson, Michael L., "Who and What Links to the Internet Archive" (2013). *Computer Science Presentations*. 11.
https://digitalcommons.odu.edu/computerscience_presentations/11

This Book is brought to you for free and open access by the Computer Science at ODU Digital Commons. It has been accepted for inclusion in Computer Science Presentations by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.

Who and What Links to the Internet Archive

Yasmin AlNoamany, Ahmed AlSum, Michele C. Weigle, Michael L. Nelson

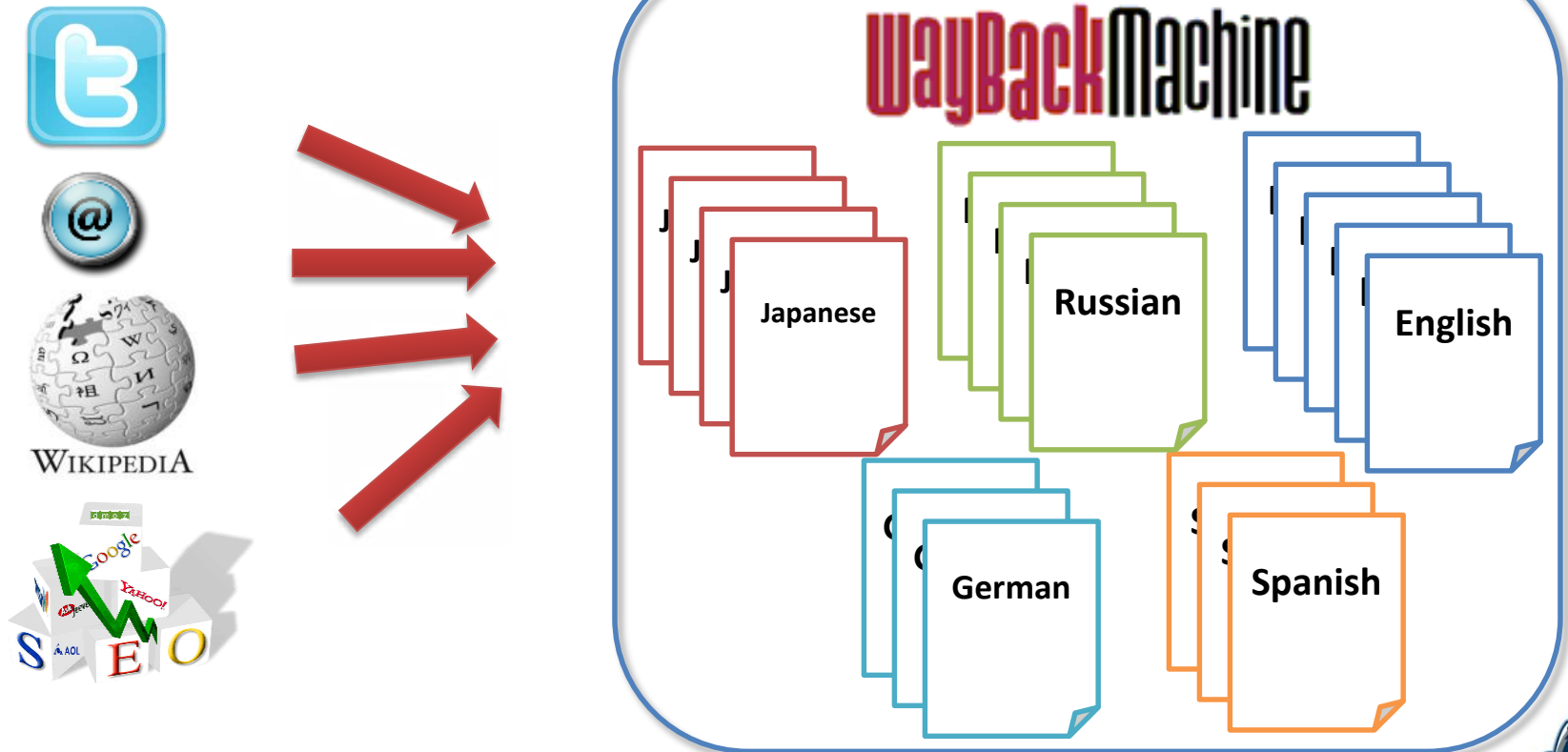
Computer Science Department

Old Dominion University, Norfolk, VA

mln@cs.odu.edu

Motivation

- What do web archive users look for and where do they come from?



Methodology

Data Set

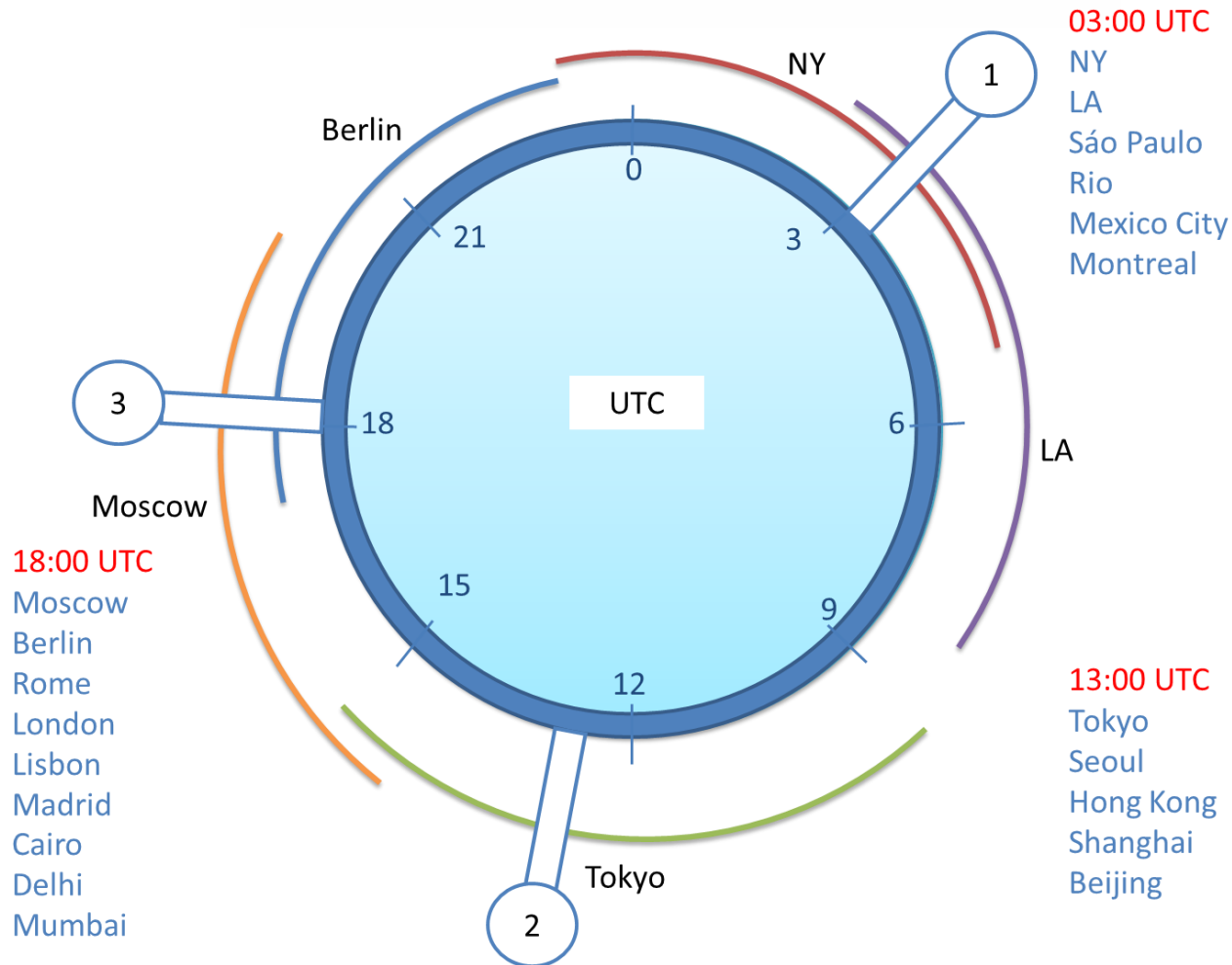
- **Six million** records from Internet Archive's Wayback Machine web server logs of **February 2, 2012**
- Data set statistics

Get	Embedded Resources	Null Referrers	2xx	3xx	4xx	5xx	Humans	Robots
99%	43%	47%	33%	51%	12%	4%	1.5%	18.8%



The percentage of humans and robots remaining after cleaning

Sample from 6pm-midnight (prime Internet hours)



Wayback Machine Access Logs

```
0.247.222.86 - - [02/Feb/2012:07:03:46 +0000] "GET
http://wayback.archive.org/web/*/http://www.cnn.com
HTTP/1.1" 200 96433 "http://www.archive.org/web/web.php"
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8)
AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77
Safari/535.7"
```

- Client IP: 0.247.222.86
- Access time: 02/Feb/2012:07:03:46 +0000
- HTTP request method: GET
- URI: http://wayback.archive.org/web/*/http://www.cnn.com
- Protocol: HTTP/1.1
- HTTP status code: 200
- Bytes sent: 96433
- Referring URI: http://www.archive.org/web/web.php
- User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7

Wayback Machine Access Logs

```
0.247.222.86 - - [02/Feb/2012:07:03:46 +0000] "GET  
http://wayback.archive.org/web/*/http://www.cnn.com  
HTTP/1.1" 200 96433 "http://www.archive.org/web/web.php"  
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like  
AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"
```

