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

LULWAH M. ALKWAI[1], MICHAEL L. NELSON and MICHELE C. WEIGLE, Old Dominion University

It is has long been suspected that web archives and search engines favor Western and English language Web pages. In this paper we quantitatively explore how well indexed and archived Arabic language Web pages are as compared to those from other languages. We began by sampling 15,092 unique URIs from three different website directories: DMOZ (multi-lingual), Raddadi, and Star28 (both primarily Arabic language). Using language identification tools, we eliminated pages not in the Arabic language (e.g., English language versions of Aljazeera pages) and culled the collection to 7,976 Arabic language Web pages. We then used these 7,976 pages and crawled the live Web and web archives to produce a collection of 300,646 Arabic language pages. We compared the analysis of Arabic language pages with that of English, Danish, and Korean language pages. First, for each language we sampled unique URIs from DMOZ, then using language identification tools we kept only pages in the desired language. Finally, we crawled the archived and live Web to collect a larger sample of pages in English, Danish, or Korean. In total for the four languages, we analyzed over 500,000 Web pages. We discovered: 1) English has a higher archiving rate than Arabic, with 72.04% archived. However, Arabic has a higher archiving rate than Danish and Korean, with 53.36% of Arabic URIs archived, followed by Danish and Korean with 35.89% and 32.81% archived respectively. 2) Most Arabic and English language pages are located in the US; only 14.84% of the Arabic URIs had an Arabic country code top-level domain (e.g., .sa) and only 10.53% had a GeoIP in an Arabic country. Most Danish language pages were located in Denmark, and most Korean language pages were located in South Korea. 3) The presence of a Web page in a directory positively impacts indexing, and presence in the DMOZ directory, specifically, positively impacts archiving in all four languages. In this work, we show that web archives and search engines favor English pages. However, it is not universally true for all Western language Web pages, because in this work we show that Arabic Web pages have a higher archival rate than Danish language Web pages.

CCS Concepts: •**Applied computing** → **Digital libraries and archives;**

Additional Key Words and Phrases: Web Archiving, Indexing, Digital Preservation, Arabic Web, English Web, Danish Web, Korean Web

---

Author's addresses: Lulwah M. Alkwai, Michael L. Nelson and Michele C. Weigle, Department of Computer Science, Old Dominion University, Norfolk, Virginia 23529 USA.

DOI: http://dx.doi.org/10.1145/0000000.0000000

[1]Department of Computer Science and Software Engineering, University of Hail, Hail, Saudi Arabia

1. INTRODUCTION

Arabic is the fourth most popular language on the Internet, trailing only English, Chinese, and Spanish [Internet World Stats 2015c]. Over the past few years, the number of Arabic-speaking Internet users has grown rapidly. In 2009, only 17% of Arabic speakers used the Internet [Internet World Stats 2009], but by the end of 2015 that had increased to almost 41% (over 135 million), approaching the world average of 45% of the population using the Internet [Internet World Stats 2015a]. In 2010, the size of the indexed Arabic Web was estimated to be about 2 billion pages [Alarifi et al. 2012]. It is not unreasonable to assume that Arabic online content is even larger today.

The Web is quickly becoming a repository for our cultural heritage, but studies have shown that the lifetime of Web pages is short (44-100 days) [Kahle 1997; Brewington and Cybenko 2000] and that resources are disappearing from the live Web [SalahEldeen and Nelson 2013b; Klein et al. 2014]. Thus, Web pages need to be preserved for future cultural and historical data mining. Web archiving is becoming recognized as an important problem [Lepore 2015], and several institutions, most notably the Internet Archive, have created archives to preserve Web pages. There are even several country and language specific archives[2], such as the BnF web archives (.fr domain)[3], the National Archives of the UK government (.uk domain)[4], and the Icelandic Web Archive (.is domain)[5].

