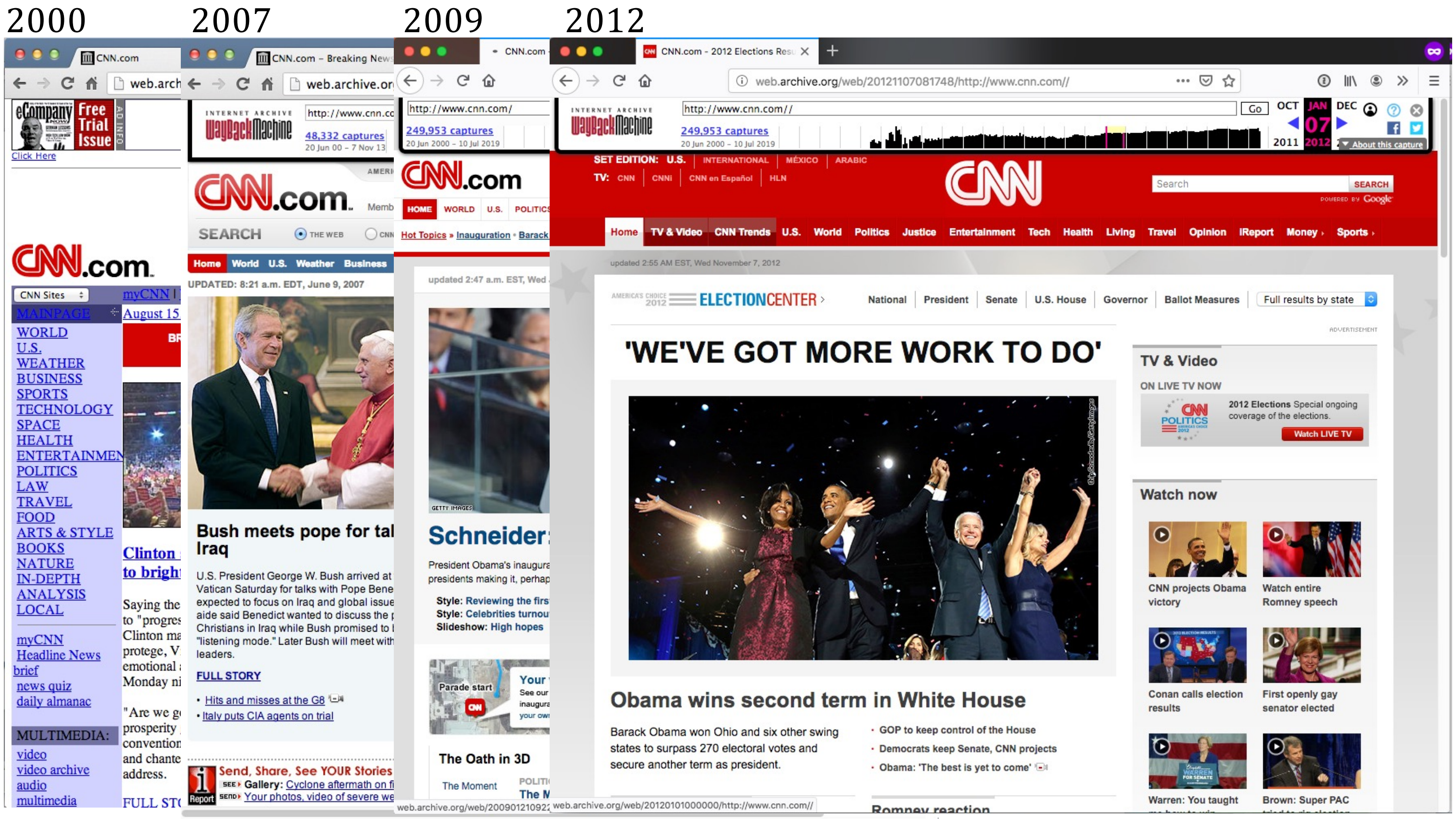
Comparing the Archival Rate of Arabic, English, Danish, and Korean Language Web Pages

Lulwah M. Alkwai, Michael L. Nelson, **Michele C. Weigle** (@weiglemc)
Web Sciences and Digital Libraries (WS-DL) Group (@WebSciDL)
Department of Computer Science
Old Dominion University
Norfolk, Virginia, USA

Published in *ACM Transactions on Information Systems (TOIS),* 36(1), July 2017
Extended version of Best Student Paper awardee at IEEE/ACM JCDL 2015

July 24, 2019 / ACM SIGIR 2019 / Paris, France

Web archives are collections of web pages of the past



web.archive.org/web/20121107081748/http://www.cnn.com//

http://www.cnn.com//

249,953 captures
20 Jun 2000 – 10 Jul 2019

SET EDITION: U.S. | INTERNATIONAL | MÉXICO | ARABIC
TV: CNN | CNNI | CNN en Español | HLN

Home | TV & Video | CNN Trends | U.S. | World | Politics | Justice | Entertainment | Tech | Health | Living | Travel | Opinion | iReport | Money | Sports

updated 2:55 AM EST, Wed November 7, 2012

ELECTION CENTER › National | President | Senate | U.S. House | Governor | Ballot Measures | Full results by state

'WE'VE GOT MORE WORK TO DO'

Obama wins second term in White House

Barack Obama won Ohio and six other swing states to surpass 270 electoral votes and secure another term as president.

- GOP to keep control of the House
- Democrats keep Senate, CNN projects
- Obama: 'The best is yet to come'

Romney reaction

TV & Video

ON LIVE TV NOW

2012 Elections Special ongoing coverage of the elections.

Watch LIVE TV

Watch now

CNN projects Obama victory

Watch entire Romney speech

Conan calls election results

First openly gay senator elected

Warren: You taught

Brown: Super PAC

UPDATED: 8:21 a.m. EDT, June 9, 2007

Bush meets pope for tal[ks on] Iraq

U.S. President George W. Bush arrived at Vatican Saturday for talks with Pope Bene[dict] expected to focus on Iraq and global issue[s] aide said Benedict wanted to discuss the [p]light of Christians in Iraq while Bush promised to [be in] "listening mode." Later Bush will meet with leaders.

FULL STORY

- Hits and misses at the G8
- Italy puts CIA agents on trial

Send, Share, See YOUR Stories

Schneider

President Obama's inaugura[l] presidents making it, perhap[s]

Style: Reviewing the firs[t]
Style: Celebrities turnou[t]
Slideshow: High hopes

The Oath in 3D

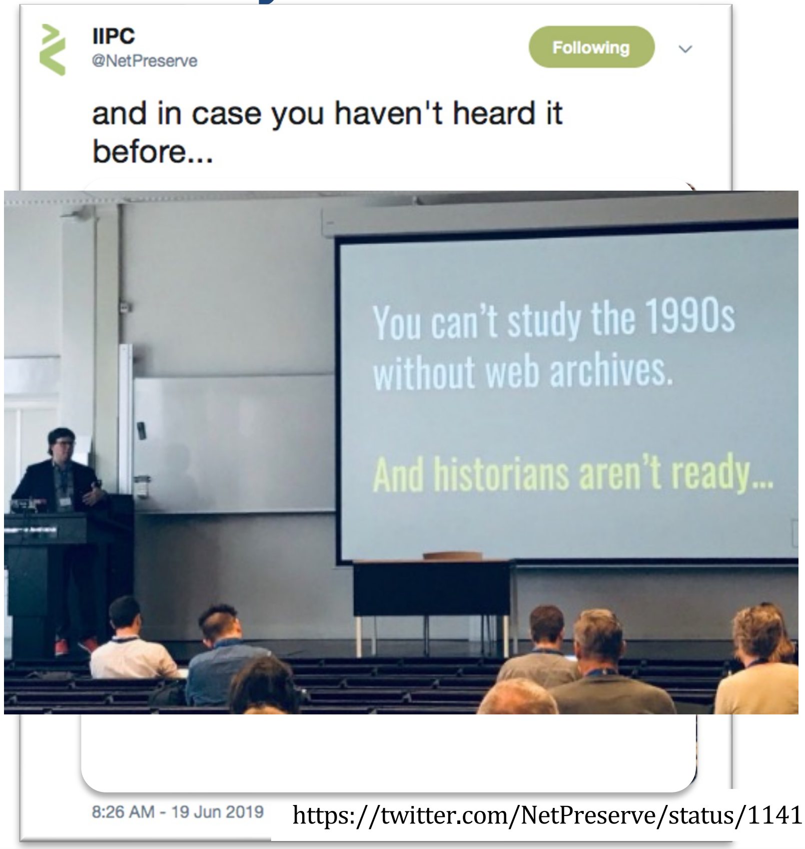
CNN.com

WORLD
U.S.
WEATHER
BUSINESS
SPORTS
TECHNOLOGY
SPACE
HEALTH
ENTERTAINMENT
POLITICS
LAW
TRAVEL
FOOD
ARTS & STYLE
BOOKS
NATURE
IN-DEPTH
ANALYSIS
LOCAL