IPs anonymized by Internet Archive

- Client IP: 0.247.222.86
- Access time: 02/Feb/2012:07:03:46 +0000
- HTTP request method: GET
- URI: http://wayback.archive.org/web/*/http://www.cnn.com
- Protocol: HTTP/1.1
- HTTP status code: 200
- Bytes sent: 96433
- Referring URI: http://www.archive.org/web/web.php
- User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7

Pre-Processing

- Data Cleaning
- Session Identification
- Robot Detection AlNoamany 2013

Data Cleaning

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20070519015308/http
://www.jcdl.org/ HTTP/1.1" 200 2137 "-"
"Mozilla/5.0"
```

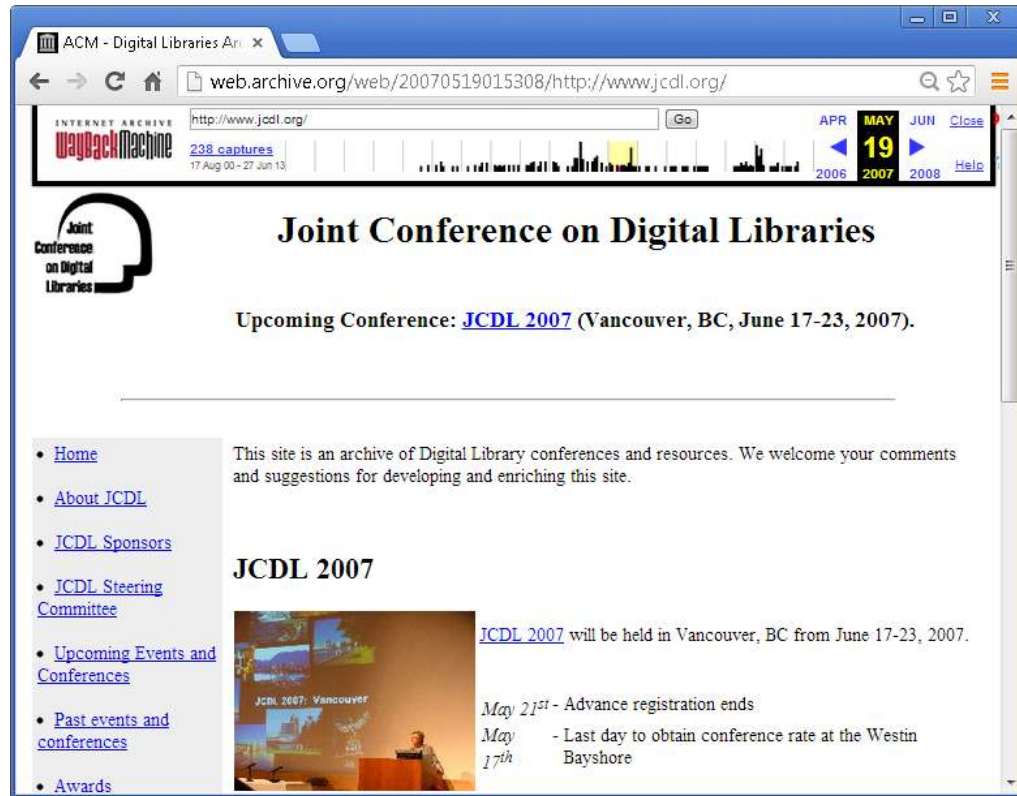
```
0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20070519015308im /h
ttp://www.jcdl.org/images/jcdl2007-edie.jpg
HTTP/1.1" 200 2137 "-" "Mozilla/5.0"
```

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://staticweb.archive.org/images/toolbar/wa
yback-toolbar-logo.png HTTP/1.1" 200 3700 "-"
"Mozilla/5.0"
```

```
0.151.147.108 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20100102003557/abou
t:blank HTTP/1.1" 302 0 "www.xx.com"
"Mozilla/4.0"
```

```
0.26.129.146 - - [02/Feb/2012:00:01:54] "GET
http://web.archive.org/web/20140004100000/http
://www.jcdl.org/ HTTP/1.1" 302 0 "-"
"Mozilla/5.0"
```

<http://web.archive.org/web/20070519015308/http://www.jcdl.org/>



ACM - Digital Libraries An x

web.archive.org/web/20070519015308/http://www.jcdl.org/

INTERNET ARCHIVE
Wayback Machine
238 captures
17 Aug 00 - 27 Jun 13

APR MAY JUN Close
2006 19 2007 2008 Help

Joint Conference on Digital Libraries

Upcoming Conference: [JCDL 2007](#) (Vancouver, BC, June 17-23, 2007).

This site is an archive of Digital Library conferences and resources. We welcome your comments and suggestions for developing and enriching this site.

- [Home](#)
- [About JCDL](#)
- [JCDL Sponsors](#)
- [JCDL Steering Committee](#)
- [Upcoming Events and Conferences](#)
- [Past events and conferences](#)
- [Awards](#)

JCDL 2007

[JCDL 2007](#) will be held in Vancouver, BC from June 17-23, 2007.

May 21st - Advance registration ends
May 17th - Last day to obtain conference rate at the Westin Bayshore

Embedded Resources

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20070519015308/http
://www.jcdl.org/ HTTP/1.1" 200 2137 "-"
"Mozilla/5.0"
```

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20070519015308im /h
ttp://www.jcdl.org/images/jcdl2007-edie.jpg
HTTP/1.1" 200 2137 "-" "Mozilla/5.0"
```

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://staticweb.archive.org/images/toolbar/wa
yback-toolbar-logo.png HTTP/1.1" 200 3700 "-"
"Mozilla/5.0"
```

```
0.151.147.108 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20100102003557/abou
t:blank HTTP/1.1" 302 0 "www.xx.com"
"Mozilla/4.0"
```

```
0.26.129.146 - - [02/Feb/2012:00:01:54] "GET
http://web.archive.org/web/20140004100000/http
://www.jcdl.org/ HTTP/1.1" 302 0 "-"
"Mozilla/5.0"
```

<http://web.archive.org/web/20070519015308/http://www.jcdl.org/>

ACM - Digital Libraries An x

web.archive.org/web/20070519015308/http://www.jcdl.org/

INTERNET ARCHIVE Wayback Machine 238 captures 17 Aug 00 - 27 Jun 13

APR MAY JUN Close 19 2006 2007 2008 Help

Joint Conference on Digital Libraries

Upcoming Conference: **JCDL 2007** (Vancouver, BC, June 17-23, 2007).

This site is an archive of Digital Library conferences and resources. We welcome your comments and suggestions for developing and enriching this site.

- Home
- About JCDL
- JCDL Sponsors
- JCDL Steering Committee
- Upcoming Events and Conferences
- Past events and conferences
- Awards

JCDL 2007

JCDL 2007 will be held in Vancouver, BC from June 17-23, 2007.

May 21st - Advance registration ends
May 17th - Last day to obtain conference rate at the Westin Bayshore

Embedded Resources

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20070519015308/http
://www.jcdl.org/ HTTP/1.1" 200 2137 "-"
"Mozilla/5.0"
```

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20070519015308im /h
ttp://www.jcdl.org/wayback-machine/2007-edie.jpg
HTTP/1.1" 200 2137 "-" "Mozilla/5.0"
```

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://staticweb.archive.org/images/toolbar/wa
yback-toolbar-logo.png HTTP/1.1" 200 3700 "-"
"Mozilla/5.0"
```

```
0.151.147.108 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20100102003557/about:blank
HTTP/1.1" 302 0 "www.xx.com"
"Mozilla/4.0"
```

```
0.26.129.146 - - [02/Feb/2012:00:01:54] "GET
http://web.archive.org/web/20140004100000/http
://www.jcdl.org/ HTTP/1.1" 302 0 "-"
"Mozilla/5.0"
```

<http://web.archive.org/web/20070519015308/http://www.jcdl.org/>

ACM - Digital Libraries An x

web.archive.org/web/20070519015308/http://www.jcdl.org/

INTERNET ARCHIVE Wayback Machine 238 captures 17 Aug 00 - 27 Jun 13

APR MAY JUN Close 19 2006 2007 2008 Help

Joint Conference on Digital Libraries

Upcoming Conference: [JCDL 2007](#) (Vancouver, BC, June 17-23, 2007).

This site is an archive of Digital Library conferences and resources. We welcome your comments and suggestions for developing and enriching this site.

- [Home](#)
- [About JCDL](#)
- [JCDL Sponsors](#)
- [JCDL Steering Committee](#)
- [Upcoming Events and Conferences](#)
- [Past events and conferences](#)
- [Awards](#)

JCDL 2007

[JCDL 2007](#) will be held in Vancouver, BC from June 17-23, 2007.