Egypt has created the Bibliotheca Alexandrina Web Archive[6], but as we show later, it does not currently contain a large set of Arabic content that is not available elsewhere. This lack of focused archiving of the Arabic Web motivates our study of how well Arabic language Web pages are being archived today. To investigate this, we obtained a sample of URIs from Arabic Web directories. For those Web pages that we determined were written in Arabic, we studied several characteristics, including GeoIP location, country code top-level domain (ccTLD), URI path depth, estimated creation date, how well the page was archived, and if the page was indexed in Google. To increase the size of our dataset, we also crawled the Arabic Web pages to collect more URIs to investigate.

In our preliminary work [Alkwai et al. 2015], and extended in this paper, we found that 46% of the Arabic URIs in our collection are not archived and 31% are not indexed by Google. Further we found that a large majority of Web pages with Arabic language content use generic top-level domains (TLDs) (especially .com) and are physically located in Western countries (with over half in the US). As expected, we found that URIs with higher path depth are less likely to be archived and indexed than URIs closer to the top-level site. In addition, we found that the presence in a directory positively impacts indexing, and presence in the DMOZ directory, specifically, positively impacts archiving.

In this work we compared the Arabic language Web page analysis with that of English, Danish, and Korean language Web pages. We used the same methods for the analysis, using updated infrastructure where possible. We obtained a sample of URIs for each language from DMOZ. We determined the language of the page, crawled the English, Danish, and Korean pages to enlarge our sample, and determined language of the resulting crawled set. For those Web pages that we determined were written in each of the languages, we studied several characteristics, including GeoIP location, ccTLD, URI path depth, estimated creation date, and if the page was archived.

[2]A list of prominent web archives is available at http://netpreserve.org/resources/member-archives
[3]http://www.bnf.fr/en/professionals/digital_legal_deposit.html
[4]http://www.nationalarchives.gov.uk
[5]http://vefsafn.is
[6]http://archive.bibalex.org

Table I: Countries with Arabic as the official language, their population, percentage of those who are Internet users, and ccTLD. Source: [Internet World Stats 2015a]

Country	Population (2015)	% are Internet Users	ccTLD	Note
Egypt	86,895,099	49.6%	.eg	
Algeria	39,542,166	18.1%	.dz	
Sudan	35,482,233	22.7%	.sd	
Morocco	32,987,206	56.0%	.ma	Co-official language, along with Berber
Iraq	32,585,692	9.2%	.iq	Co-official language, along with Kurdish
Saudi Arabia	27,345,986	60.5%	.sa	
Yemen	26,052,966	20.0%	.ye	
Syria	22,597,531	26.2%	.sy	
South Sudan	11,562,695	0%	.ss	
Tunisia	10,937,521	43.8%	.tn	
Somalia	10,428,043	1.5%	.so	Co-official language, along with Somali
United Arab Emirates	9,206,000	88.0%	.ae	
Jordan	6,528,061	44.2%	.jo	
Libya	6,244,174	16.5%	.ly	
Lebanon	4,136,895	70.5%	.lb	
Mauritania	3,516,806	6.2%	.mr	
Oman	3,219,775	66.4%	.om	
Kuwait	2,742,711	75.5%	.kw	
Palestine	2,731,052	55.4%	.ps	
Qatar	2,123,160	85.3%	.qa	
Bahrain	1,314,089	90.0%	.bh	
Djibouti	810,179	9.5%	.dj	Co-official language, along with French
Comoros	766,865	6.5%	.km	Co-official language, along with French and Comorian

As expected, we found that English has a higher archiving rate than Arabic with 73.30% compared to 53%, then followed by Danish and Korean, with 39.59% and 41.89% archived, respectively. Further we found that a large majority of Web pages with both English and Arabic language content use generic TLDs (especially .com) and are physically located in Western countries (with over half in the US). This is in contrast to Danish and Korean language Web pages, where the majority are located in Denmark and South Korea, with .dk and .kr as the most common TLDs, respectively. As with Arabic Web pages, for English, Danish, and Korean, we found that those Web pages listed in the DMOZ directory were more likely to be archived than those not found in DMOZ.

2. RELATED WORK

There has been previous work on the coverage of web archives, including a study of international bias in archiving and studies of national domains. Little, though, has been done specifically in terms of Arabic language content.