myCNN
Headline News brief
news quiz
daily almanac

MULTIMEDIA:
video
video archive
audio
multimedia

Web archives are essential for studying recent history and culture



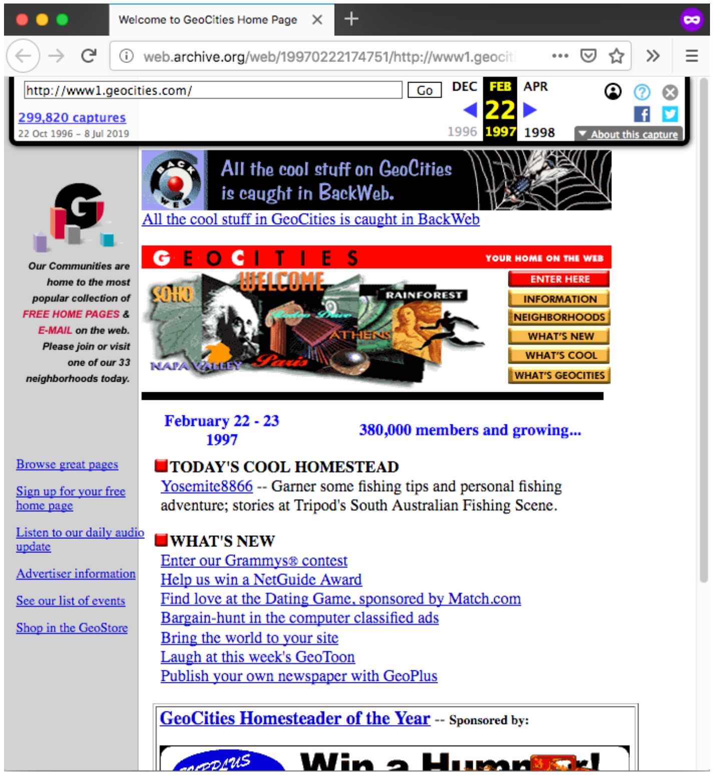
IIPC
@NetPreserve

and in case you haven't heard it before...

You can't study the 1990s without web archives.
And historians aren't ready...

8:26 AM - 19 Jun 2019

https://twitter.com/NetPreserve/status/1141321443373920256 *(photo cropped and enlarged)*



299,820 captures
22 Oct 1996 – 8 Jul 2019

All the cool stuff in GeoCities is caught in BackWeb

Our Communities are home to the most popular collection of FREE HOME PAGES & E-MAIL on the web. Please join or visit one of our 33 neighborhoods today.

February 22 - 23 1997

380,000 members and growing...

- Browse great pages
- Sign up for your free home page
- Listen to our daily audio update
- Advertiser information
- See our list of events
- Shop in the GeoStore

TODAY'S COOL HOMESTEAD

Yosemite8866 -- Garner some fishing tips and personal fishing adventure; stories at Tripod's South Australian Fishing Scene.

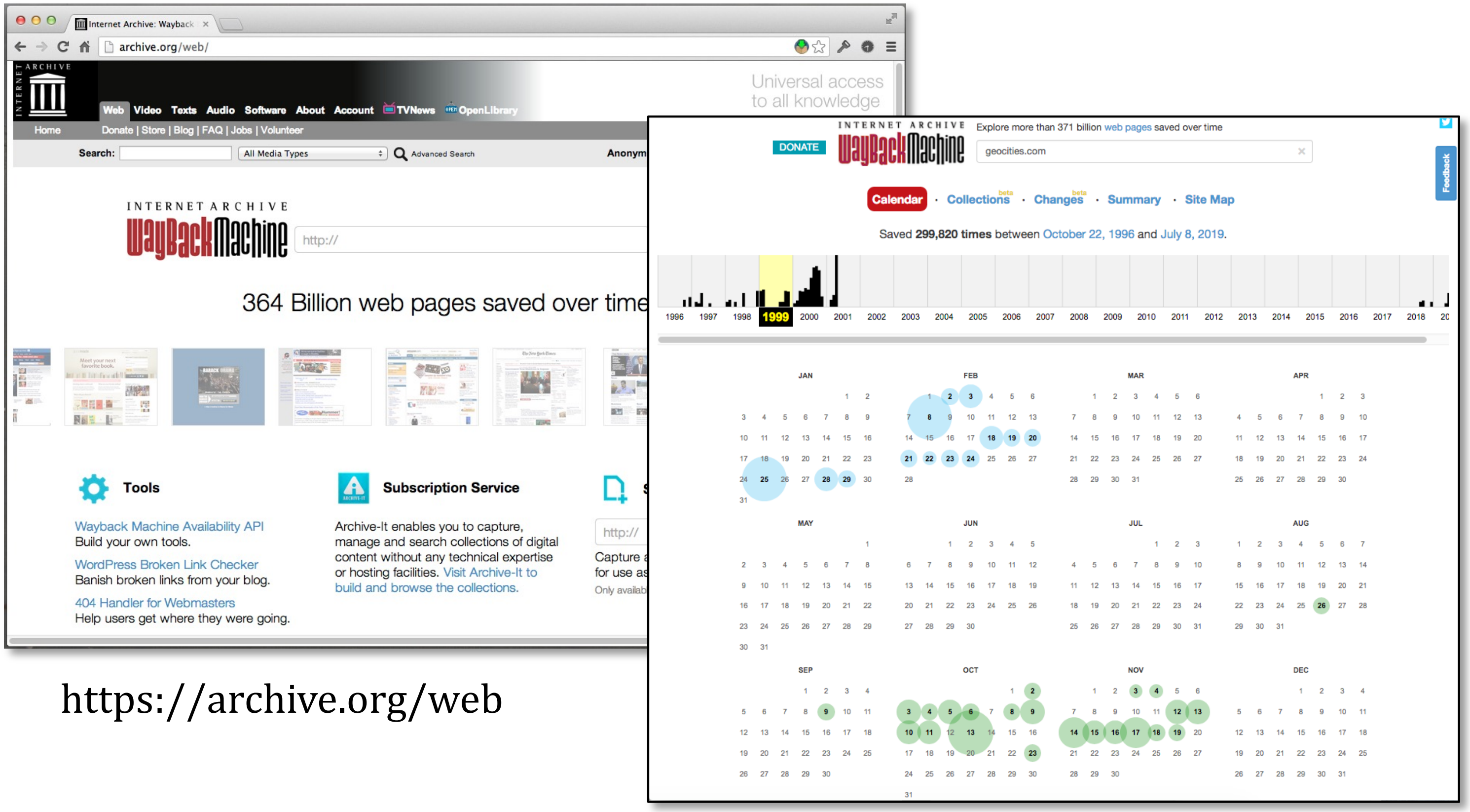
WHAT'S NEW

Enter our Grammys® contest
Help us win a NetGuide Award
Find love at the Dating Game, sponsored by Match.com
Bargain-hunt in the computer classified ads
Bring the world to your site
Laugh at this week's GeoToon
Publish your own newspaper with GeoPlus

GeoCities Homesteader of the Year -- **Sponsored by:**

http://web.archive.org/web/19970222174751/http://www1.geocities.com/

The Internet Archive holds the largest web archive



364 Billion web pages saved over time

Wayback Machine Availability API
Build your own tools.