May 21st - Advance registration ends
May 17th - Last day to obtain conference rate at the Westin Bayshore

Static Resources

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20070519015308/http
://www.jcdl.org/ HTTP/1.1" 200 2137 "-"
"Mozilla/5.0"
```

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20070519015308im /h
ttp://www.jcdl.org/wayback-machine/2007-edie.jpg
HTTP/1.1" 200 2137 "-" "Mozilla/5.0"
```

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://staticweb.archive.org/images/toolbar/wa
yback-toolbar-logo.png HTTP/1.1" 200 3700 "-"
"Mozilla/5.0"
```

```
0.151.147.108 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20100102003557/abou
t:blank HTTP/1.1" 302 0 "www.xx.com"
"Mozilla/4.0"
```

```
0.26.129.146 - - [02/Feb/2012:00:01:54] "GET
http://web.archive.org/web/20140004100000/http
://www.jcdl.org/ HTTP/1.1" 302 0 "-"
"Mozilla/5.0"
```

<http://web.archive.org/web/20070519015308/http://www.jcdl.org/>

ACM - Digital Libraries An x

web.archive.org/web/20070519015308/http://www.jcdl.org/

INTERNET ARCHIVE
Wayback Machine

238 captures
17 Aug 00 - 27 Jun 13

APR MAY JUN Close
2006 19 2007 2008 Help

Joint Conference on Digital Libraries

Upcoming Conference: [JCDL 2007](#) (Vancouver, BC, June 17-23, 2007).

This site is an archive of Digital Library conferences and resources. We welcome your comments and suggestions for developing and enriching this site.

- [Home](#)
- [About JCDL](#)
- [JCDL Sponsors](#)
- [JCDL Steering Committee](#)
- [Upcoming Events and Conferences](#)
- [Past events and conferences](#)
- [Awards](#)

JCDL 2007

[JCDL 2007](#) will be held in Vancouver, BC from June 17-23, 2007.

May 21st - Advance registration ends
May 17th - Last day to obtain conference rate at the Westin Bayshore

Invalid Requests

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20070519015308/http
://www.jcdl.org/ HTTP/1.1" 200 2137 "-"
"Mozilla/5.0"
```

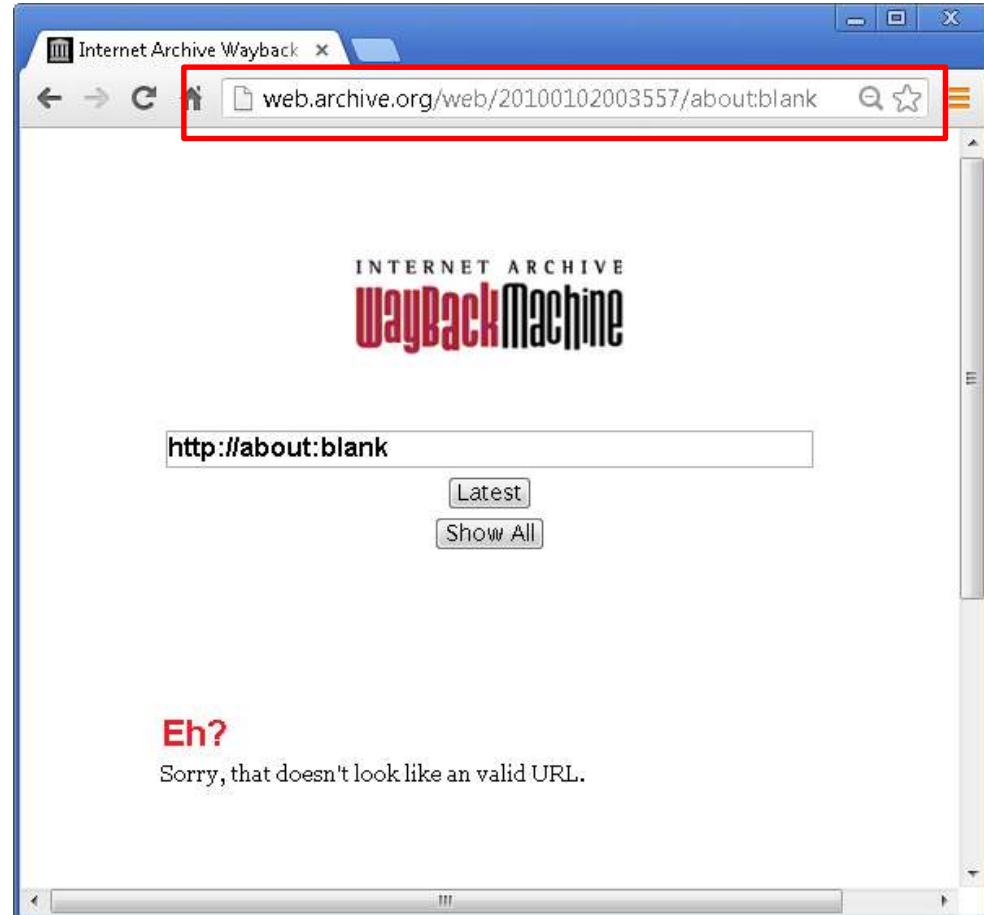
```
0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20070519015308im /h
http://www.jcdl.org/day/2007-edie.jpg
HTTP/1.1" 200 2137 "-" "Mozilla/5.0"
```

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://staticweb.archive.org/images/toolbar/wa
yback-toolbar-logged.png HTTP/1.1" 200 3700 "-"
"Mozilla/5.0"
```

```
0.151.147.108 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20100102003557/abou
t:blank HTTP/1.1" 302 0 "www.xx.com"
"Mozilla/4.0"
```

```
0.26.129.146 - - [02/Feb/2012:00:01:54] "GET
http://web.archive.org/web/20140004100000/http
://www.jcdl.org/ HTTP/1.1" 302 0 "-"
"Mozilla/5.0"
```

<http://web.archive.org/web/20100102003557/about:blank>



Invalid Requests

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET  
http://web.archive.org/web/20070519015308/http://www.jcdl.org/ HTTP/1.1" 200 2137 "-"  
"Mozilla/5.0"
```

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET  
http://web.archive.org/web/20070519015308im/http://www.jcdl.org/wayback-machine/2007-edie.jpg  
HTTP/1.1" 200 2137 "-" "Mozilla/5.0"
```

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET  
http://staticweb.archive.org/images/toolbar/wayback-toolbar-logged-in.png  
HTTP/1.1" 200 3700 "-"  
"Mozilla/5.0"
```

```
0.151.147.108 [02/Feb/2012:00:01:03] "GET  
http://web.archive.org/web/20100102003557/about:blank HTTP/1.1" 200 0 "http://www.xx.com"  
"Mozilla/4.0"
```

```
0.26.129.146 - - [02/Feb/2012:00:01:54] "GET  
http://web.archive.org/web/20140004100000/http://www.jcdl.org/ HTTP/1.1" 302 0 "-"  
"Mozilla/5.0"
```

```
http://web.archive.org/web/20100102003557/about:blank
```

The screenshot shows a browser window titled "Internet Archive Wayback" with the address bar containing web.archive.org/web/20100102003557/about:blank. The page content displays the "INTERNET ARCHIVE WayBack Machine" logo and a search bar with the text "http://about:blank". Below the search bar are buttons for "Latest" and "Show All". At the bottom, a red error message reads "Eh? Sorry, that doesn't look like a valid URL." A red box highlights the address bar.

Requests that had 3xx Status Code

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET  
http://web.archive.org/web/20070519015308/http://www.jcdl.org/ HTTP/1.1" 200 2137 "-"  
"Mozilla/5.0"
```

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET  
http://web.archive.org/web/20070519015308im/http://www.jcdl.org/ HTTP/1.1" 200 2137 "-"  
"Mozilla/5.0"
```

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET  
http://staticweb.archive.org/images/toolbar/wayback-toolbar-logged.png HTTP/1.1" 200 3700 "-"  
"Mozilla/5.0"
```

```
0.151.147.108 [02/Feb/2012:00:01:03] "GET  
http://web.archive.org/web/20100102003557/about:blank HTTP/1.1" 200 0 "http://www.xx.com"  
"Mozilla/4.0"
```

```
0.26.129.146 - - [02/Feb/2012:00:01:54] "GET  
http://web.archive.org/web/20140004100000/http://www.jcdl.org/ HTTP/1.1" 302 0 "-"  
"Mozilla/5.0"
```

<http://web.archive.org/web/20130114160045/http://www.jcdl.org/>

Joint Conference on Digital Libraries

Upcoming Conference: JCDL 2013: July 22-26, Indianapolis, IN

Home About Sponsors Steering Committee Upcoming Events Past Conferences

Welcome to JCDL

The Joint Conference on Digital Libraries (JCDL) is a major international forum focusing on digital libraries and associated technical, practical, and social issues. JCDL enhances the tradition of conference excellence already established by the ACM and IEEE-CS by combining the annual events that these professional societies have sponsored on an annual basis, the ACM Digital Libraries Conferences and the IEEE-CS Advances in Digital Libraries Conferences.

Upcoming Events

- RCDL 2013: 15th All-Russia Conference, Yaroslavl, RU - 17, 2013
- JCDL 2013: 13th ACM/IEEE Conference on Digital Libraries, Indianapolis, IN, July 22-26, 2013

Requests that had 3xx Status Code

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET  
http://web.archive.org/web/20070519015308/http://www.jcdl.org/ HTTP/1.1" 200 2137 "-"  
"Mozilla/5.0"
```

```
http://web.archive.org/web/20130114160045/  
http://www.jcdl.org/
```

```
curl -I "http://web.archive.org/web/20140004100000/http://www.jcdl.org/"
```

```
HTTP/1.1 302 Moved Temporarily
```

```
Server: Tengine/1.4.3
```

```
Date: Tue, 02 Jul 2013 19:48:59 GMT
```

```
Content-Type: application/octet-stream
```

```
Content-Length: 0
```

```
Connection: keep-alive
```

```
set-cookie: wayback_server=10; Domain=archive.org; Path=/; Expires=Thu, 01-Aug-13 19:48:59 GMT;
```

```
Location: /web/20130114160045/http://www.jcdl.org/
```

```
0.26.129.146 - - [02/Feb/2012:00:01:54] "GET  
http://web.archive.org/web/20140004100000/http://www.jcdl.org/ HTTP/1.1" 302 0 "-"  
"Mozilla/5.0"
```

Welcome to JCDL

The Joint Conference on Digital Libraries (JCDL) is a major international forum focusing on digital libraries and associated technical, practical, and social issues. JCDL enhances the tradition of conference excellence already established by the ACM and IEEE-CS by combining the annual events that these professional societies have sponsored on an annual basis, the ACM Digital Libraries Conferences and the IEEE-CS Advances in Digital Libraries Conferences.