In 2010, Ainsworth et al. [Ainsworth et al. 2011] investigated how much of the Web was archived. They collected a sample of URIs from four different sources (DMOZ, Delicious, Bitly, and search engine indexes). The resulting archival percentages ranged from 16% to 79%. A follow-on study in 2013 [AlSum 2014] showed that the archival percentage range had increased to from 33% to 95%. However, these studies were not focused on content from specific countries or content in specific languages.

Thelwall and Vaughn [Thelwall and Vaughan 2004] studied the coverage of archiving at the Internet Archive and focused on content from four different countries: China, Singapore, Taiwan, and the United States. They found large national differences in the

archive coverage of the Web. This work focused on content location rather than content language and TLD.

Baeza-Yates et al. [Baeza-Yates et al. 2007] characterized national Web domains based on 120 million pages from 24 different countries. They found that some characteristics, such as URI path length and distribution of HTTP response codes (e.g., 200 OK, 404 Not Found, etc.), were similar across different country domains. Yet they noted that not all sites in a country use the ccTLD (e.g., .us is seldom used in the United States), so other methods for determining if a site belongs to a particular country may be required.

Gomes and Silva [Gomes and Silva 2005] studied the Portuguese Web, including websites related to Portugal or of interest to Portuguese people. They filtered sites based on domain .pt, but also acknowledged that some sites would use other TLDs (such as .com, .net, .org) and so also considered sites that had content in the Portuguese language.

In 2009, the Language Technologies Institute at Carnegie Mellon University created the ClueWeb09 dataset [Callan et al. 2009], to support research on information retrieval. ClueWeb09 consists of about 1 billion Web pages in ten languages including Arabic. The Arabic dataset, named ArClueWeb09, consists of over 29M records in WARC format. However, the collected 2009 Arabic Web crawl has become somewhat dated, and we believe that the Arabic Web has grown in the last several years. In Section 5.1.1, we estimated the creation dates of our Arabic dataset. We found that 73% of the archived Arabic Web pages in our dataset were created after 2009. We believe that our dataset thus better reflects today's Arabic Web. The use of the Web globally has changed greatly since 2009. For instance, Facebook and Twitter, two of the most popular sites, have about 7 times more users today [Facebook 2016; Twitter 2016] than they did in 2009 [Kaplan and Haenlein 2010; Kwak et al. 2010].

An Arabic Web crawl was released after our analysis, named ArabicWeb16 [Suwaileh et al. 2016]. This is a public Web crawl of roughly 150M Arabic Web pages, which could be used for future Arabic Web analysis.

A recent investigation into the unarchived Web [Huurdeman et al. 2014] on the Dutch Web archive has shown that the archived Web can be a rich source of links to potentially unarchived content. In our work, we crawl archived pages to increase the size and variety of our dataset.

To further discuss Web archiving, we must introduce terminology from the Memento framework. Memento [Van de Sompel et al. 2013; Van de Sompel et al. 2009] is an HTTP protocol extension which links information from multiple web archives. We can use Memento to obtain a list of archived versions of resources, or mementos, from various different archives. In this paper, we use the following Memento terminology:

— URI-R - the original resource as it used to appear on the live Web. A URI-R may have 0 or more mementos (URI-Ms).
— URI-M - an archived snapshot of the URI-R at a specific date and time, which is called the Memento-Datetime, e.g., $URI\text{-}M_i = URI\text{-}R@t_i$.
— TimeMap - a resource that provides a list of mementos (URI-Ms) for a URI-R, ordered by their Memento-Datetimes.

3. EXPERIMENTAL SETUP

This section describes our experimental setup: selecting Arabic seed URIs, determining language, and crawling Arabic seed URIs.