WordPress Broken Link Checker
Banish broken links from your blog.

404 Handler for Webmasters
Help users get where they were going.

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.

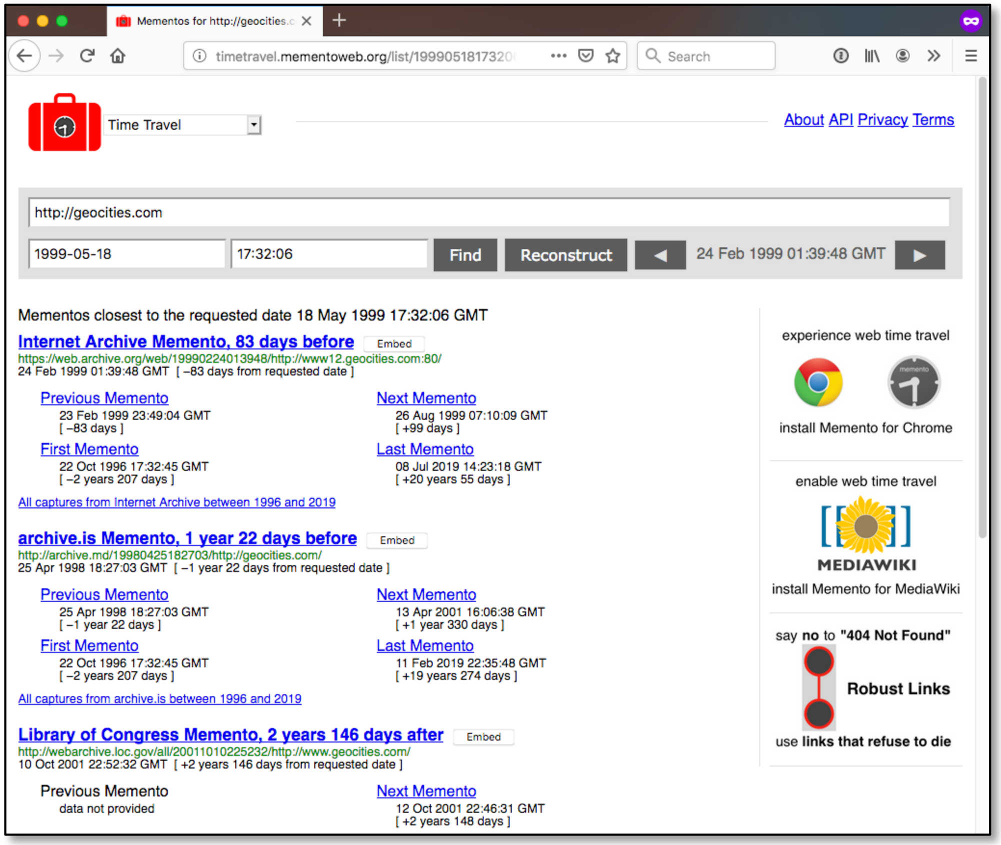
Capture a
for use as
Only availab

Explore more than 371 billion web pages saved over time

Saved 299,820 times between October 22, 1996 and July 8, 2019.

https://archive.org/web

But it’s not the only one


Mementos for http://geocities.c
timetravel.mementoweb.org/list/1999051817320
Search
Time Travel
About API Privacy Terms

http://geocities.com

1999-05-18 | 17:32:06 | Find | Reconstruct | ◀ | 24 Feb 1999 01:39:48 GMT | ▶

Mementos closest to the requested date 18 May 1999 17:32:06 GMT

Internet Archive Memento, 83 days before Embed
https://web.archive.org/web/19990224013948/http://www12.geocities.com:80/
24 Feb 1999 01:39:48 GMT [–83 days from requested date]

Previous Memento
23 Feb 1999 23:49:04 GMT
[–83 days]

Next Memento
26 Aug 1999 07:10:09 GMT
[+99 days]

First Memento
22 Oct 1996 17:32:45 GMT
[–2 years 207 days]

Last Memento
08 Jul 2019 14:23:18 GMT
[+20 years 55 days]

All captures from Internet Archive between 1996 and 2019

archive.is Memento, 1 year 22 days before Embed
http://archive.md/19980425182703/http://geocities.com/
25 Apr 1998 18:27:03 GMT [–1 year 22 days from requested date]

Previous Memento
25 Apr 1998 18:27:03 GMT
[–1 year 22 days]

Next Memento
13 Apr 2001 16:06:38 GMT
[+1 year 330 days]

First Memento
22 Oct 1996 17:32:45 GMT
[–2 years 207 days]

Last Memento
11 Feb 2019 22:35:48 GMT
[+19 years 274 days]

All captures from archive.is between 1996 and 2019

Library of Congress Memento, 2 years 146 days after Embed
http://webarchive.loc.gov/all/20011010225232/http://www.geocities.com/
10 Oct 2001 22:52:32 GMT [+2 years 146 days from requested date]

Previous Memento
data not provided

Next Memento
12 Oct 2001 22:46:31 GMT
[+2 years 148 days]

experience web time travel

install Memento for Chrome

enable web time travel

MEDIAWIKI

install Memento for MediaWiki

say no to "404 Not Found"

Robust Links

use links that refuse to die

http://timetravel.mementoweb.org/list/19990518173206/http://geocities.com

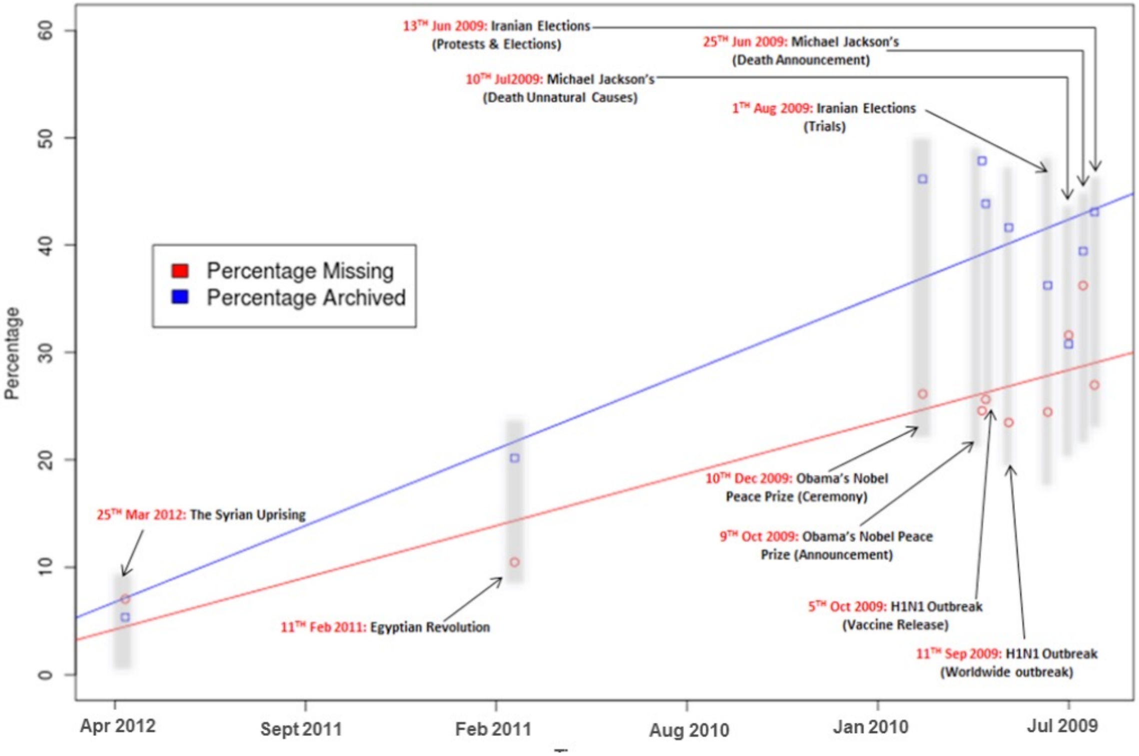
The National Archives
Vefsafn.is
Landsbókasafn Íslands - Háskólabókasafn
LIBRARY OF CONGRESS
UKWA
UK WEB ARCHIVE
NATIONAL ARCHIVES
archive.is
webpage capture
PORTUGUESE WEB ARCHIVE
Webrecorder
spletni ARHIV NUK
ARCHIVE-IT
EESTI VEEBIARHIIV

https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives

We’ve studied recent (2010s) events in the Middle East

- Iranian Elections and Protests - Jun 2009
  - SalahEldeen and Nelson, TPDL 2012
- Egyptian Revolution - Jan 20 - Mar 1, 2011
  - SalahEldeen and Nelson, TPDL 2012
  - AlNoamany, Weigle, Nelson, ACM WebSci 2017
- Syrian Uprising - Mar 2012
  - SalahEldeen and Nelson, TPDL 2012
- Egypt’s Presidential Election - 2012
  - AlNoamany, Weigle, Nelson, TPDL 2015, IJDL 2016