Upcoming Events

- [RCDL 2013: 15th All-Rus Conference, Yaroslavl, Ru - 17, 2013](#)
- [JCDL 2013: 13th ACM/IEEE Conference on Digital Lib Indianapolis, IN, July 22-2](#)
- [JCDL 2013: 13th ACM/IEEE Conference on Digital Lib Indianapolis, IN, July 22-2](#)

Requests that had 3xx Status Code

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET  
http://web.archive.org/web/20070519015308/http://www.jcdl.org/ HTTP/1.1" 200 2137 "-"  
"Mozilla/5.0"
```

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET  
http://web.archive.org/web/20070519015308im/http://www.jcdl.org/ HTTP/1.1" 200 2137 "-"  
"Mozilla/5.0"
```

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET  
http://staticweb.archive.org/images/toolbar/wayback-toolbar-logged.png HTTP/1.1" 200 3700 "-"  
"Mozilla/5.0"
```

```
0.151.147.108 [02/Feb/2012:00:01:03] "GET  
http://web.archive.org/web/20100102003557/about:blank HTTP/1.1" 200 0 "http://www.xx.com"  
"Mozilla/4.0"
```

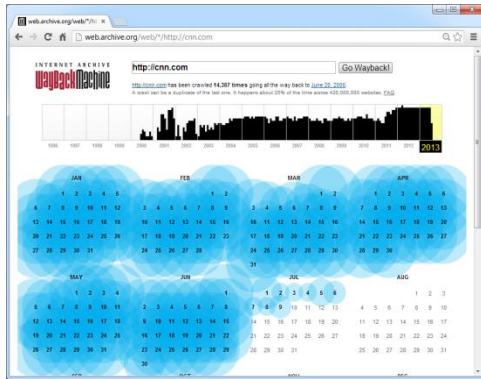
```
0.26.129.146 [02/Feb/2012:00:01:54] "GET  
http://web.archive.org/web/20140004100000/http://www.jcdl.org/ HTTP/1.1" 302 0 "-"  
"Mozilla/5.0"
```

<http://web.archive.org/web/20130114160045/http://www.jcdl.org/>

The screenshot shows a browser window with the URL <http://web.archive.org/web/20130114160045/http://www.jcdl.org/> in the address bar. The page content includes the Wayback Machine logo, the conference title "Joint Conference on Digital Libraries", and a list of upcoming events. The events listed are:

- RCDL 2013: 15th All-Russia Conference, Yaroslavl, RU - 17, 2013
- JCDL 2013: 13th ACM/IEEE Conference on Digital Libraries, Indianapolis, IN, July 22-26, 2013

Session: Set of Web Pages Requested by a Particular User



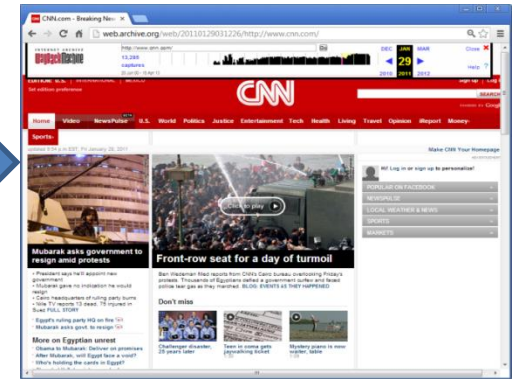
p1

1 mins



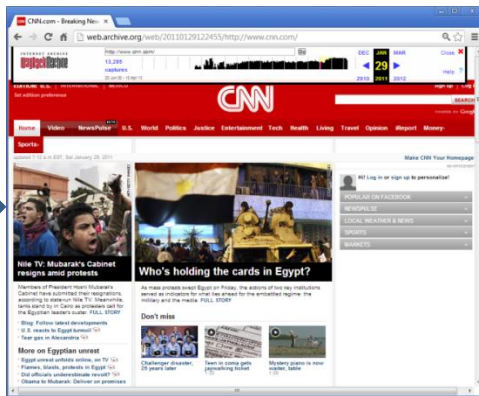
p2

4 mins



p3

3 mins



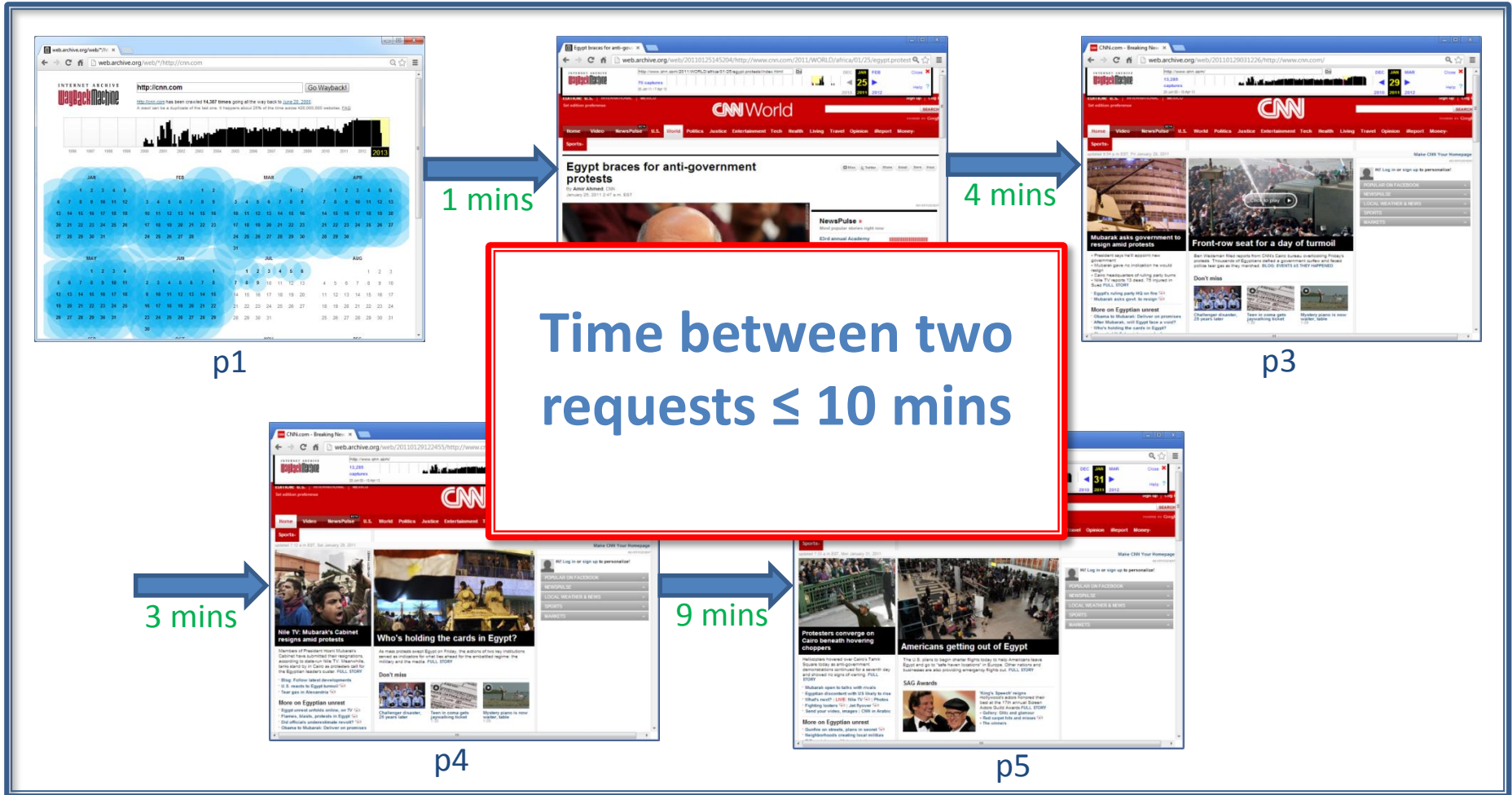
p4

9 mins



p5

Session: Set of Web Pages Requested by a Particular User

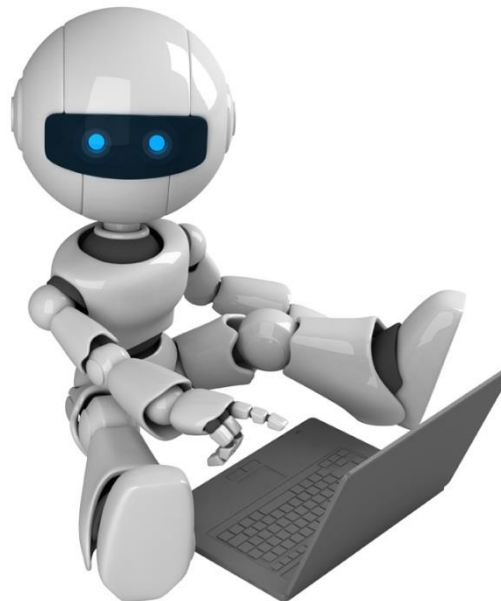


Session Identification

- Threshold timeout: 10 minutes Liu et al. 2007, Spiliopoulou et al. 2003
- Grouping: based on the IP and User-Agent

Robot Detection is a Big Challenge

I'm not a
robot

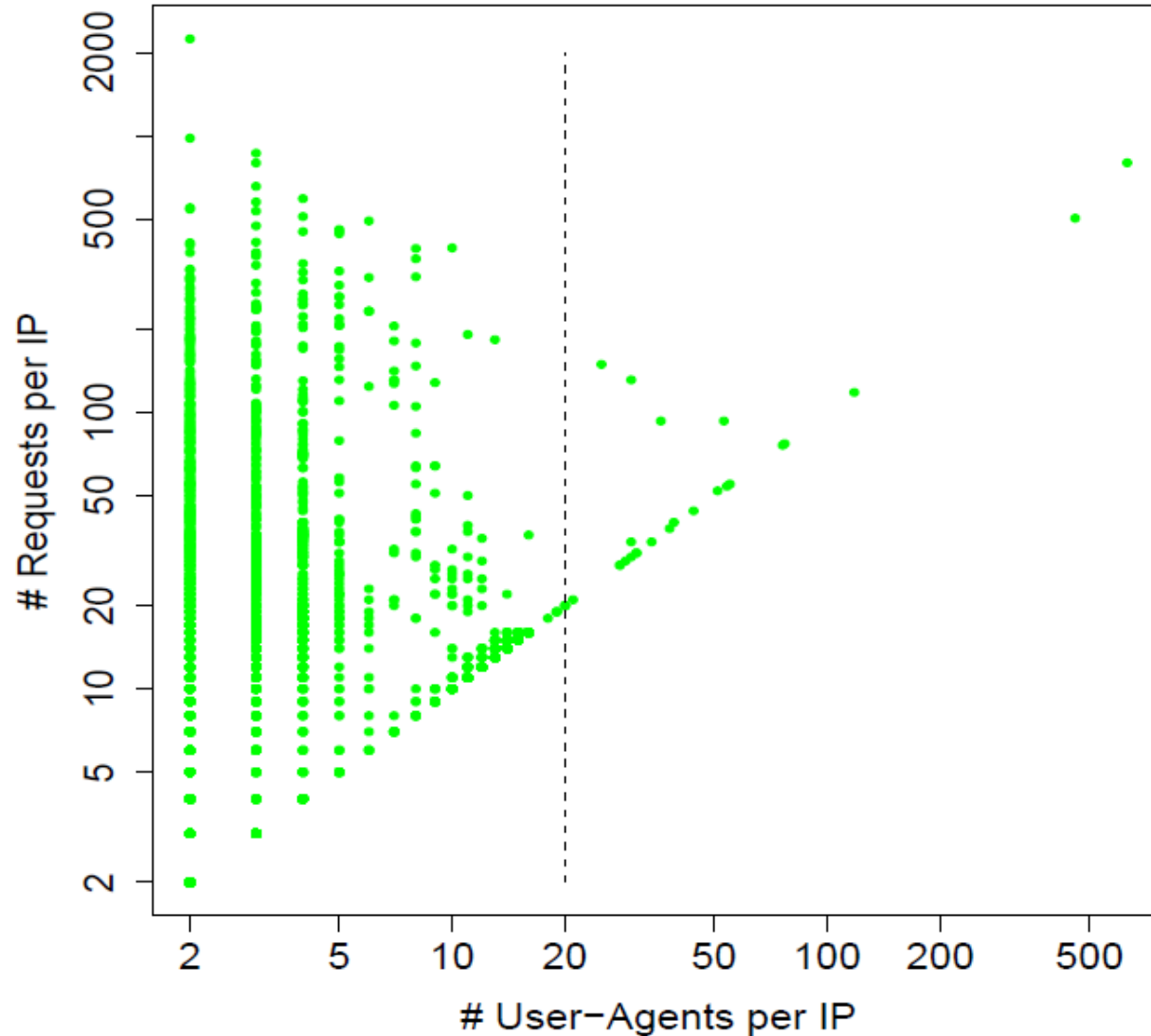


User-Agent Check

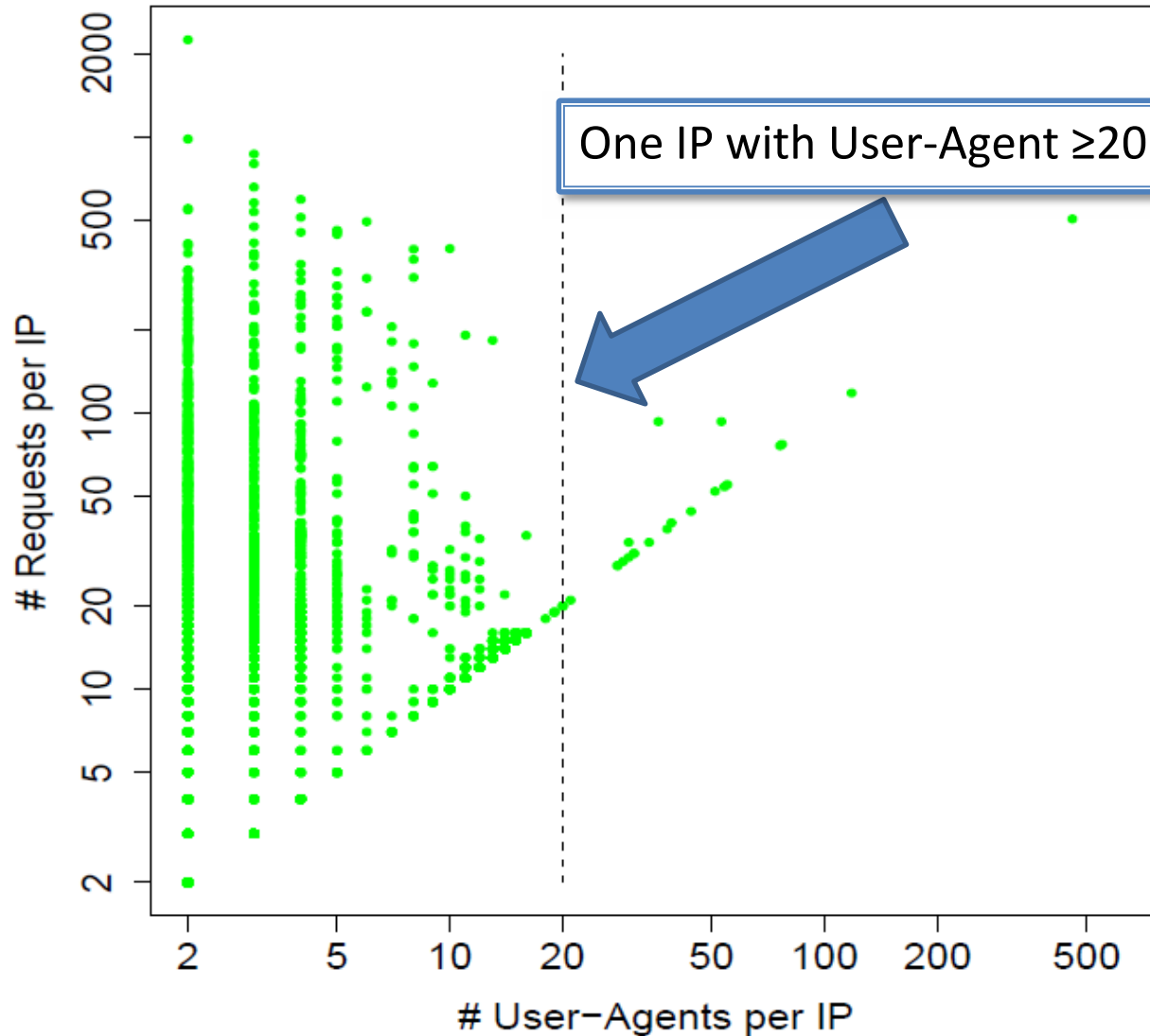
```
0.182.141.149 - -  
[02/Feb/2012:00:01:51 +0000] "GET  
http://wayback.archive.org/web/199906  
01000000*/http://www.belizefirst.com/  
HTTP/1.0" 200 98507 "-"
```

```
"Python-urllib/1.17"
```


Number of User-Agents per IP



Number of User-Agents per IP



Robots.txt File

- Session that contains an access for robots.txt is a robot

```
0.182.141.149 - - [02/Feb/2012:06:20:46 +0000] "GET
http://web.archive.org/robots.txt HTTP/1.0" 200 125 "-"
"Mozilla/5.0 (compatible; MJ12bot/v1.4.1;
http://www.majestic12.co.uk/bot.php?+)"
```

```
0.182.141.149 - - [02/Feb/2012:06:20:19 +0000] "GET
http://wayback.archive.org/web/*/http://www.devilscafe.in
HTTP/1.1" 404 2168 "-" "Mozilla/5.0 (compatible;
MJ12bot/v1.4.1; http://www.majestic12.co.uk/bot.php?+)"
```