3.1. Selecting Arabic Seed URIs

First, we searched for Arabic website directories and took the top three based on Alexa ranking[7]. Between March-May 2014, we collected all URIs from these three Arabic website directories: 1) the Arabic DMOZ listing, registered in the US in 1999, 2) Raddadi, a well-known Arabic directory, registered in Saudi Arabia in 2000, and 3) Star28, an Arabic directory, registered in Lebanon in 2004. Table II shows the number of collected URIs from these three sources. From the 15,743 URIs collected, we found 15,092 unique URIs (only 651 were duplicates). Using cs.odu.edu machines we tested the existence of each seed URI on the live Web and found 11,014 that returned HTTP 200 OK status code (some after redirection). We downloaded the contents of each page that was found on the live Web.

Table II: Arabic seed source count

Name	URI	Initial seed URIs
DMOZ	dmoz.org/World/Arabic/	4,086
Raddadi	raddadi.com	3,271
Star28	star28.com	8,386
Total		15,743

3.2. Determining Language

Table I, sorted by population, lists each country where Arabic is an official language, its population, the percentage of its population that are Internet users, its ccTLD, and if other languages are spoken. Although we gathered Web pages from Arabic language directories, it is likely that some of these were written in other languages. We were interested in further analyzing only pages written in Arabic, so we used several methods to determine the language of each of the 11,014 live Web seed URIs.

One of the challenges is to find a reliable language test to determine language. No test will result in 100% confidence, so in order to detect the language of a Web page, we tested four different methods. The language tests we performed were as follows:

— **HTTP Content-Language** - If the HTTP response header contained Content-Language:ar, where ar is the ISO 639-2 code[8] for Arabic, we considered the Web page to be written in Arabic.
— **HTML title tag** - The HTML title tag is often a good indicator of the language of a Web page's content [Noruzi 2007]. We extracted the title tag of each Web page and used the guess-language Python library [Johnson 2010] to determine the language.
— **Trigram method** - The trigram technique uses letter trigrams, sequences of three letters, to determine language [Beesley 1988]. The identification is performed through basic trigram lookups paired with unicode character set recognition. We extracted the text of each Web page using BeautifulSoup [Richardson 2013], and then we used the Python-Language-Detector tool [Graves 2012], which implements the trigram method on the extracted text and title.
— **Language detection API client** - The Language Detection API[9] is a Web service that detects 160 different languages. We ran the test on the extracted text and title from the HTML of each Web page using BeautifulSoup [Richardson 2013].

[7]http://www.alexa.com
[8]https://www.loc.gov/standards/iso639-2/php/code_list.php
[9]https://detectlanguage.com

The reliability of the tests to determine if a Web page is in Arabic was measured by having a native reader (the first author) quickly evaluate a sample of pages. Next, we measured the number of URIs reported as Arabic. The number of URIs where Content-Language was not available was 1,200, that is 10.90% of the dataset, and the number of blank page titles or titles that had undetermined encoding was 1,763, which is 16% of the dataset. Figure 1 shows the intersection between the four language tests. We found only 872 of the URIs tested as Arabic language in all four tests. We decided to consider the Web page part of the Arabic Web if it passes any one of the language tests. After running all of the tests on the 11,014 live Web pages, we found 7,976 that passed at least one of the language tests. We consider this set to be our Arabic seed URIs [Alkwai 2017a].

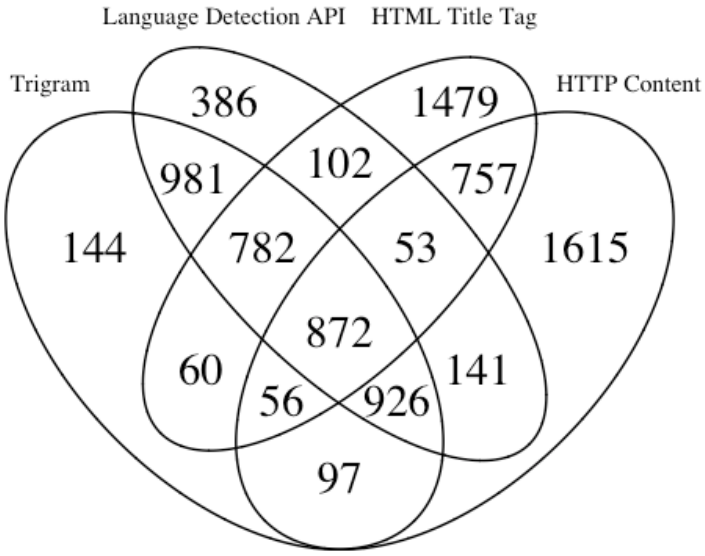


Fig. 1: Language test intersection testing for Arabic language on Arabic seed URIs