But, we can only study the past Web that already exists in the archives



11% of resources shared in social media disappear each year

Hany M. SalahEldeen, Michael L. Nelson, Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost?, *Proceedings of TPDL* 2012.

How well is the Arabic Web archived?

- Arabic is the 4th most popular language on the Internet
- Anecdotally known that web archives and search engines favor Western and English language pages

We investigated the state of archival of Arabic language web pages in 2014-2015

- Gathered Arabic language web pages
- Analyzed domains, TLDs, GeoIP
- Analyzed presence in Google's index and web archives
- Compared this to the archival and indexing of English, Danish, and Korean language web pages

How to gather and detect Arabic language web pages?

Gathered URIs from Arabic website directories

Directory	Registered Country	Year Estab.	Directory URI	URIs
DMOZ	US	1999	Dmoz.org/world/arabic	4,086
Raddadi	Saudi Arabia	2000	Raddadi.com	3,271
Star28	Lebanon	2004	Star28.com	8,386
Total				**15,743**

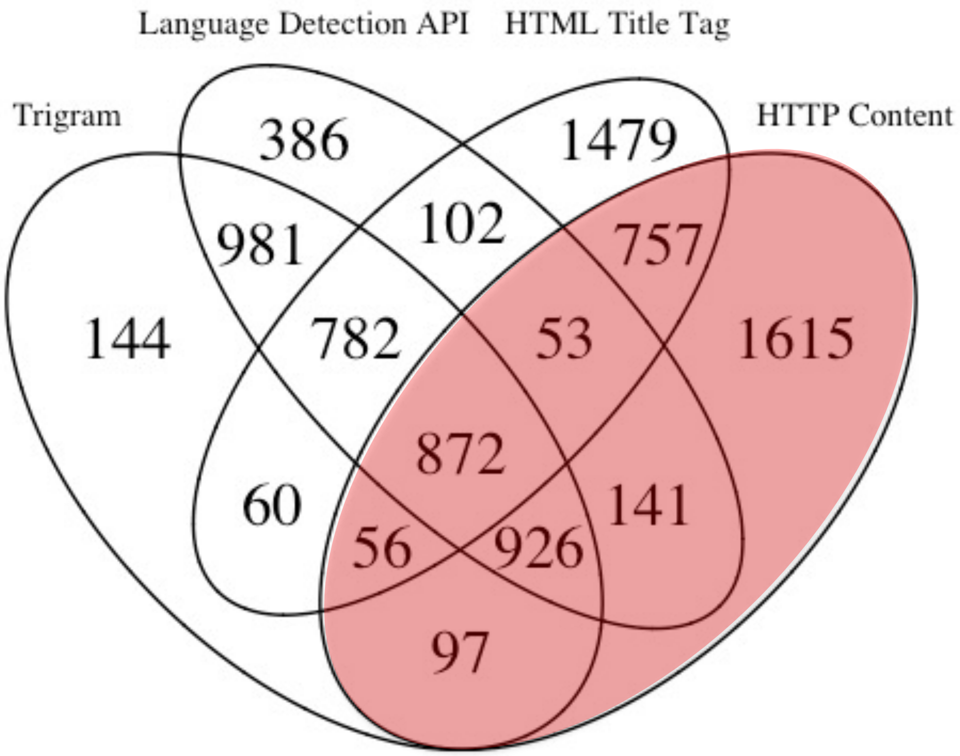
- 15,743 seed URIs
  - 15,092 were unique
  - **11,014** were live on the Web in March-May 2014

We used four methods to determine the language of the 11,014 URIs

- HTTP Content-Language header
- HTML Title tag https://github.com/kent37/guess-language
- Language Detection API http://detectlanguage.com
- Trigrams https://github.com/decultured/Python-Language-Detector

Sample of test results evaluated by a native speaker (first author)

HTTP Content-Language classified 41% of the seed URIs as Arabic



HTTP response header

ex:

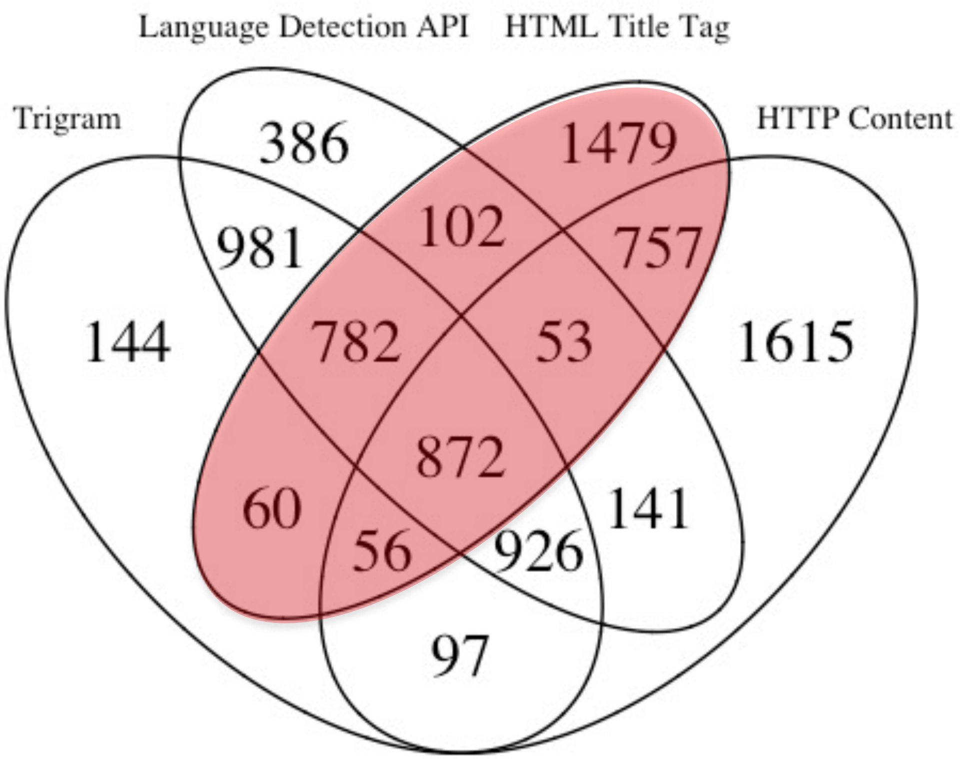
```
Content-Language: ar
```

HTML tag

ex:

```
<html dir="rtl"
xmlns="http://www.w3.org/19
99/xhtml" xml:lang="ar"
lang="ar">
```