```
0.182.141.149 - - [02/Feb/2012:06:21:19 +0000] "GET
http://wayback.archive.org/web/*/http://www.genie.co.il
HTTP/1.1" 200 96205 "-" "Mozilla/5.0 (compatible;
MJ12bot/v1.4.1; http://www.majestic12.co.uk/bot.php?+)"
```

6 Requests, 2 Seconds → Robot

```
0.182.141.149 - - [02/Feb/2012:07:00:01 +0000] "GET
http://wayback.archive.org/web/*/http://www.cnn.com HTTP/1.1" 200 106433 "-"
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8)

0.182.141.149 - - [02/Feb/2012:07:00:01 +0000] "GET
http://wayback.archive.org/web/*/http://www.bbc.com HTTP/1.1" 200 566433 "-"
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8)

0.182.141.149 - - [02/Feb/2012:07:00:02 +0000] "GET
http://wayback.archive.org/web/*/http://www.google.com HTTP/1.1" 200 96433 "-"
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8)

0.182.141.149 - - [02/Feb/2012:07:00:02 +0000] "GET
http://wayback.archive.org/web/*/http://www.yahoo.com HTTP/1.1" 200 933333 "-"
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8)

0.182.141.149 - - [02/Feb/2012:07:00:02 +0000] "GET
http://wayback.archive.org/web/*/http://www.bing.com HTTP/1.1" 200 964333 "-"
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8)

0.182.141.149 - - [02/Feb/2012:07:00:3 +0000] "GET
http://wayback.archive.org/web/*/http://www.jcdl.org HTTP/1.1" 200 123233 "-"
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8)
```

3 Requests, 520 Seconds (9 Minutes) → Human

```
0.11.160.13 - - [02/Feb/2012:07:00:00 +0000] "GET
http://wayback.archive.org/web/*/http://www.cnn.com HTTP/1.1" 200 106433 "-"
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8)

0.11.160.13 - - [02/Feb/2012:07:03:46 +0000] "GET
http://wayback.archive.org/web/20100330042821/http://www.cnn.com HTTP/1.1" 200
566433 " http://wayback.archive.org/web/*/http://www.cnn.com" "Mozilla/5.0
(Macintosh; Intel Mac OS X 10_6_8)