HTML Title tag classified 38% of the seed URIs as Arabic



Extract text from HTML title tag

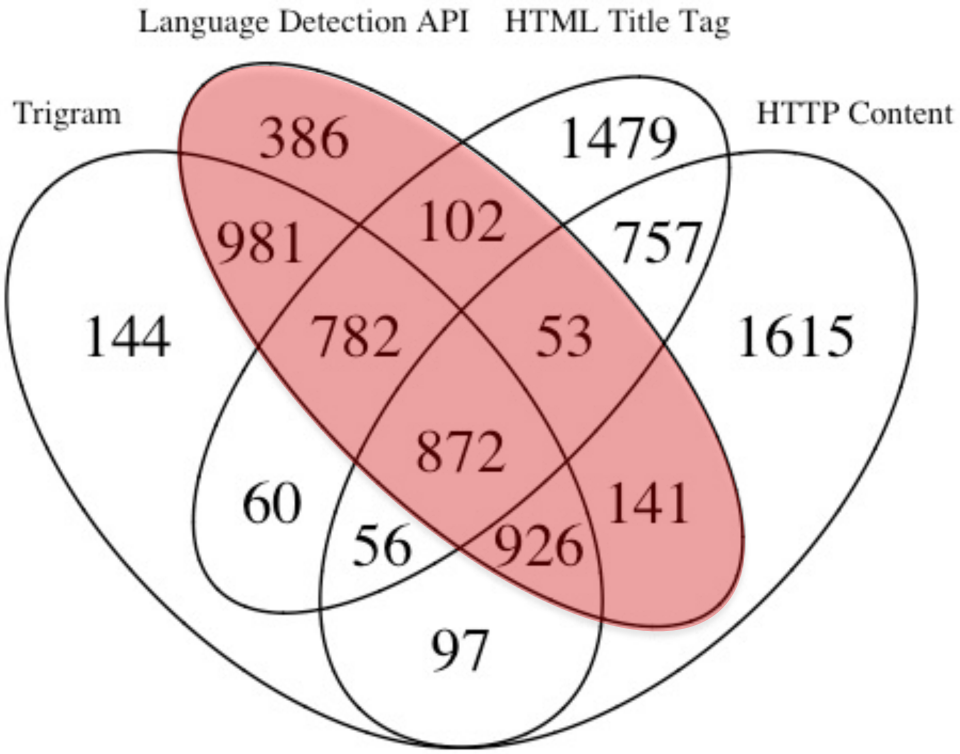
ex:

`<TITLE>دليل العرب الشامل</TITLE>`

guess-language library

https://github.com/kent37/guess-language

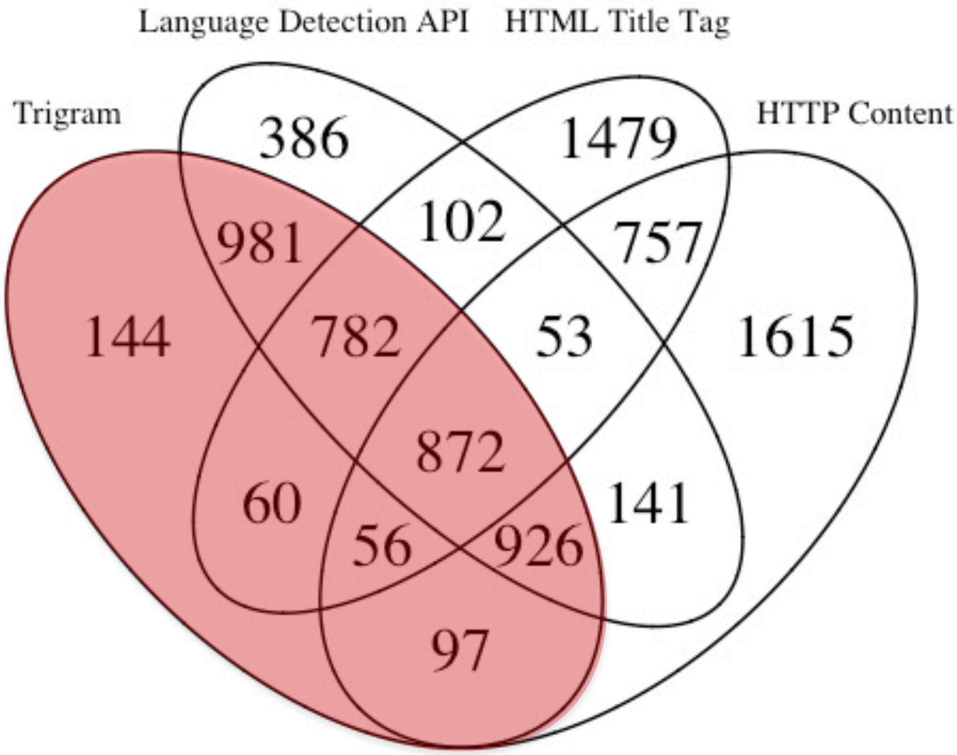
The Language Detection API test classified 39% of the seed URIs as Arabic



Extract title and text

Language Detection API
http://detectlanguage.com

The Trigram test classified 36% of the seed URIs as Arabic

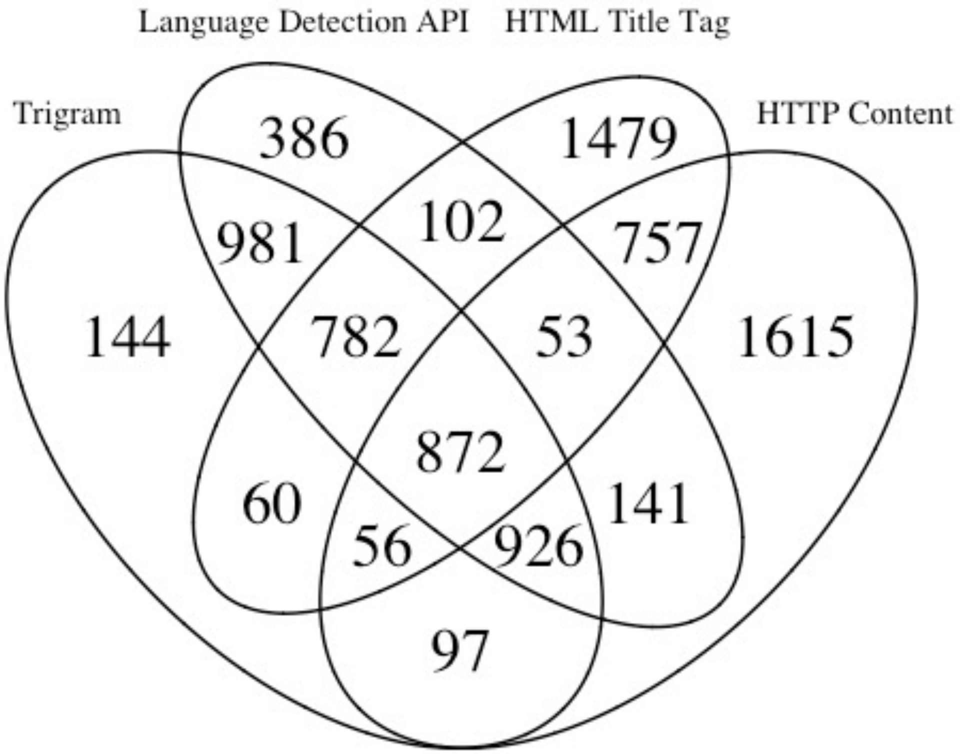


Extract title and text

Test sequences of three letters (trigrams)

Python-Language-Detector tool
https://github.com/decultured/Python-Language-Detector

We took the union to obtain 7976 Arabic seed URIs

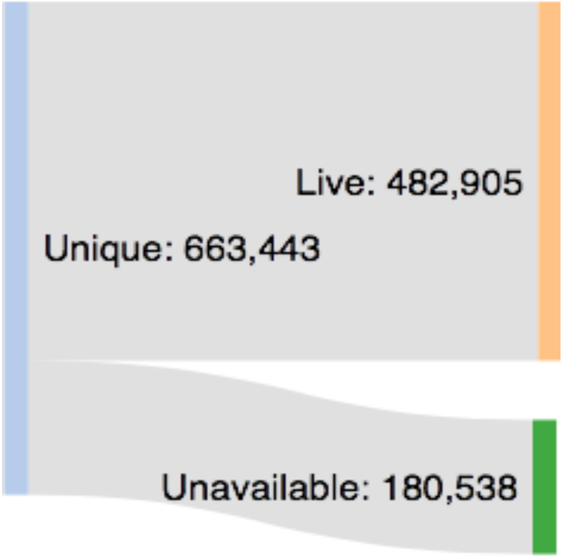


72.4% of the seed URIs were determined to be in Arabic

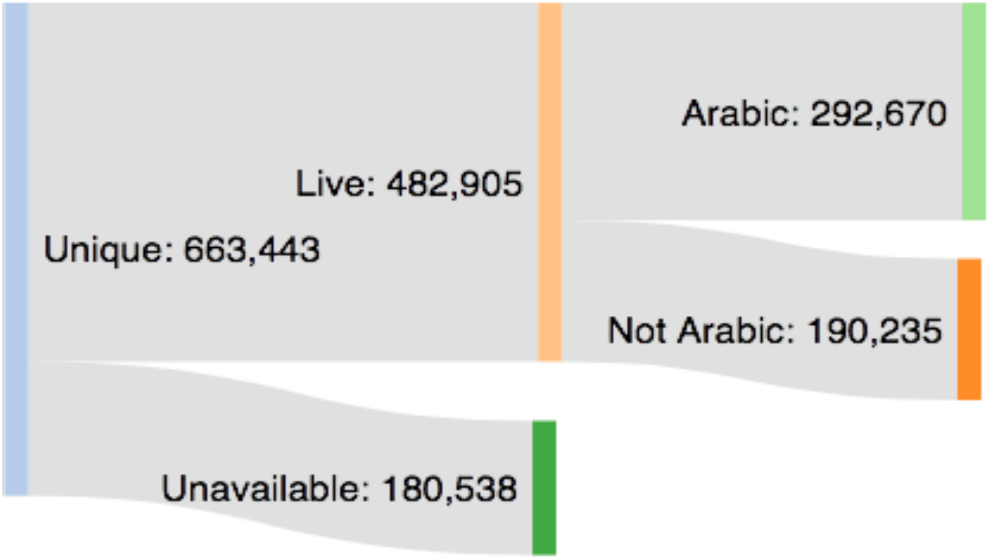
We expanded the dataset by crawling the live Web and the past Web

- Crawled the 7976 live Arabic seed URIs, 2 levels deep
  - gathered all URIs linked from each seed URI
  - then, gathered all URIs linked from those URIs
  - 575,242 additional URIs
- Crawled the most recent memento of the Arabic seed URIs
  - 515,821 additional URIs
- Total of 663,443 unique crawled URIs gathered

482,905 were live



292,670 were live and Arabic



Total Arabic Dataset = **300,646 URIs**

7,976 seed URIs

+ 292,670 crawled URIs

What are the characteristics of our Arabic language dataset?

Dataset has 17,536 unique domains

Rank	Domain	URIs	GeoIP	Category
1	Alarab.net	284	US	News
2	Aljarida.com	248	US	News
3	Arabic.cnn.com	245	US	News
4	Alarabiya.net	231	US	News
5	Ar.wikipedia.org	230	US	Encyclopedia
6	Aljazeera.net	213	US	News
7	Moheet.com	142	US	News
8	Facebook.com	133	US	Social
9	Al-sharq.com	132	US	Middle East Portal
10	Lakii.com	123	US	General Portal
17	Kuwaitclub.com.kw	71	Kuwait	Sport

First Arabic GeoIP is at rank 17

Rank	Domain	URIs	GeoIP	Category
1	Alarab.net	284	US	News
2	Aljarida.com	248	US	News
3	Arabic.cnn.com	245	US	News
4	Alarabiya.net	231	US	News
5	Ar.wikipedia.org	230	US	Encyclopedia
6	Aljazeera.net	213	US	News
7	Moheet.com	142	US	News
8	Facebook.com	133	US	Social
9	Al-sharq.com	132	US	Middle East Portal
10	Lakii.com	123	US	General Portal
17	Kuwaitclub.com.kw	71	Kuwait	Sport

6 top domains are news websites

Rank	Domain	URIs	GeoIP	Category
1	Alarab.net	284	US	News
2	Aljarida.com	248	US	News
3	Arabic.cnn.com	245	US	News
4	Alarabiya.net	231	US	News
5	Ar.wikipedia.org	230	US	Encyclopedia
6	Aljazeera.net	213	US	News
7	Moheet.com	142	US	News
8	Facebook.com	133	US	Social
9	Al-sharq.com	132	US	Middle East Portal
10	Lakii.com	123	US	General Portal
17	Kuwaitclub.com.kw	71	Kuwait	Sport

Popular Western domains are in the Top 10

Rank	Domain	URIs	GeoIP	Category
1	Alarab.net	284	US	News
2	Aljarida.com	248	US	News
3	Arabic.cnn.com	245	US	News
4	Alarabiya.net	231	US	News
5	Ar.wikipedia.org	230	US	Encyclopedia
6	Aljazeera.net	213	US	News
7	Moheet.com	142	US	News
8	Facebook.com	133	US	Social
9	Al-sharq.com	132	US	Middle East Portal
10	Lakii.com	123	US	General Portal
17	Kuwaitclub.com.kw	71	Kuwait	Sport

Over half have a .com TLD

TLD	Percent
com	57.97%
net	15.07%
org	6.40%
gov.sa	1.94%
info	1.68%
edu.sa	1.27%
ws	1.16%
org.sa	0.97%
com.sa	0.80%
gov.eg	0.80%
Other	11.94%

Only ~10% have an Arabic ccTLD

TLD	Percent
com	57.97%
net	15.07%
org	6.40%
gov.sa	1.94%
info	1.68%
edu.sa	1.27%
ws	1.16%
org.sa	0.97%
com.sa	0.80%
gov.eg	0.80%
Other	11.94%

ccTLD	Country	Percent
.sa	Saudi Arabia	5.33%
.eg	Egypt	2.00%
.jo	Jordan	2.00%
.ae	United Arab Emirates	1.06%
.kw	Kuwait	0.82%

Most are geo-located in the US

Geo-location	Percent
United States	57.97%
Arabic Countries	10.53%
Germany	9.75%
Netherlands	5.29%
France	4.37%
Canada	3.31%
United Kingdom	3.07%
Other	5.71%

Within Arabic countries, most are geo-located in Saudi Arabia

Geo-location	Percent
United States	57.97%
Arabic Countries	10.53%
Germany	9.75%
Netherlands	5.29%
France	4.37%
Canada	3.31%
United Kingdom	3.07%
Other	5.71%

Geo-location	Percent
Saudi Arabia	4.75%
Egypt	1.97%
Jordan	1.42%
Kuwait	0.71%
United Arab Emirates	0.67%

How well are these Arabic web pages indexed and archived?

53.77% of Arabic language web pages are archived

- Used Memento aggregator to determine if archived
  - checks multiple web archives
- 161,678 URIs were archived
- 97% of those were found in the Internet Archive
- Mementos also found in 9 other archives