0.11.160.13 - - [02/Feb/2012:07:08:00 +0000] "GET
http://wayback.archive.org/web/*/http://www.cnn.com HTTP/1.1" 200 96433 "
http://wayback.archive.org/web/*/http://www.cnn.com" "Mozilla/5.0 (Macintosh;
Intel Mac OS X 10_6_8)
```

Image-to-HTML Ratio

If I download these, I'm not a robot

Internet Archive
waybackmachine

233 captures
17 Aug 00 - 27 Jun 13

DEC JAN MAR
14
2011 2013 2014

Joint Conference on Digital Libraries

Upcoming Conference: [JCDL 2013: July 22-26, Indianapolis, IN](#)

Home About Sponsors Steering Committee Upcoming Events Past Conferences

Welcome to JCDL

The Joint Conference on Digital Libraries (JCDL) is a major international forum focusing on digital libraries and associated technical, practical, and social issues. JCDL enhances the tradition of conference excellence already established by the ACM and IEEE-CS by combining the annual events that these professional societies have sponsored on an annual basis, the ACM Digital Libraries Conferences and the IEEE-CS Advances in Digital Libraries Conferences.

Upcoming Events

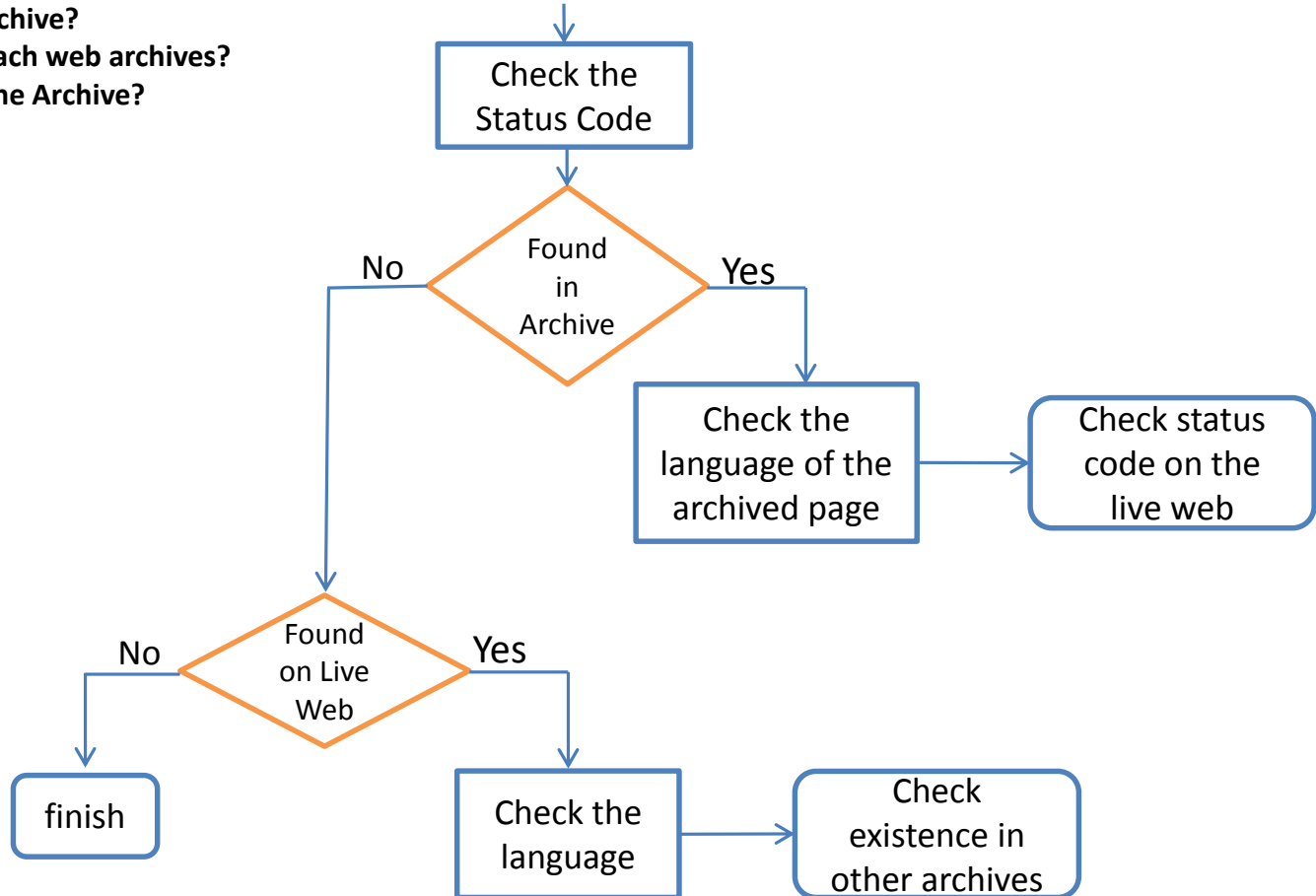
- [RCDL 2013: 15th All-Russian Conference, Yaroslavl, Russia, July 15-17, 2013](#)
- [JCDL 2013: 13th ACM/IEEE Conference on Digital Libraries, Indianapolis, IN, July 22-26, 2013](#)

Image-to-HTML Ratio

- The ratio between the number of image files and the number of HTML files per session
- Robots sessions are less than **1:10** image to HTML ratio (Stassopoulou et al. 2005)



- Who link to the archive?
- How do people reach web archives?
- Why they link to the Archive?
- Deep links?



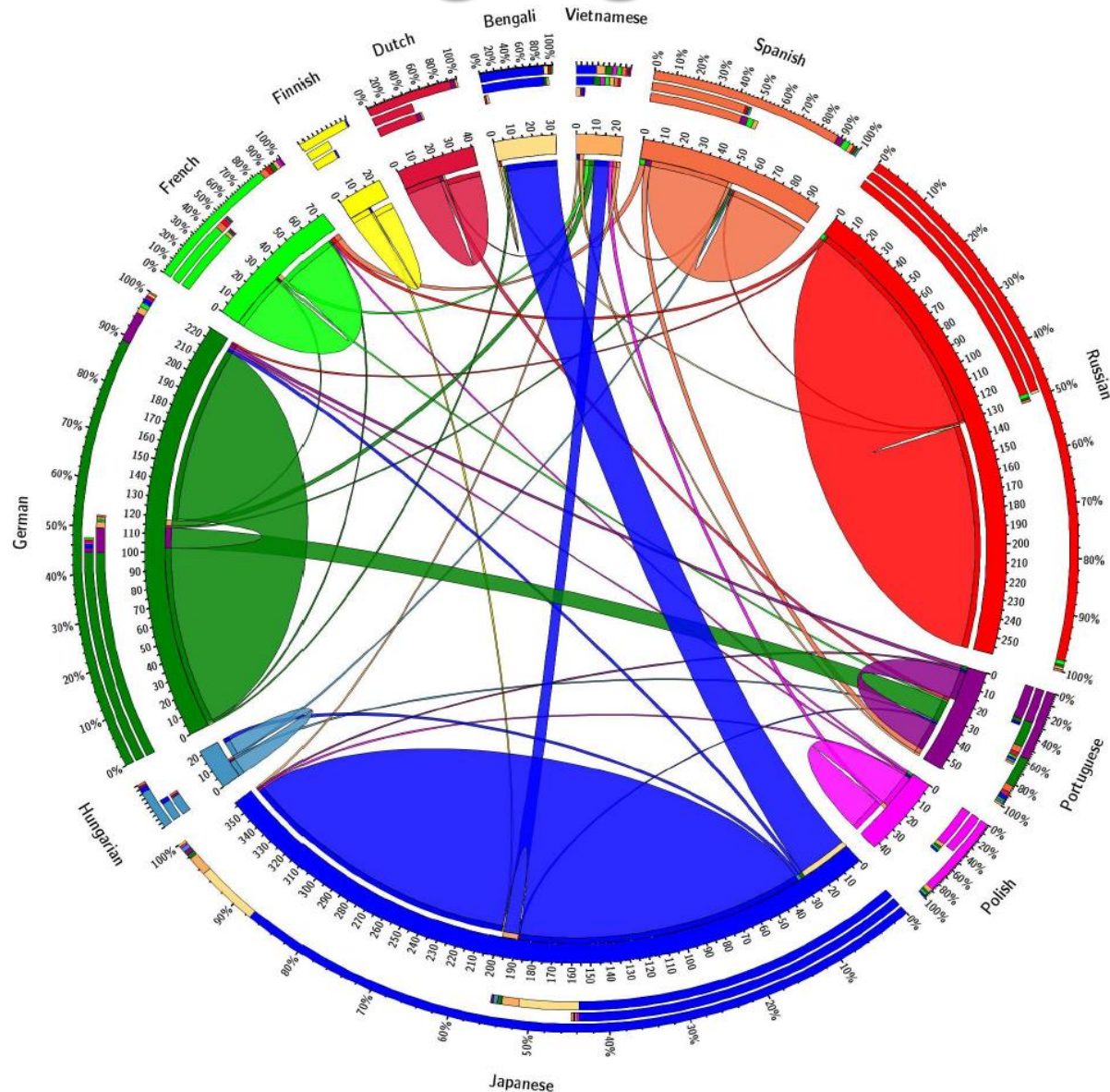
Languages for Pages in the Archive

Language	Humans	Language	Robots
English	71.7%	English	72.4%
Japanese	5.5%	Russian	7.0%
German	3.6%	German	3.1%
Vietnamese	2.9%	Spanish	1.9%
Russian	2.3%	French	1.8%
Portuguese	2.1%	Vietnamese	1.7%
French	2.1%	Japanese	1.5%
Spanish	1.9%	Polish	1.5%
Bengali	1.8%	Portuguese	1.3%
Italian	0.9%	Thai	1.1%

Languages for Requested Pages NOT in the Archive


Language	Humans	Language	Robots
English	66.9%	English	62.2%
Russian	7.9%	Russian	11.1%
German	5.4%	German	3.8%
Japanese	5.1%	Indonesian	3.1%
Spanish	2.5%	Polish	2.5%
Polish	2.3%	Vietnamese	2.2%
Romanian	1.6%	Spanish	2.0%
French	1.2%	Thai	1.9%
Italian	0.8%	French	1.8%
Portuguese	0.7%	Dutch	1.1%

Most Languages Self-Link



The Existence of the Archived Pages on the Live Web

	Humans	Robots
URI-Rs available on live web	36.4%	62.5%
URI-Rs missing from live web	63.6%	37.5%



Humans come to the archive because they can't find web pages on the live web

The Existence of Unarchived Pages on the Live Web

	Humans	Robots
URI-Rs available on live web	25.4%	33.2%
URI-Rs missing from live web	74.6%	66.8%

Existence in Other Web Archives

Web Archive	#URI-R	#URI-M
Internet Archive (2013)	56,503	1,657,264
The National Archives	787	15,354
ArchiefWeb	47	18,347
Archive-It	41	4,682
UK Web Archive	38	12,277
Library of Congress	35	1,092
WebCite	29	1,104

The number of the requested pages not in the archive is **211,825** (2012)

82% of Human Sessions Have Referring URIS

WebSite	Percentage	Description
en.wikipedia.org	12.9%	Wikipedia