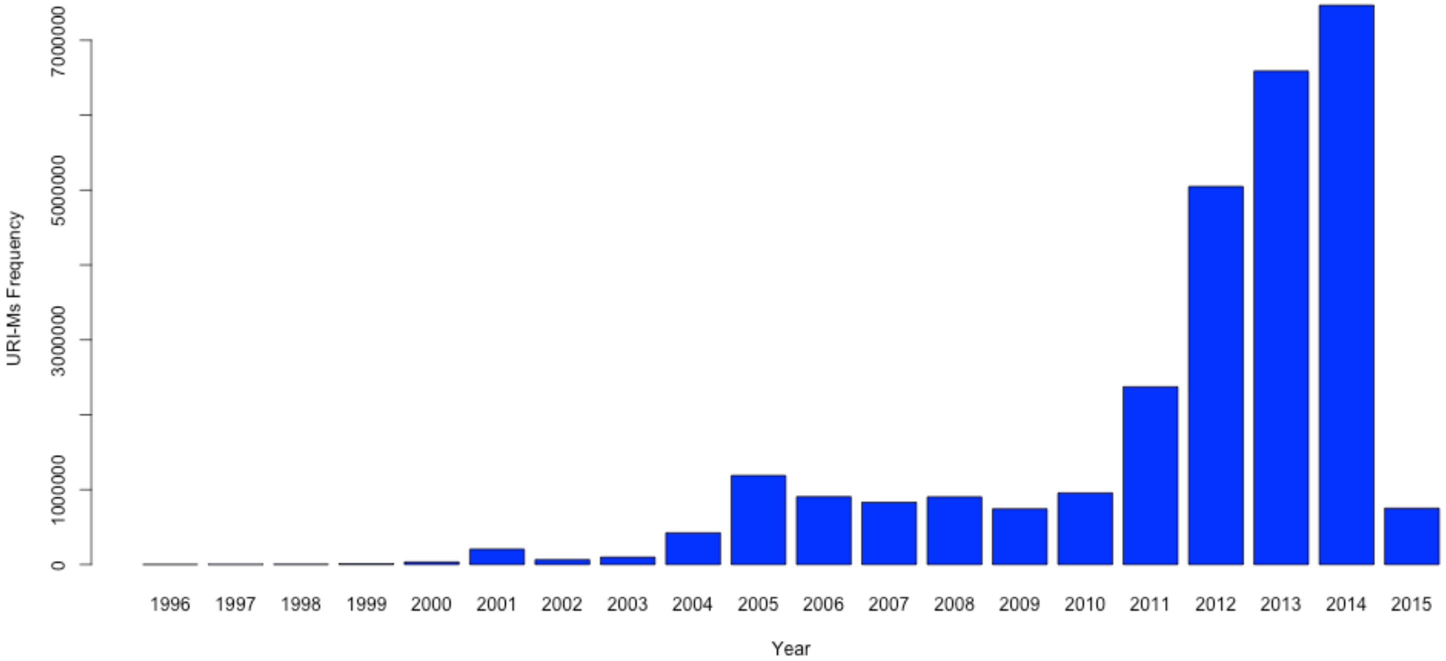
6 out of the 10 most archived are news websites

URI-Rs	Mementos	Category
gulfup.com	10,987	File Sharing
masrawy.com	9,144	Egyptian portal
arabic.cnn.com	9,022	News
aljazeera.net	8,906	News
maktoob.yahoo.com	8,478	Search Engine
shorooknews.com	7,548	News
arabnews.com	6,274	News
bbc.co.uk/arabic	6,268	News
ahram.org.eg	5,347	News
google.com.sa	4,968	Search Engine

Many are Arabic versions of globally popular sites

URI-Rs	Mementos	Category
gulfup.com	10,987	File Sharing
masrawy.com	9,144	Egyptian portal
arabic.cnn.com	9,022	News
aljazeera.net	8,906	News
maktoob.yahoo.com	8,478	Search Engine
shorooknews.com	7,548	News
arabnews.com	6,274	News
bbc.co.uk/arabic	6,268	News
ahram.org.eg	5,347	News
google.com.sa	4,968	Search Engine

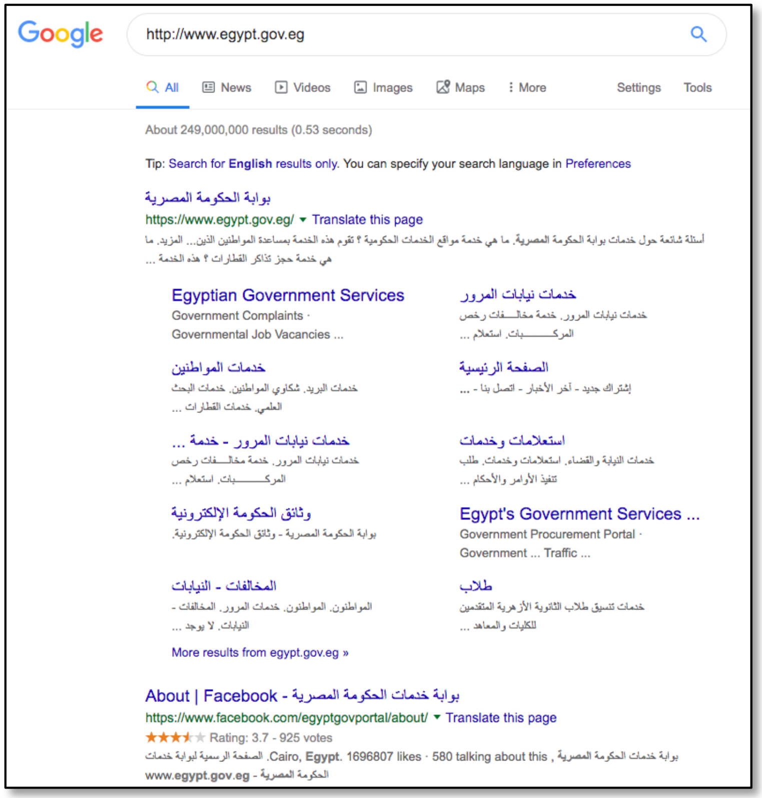
Most mementos are from recent years



Analysis done in 2015

We looked at indexing of Arabic seed URIs

- Used Google Custom Search API to query
  - limited to 1000 queries/day
  - tested only seeds



http://www.egypt.gov.eg

About 249,000,000 results (0.53 seconds)

Tip: Search for **English** results only. You can specify your search language in Preferences

بوابة الحكومة المصرية
https://www.egypt.gov.eg/ ▾ Translate this page
أسئلة شائعة حول خدمات بوابة الحكومة المصرية. ما هي خدمة مواقع الخدمات الحكومية ؟ تقوم هذه الخدمة بمساعدة المواطنين الذين... المزيد. ما
هي خدمة حجز تذاكر القطارات ؟ هذه الخدمة ...

Egyptian Government Services
Government Complaints ·
Governmental Job Vacancies ...

خدمات نيابات المرور
خدمات نيابات المرور. خدمة مخالـــفات رخص
المركـــــــــبات. استعلام ...

خدمات المواطنين
خدمات البريد. شكاوي المواطنين. خدمات البحث
العلمي. خدمات القطارات ...

الصفحة الرئيسية
إشتراك جديد - آخر الأخبار - اتصل بنا - ...

خدمات نيابات المرور - خدمة ...
خدمات نيابات المرور. خدمة مخالـــفات رخص
المركـــــــــبات. استعلام ...

استعلامات وخدمات
خدمات النيابة والقضاء. استعلامات وخدمات. طلب
تنفيذ الأوامر والأحكام ...

وثائق الحكومة الإلكترونية
بوابة الحكومة المصرية - وثائق الحكومة الإلكترونية.

Egypt's Government Services ...
Government Procurement Portal ·
Government ... Traffic ...

المخالفات - النيابات
المواطنون. المواطنون. خدمات المرور. المخالفات -
النيابات. لا يوجد ...

طلاب
خدمات تنسيق طلاب الثانوية الأزهرية المتقدمين
للكليات والمعاهد ...

More results from egypt.gov.eg »

About | Facebook - بوابة خدمات الحكومة المصرية
https://www.facebook.com/egyptgovportal/about/ ▾ Translate this page
★★★★☆ Rating: 3.7 - 925 votes
بوابة خدمات الحكومة المصرية , 580 talking about this · 1696807 likes .Cairo, **Egypt**. الصفحة الرسمية لبوابة خدمات
www.egypt.gov.eg - الحكومة المصرية

69% of Arabic language seed URIs were indexed by Google

Arabic Seed Dataset	
(Live, Indexed, Archived)	Percent
(1, 1, 1)	43.34%
(1, 1, 0)	25.59%
(1, 0, 1)	15.27%
(1, 0, 0)	15.76%

All seeds were live

- 82% of those listed in DMOZ were indexed
- 74.6% of the top-level seeds (path depth = 0) were indexed

58% of seeds were archived

Arabic Seed Dataset	
(Live, Indexed, Archived)	**Percent**
(1, 1, 1)	43.34%
(1, 1, 0)	25.59%
(1, 0, 1)	15.27%
(1, 0, 0)	15.76%

But, 42% were not archived

43% were indexed and archived

Good!
(discovered
and saved)

Arabic Seed Dataset	
(Live, Indexed, Archived)	Percent
(1, 1, 1)	43.34%
(1, 1, 0)	25.59%
(1, 0, 1)	15.27%
(1, 0, 0)	15.76%

31% were not indexed by Google

Arabic Seed Dataset	
(Live, Indexed, Archived)	Percent
(1, 1, 1)	43.34%
(1, 1, 0)	25.59%
(1, 0, 1)	15.27%
(1, 0, 0)	15.76%

16% were neither indexed nor archived

Arabic Seed Dataset

(Live, Indexed, Archived)	Percent
(1, 1, 1)	43.34%
(1, 1, 0)	25.59%
(1, 0, 1)	15.27%
(1, 0, 0)	15.76%

Bad!
(undiscovered
and not saved)

How does this compare to other languages?

We chose English, Danish, and Korean

- English
  - most popular language on the Internet
  - over 65 countries have English as an official language
- Danish
  - European language
  - 96% of population in Denmark uses the Internet
  - Denmark has government initiative to archive Danish cultural heritage on the Web
- Korean
  - Asian language
  - 92% of population in South Korea uses the Internet (highest in Asian countries)

We gathered seed URIs from DMOZ

- English
  - sample of size 10,000
- Danish
  - sample of size 10,000
- Korean
  - all 2,347 URIs in DMOZ

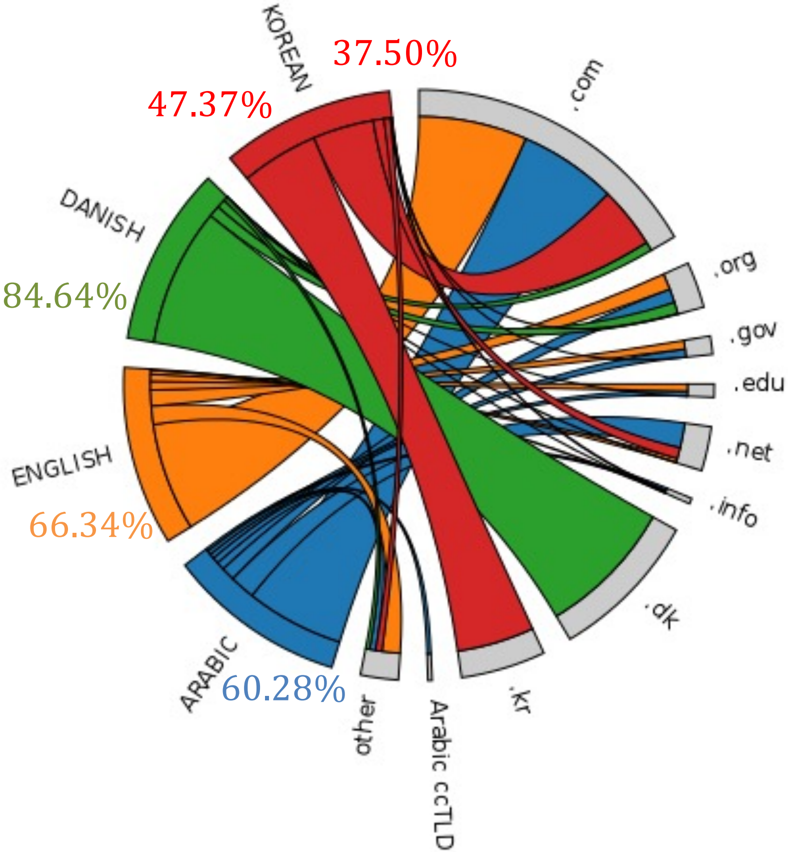
Crawled the seed URIs to expand the dataset

	Arabic	English	Danish	Korean
Seeds	15,742	10,000	10,000	2,347
Live	11,014	9,384	9,245	2,070
Language	7,976	8,576	6,331	1,157
Crawled	663,443	224,249	174,369	16,016
Live	482,905	176,261	131,484	11,099
Language	292,670	137,950	99,019	7,965
Total	**300,646**	**146,526**	**105,350**	**9,482**

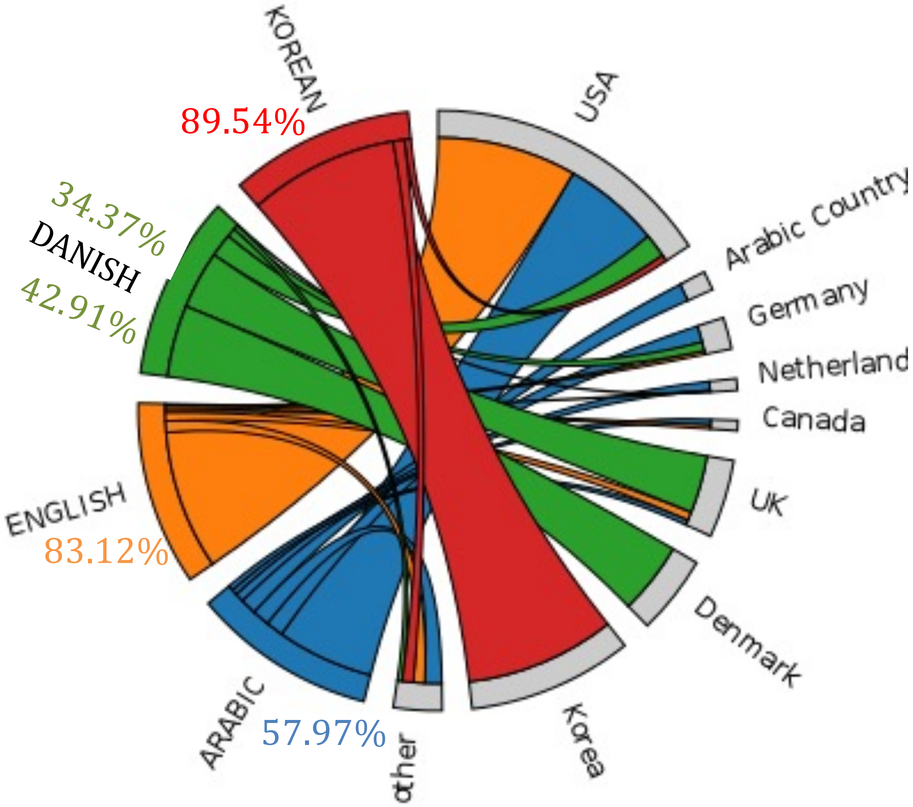
562,004 total

December 2015 - March 2016

English and Arabic web pages were .com, Danish and Korean were ccTLD



English and Arabic web pages were located in the US, Danish and Korean were in their countries



Arabic is archived less than English, but more than Danish and Korean

	Arabic	English	Danish	Korean
Seeds and Crawled	300,646	146,526	105,350	9,482
Archived	161,678	107,398	41,703	3,972
Percent	53.77%	73.30%	* 39.59%	41.89%

* Danish government archive is dark (not publicly available)

>90% of pages listed in DMOZ were archived, regardless of language

	Arabic	English	Danish	Korean
DMOZ Seeds	2,904	8,576	6,331	1,157
Archived	2,774	8,014	6,164	1,358
Percent	95.52%	93.44%	97.36%	89.52%

* DMOZ closed in 2017, but curlie.org began in 2018 with DMOZ data

There are a few caveats

- DMOZ is no longer operational
- Some web pages are multilingual
- Not easy to characterize “the Arabic Web”
  - 67% of Arabic dataset had neither Arabic ccTLD nor Arabic geo-location
- The language of a web page may shift over time

How well are Arabic language web pages archived?

Only about half of Arabic language web pages are archived

- Analyzed over 500,000 web pages (2014-2016)
- Arabic is archived less than English, but more than Danish and Korean
- Most Arabic and English pages are geo-located in the US, Danish in Denmark, and Korean in South Korea
- Web pages present in DMOZ are highly likely to be archived, regardless of language