archive.org	11.9%	IA Home Page
reddit.com	10.2%	Social News Web Site
google.TLD	9.9%	Search Engine
info-poland.buffalo.edu	1.5%	Polish Studies
de.wikipedia.org	1.4%	Wikipedia
cracked.com	1.2%	Humor Site
snopes.com	1.1%	Urban Legends Reference Pages
facebook.com	0.9%	Social Media
crochetpatterncentral.com	0.9%	Crocheting Hobbies

Many European Domains Link to IA

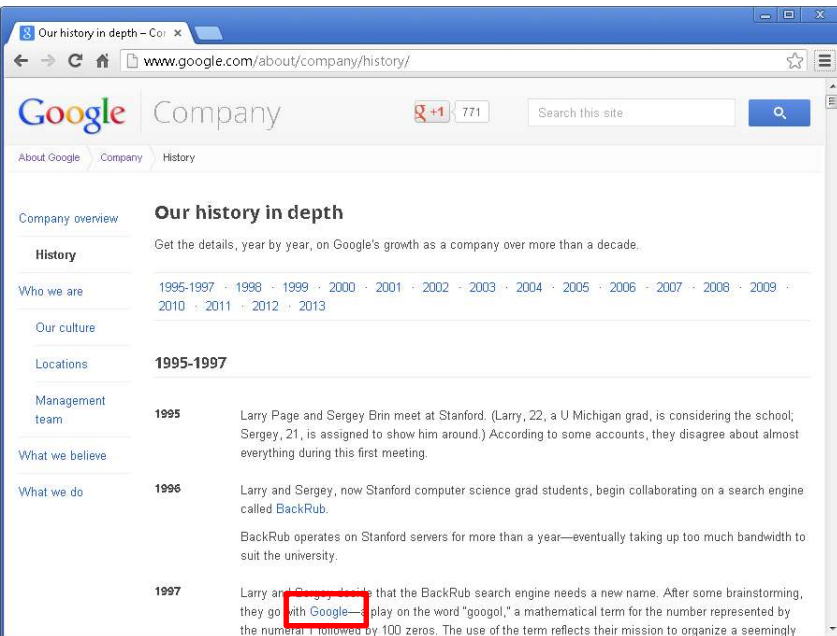
TLD	.com	.org	.net	.jp	.ru	.de	.edu	.to	.uk	.info
Percentage	45.4%	33.9%	8.4%	1.8%	1.4%	1.4%	1.1%	0.7%	0.6%	0.5%

The top 10 TLDs of the referrers.

ccTLD	.com	.uk	.de	.ca	.jp	.pl	.nl	.ru	.fr	.br
Percentage	56.7%	6.0%	5.3%	4.8%	3.7%	2.2%	1.9%	1.7%	1.5%	1.4%

The top 10 ccTLDs of Google search referrers.

Most of the Links (86%) Are to Mementos



Our history in depth - Co: x
www.google.com/about/company/history/

Google Company

About Google Company History

Our history in depth

Get the details, year by year, on Google's growth as a company over more than a decade.

History

Who we are: 1995-1997 · 1998 · 1999 · 2000 · 2001 · 2002 · 2003 · 2004 · 2005 · 2006 · 2007 · 2008 · 2009 · 2010 · 2011 · 2012 · 2013

Our culture

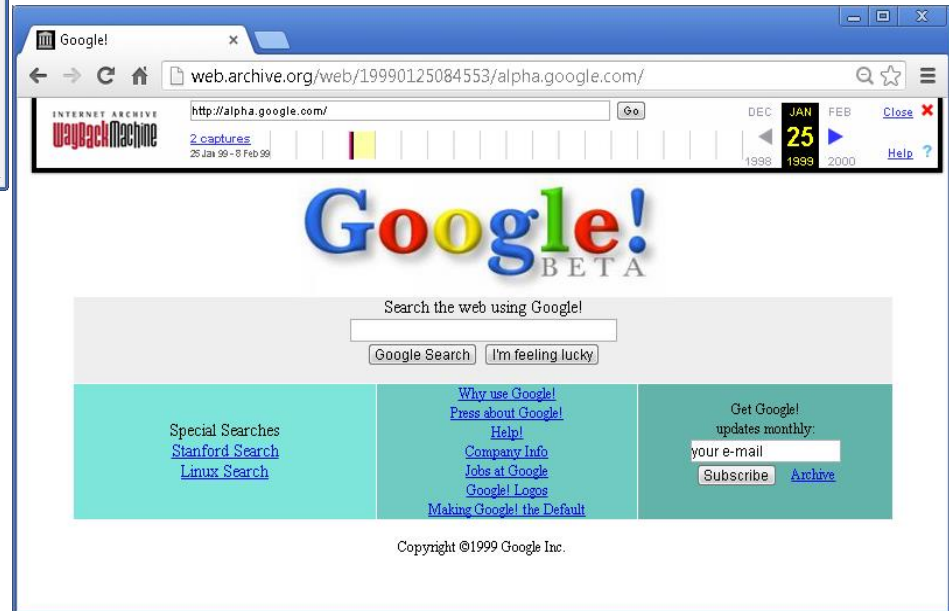
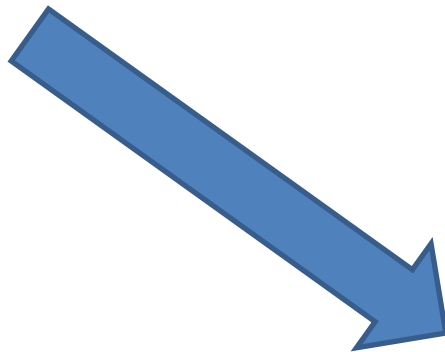
Locations

1995-1997

1995 Larry Page and Sergey Brin meet at Stanford. (Larry, 22, a U Michigan grad, is considering the school; Sergey, 21, is assigned to show him around.) According to some accounts, they disagree about almost everything during this first meeting.

1996 Larry and Sergey, now Stanford computer science grad students, begin collaborating on a search engine called BackRub. BackRub operates on Stanford servers for more than a year—eventually taking up too much bandwidth to suit the university.

1997 Larry and Sergey realize that the BackRub search engine needs a new name. After some brainstorming, they go with Google—play on the word "googol," a mathematical term for the number represented by the numeral 1 followed by 100 zeros. The use of the term reflects their mission to organize a seemingly



Google!

web.archive.org/web/19990125084553/alpha.google.com/

INTERNET ARCHIVE WaybackMachine

http://alpha.google.com/

2 captures 25 Jan 99 - 8 Feb 99

DEC JAN FEB Close

1998 1999 2000 Help ?

Google!

BETA

Search the web using Google!

Google Search I'm feeling lucky

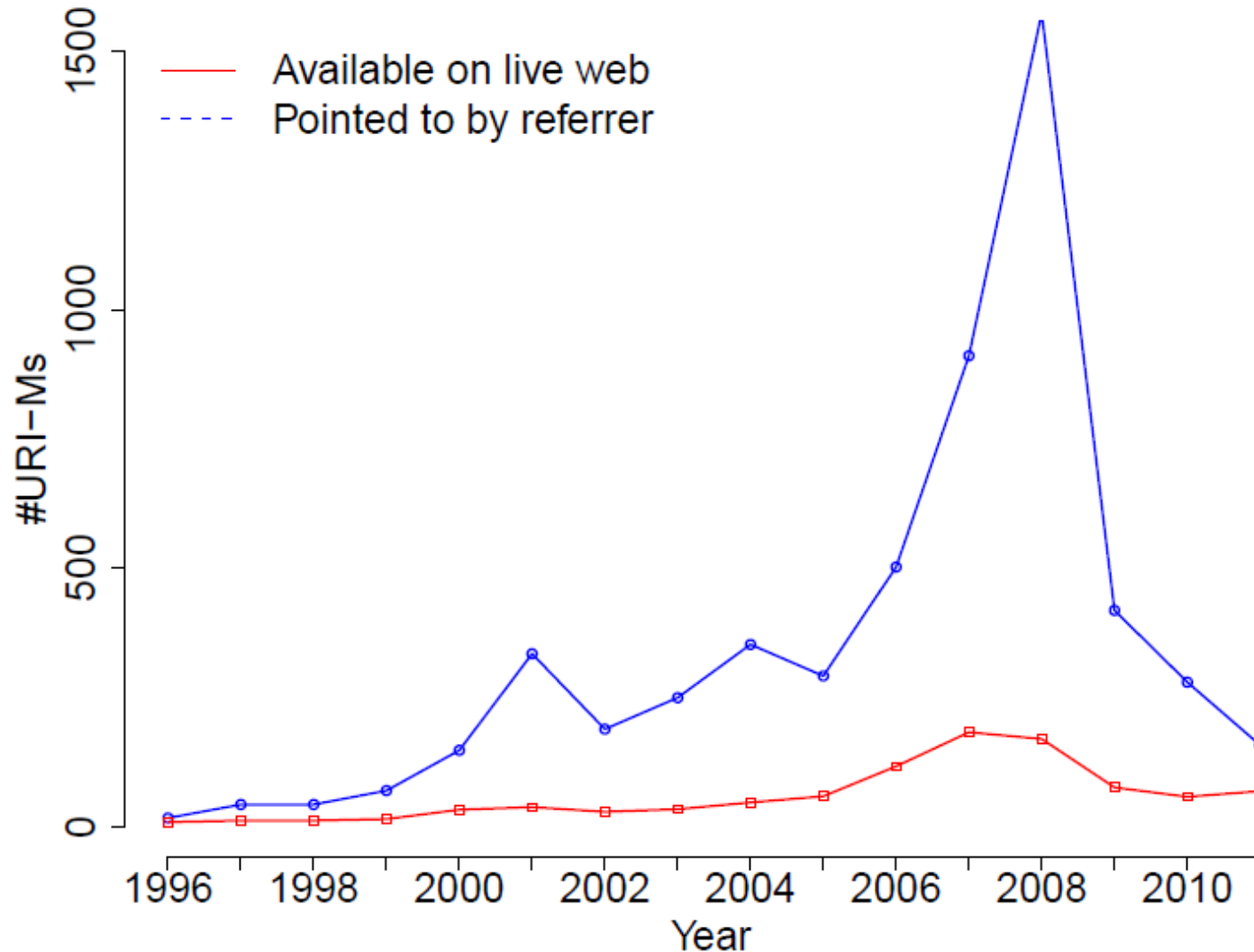
Special Searches: Stanford Search, Linux Search

Why use Google! Press about Google! Help! Company Info Jobs at Google Google! Logos Making Google! the Default

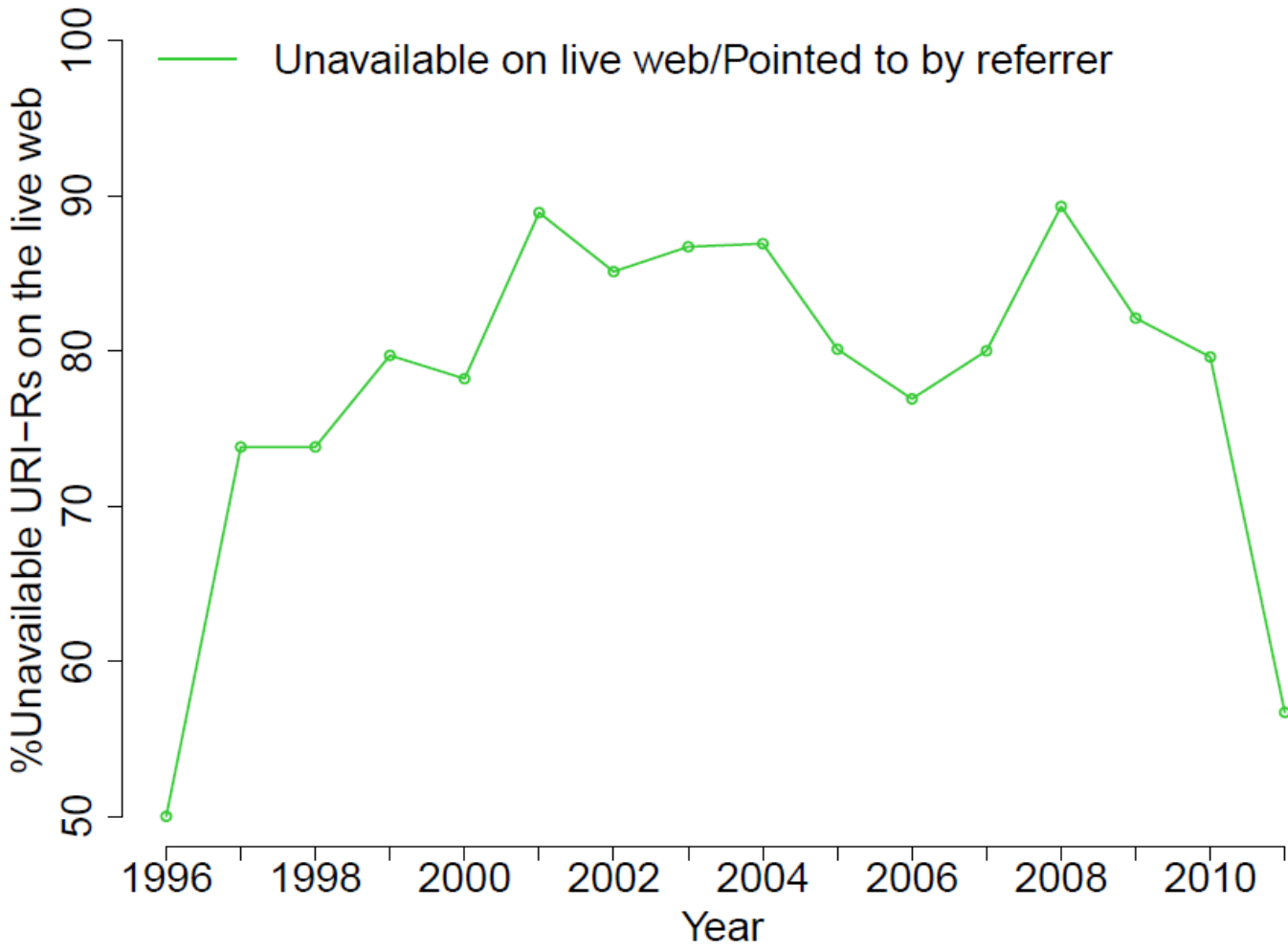
Get Google! updates monthly: your e-mail Subscribe Archive

Copyright ©1999 Google Inc.

Significant Bias for the Recent Past



For 83% of Externally Linked Mementos, Corresponding Original URI is 404 on Live Web



Conclusions

- English is the most common language, followed by many European languages, and Japanese & Vietnamese
- Languages self-link (and link to English)
- 82% of human sessions have referrals
- 86% of the referring web pages link deeply to mementos
- 83% of the links to these mementos are because their corresponding URI-Rs do not exist on the live web