3.3. Crawling Arabic Seed URIs

To increase the size of our dataset, we crawled the Arabic seed URIs between January-March 2014. Our first pass was to gather additional URIs linked from the live Web versions of our seed URIs. We crawled the seed URIs two levels deep, meaning that we gathered all of the URIs linked to from each seed URI and then gathered all of the URIs linked to by the URIs in the first level crawl. This resulted in collecting 575,242 URIs, all of which were available on the live Web.

To gather even more URIs, we crawled the Arabic seed URIs that had at least one archived version, or memento. We crawled the most recent memento and gathered 515,821 URIs. Of these, only 335,283 were available on the live Web.

Combining the two sets (crawled live and crawled archived), we obtained a total of 663,443 unique URIs. We ran each of these through our Arabic language tests, resulting in 292,670 Arabic URIs obtained from crawling our Arabic seeds.

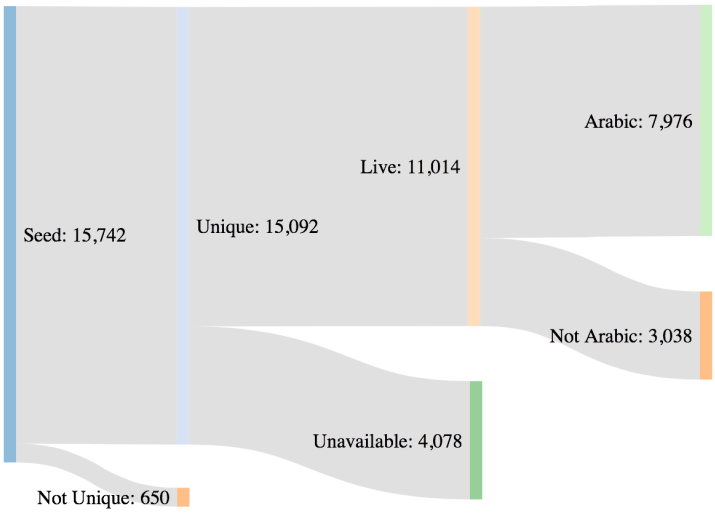


Fig. 2: Filtering Arabic seed URIs

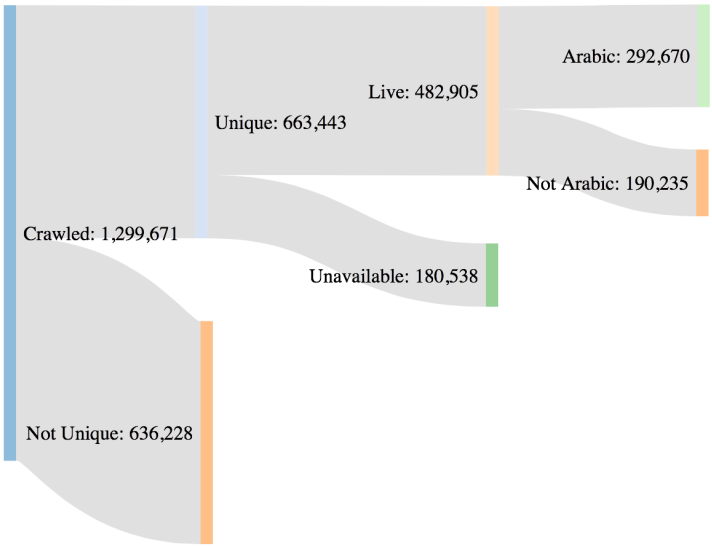


Fig. 3: Filtering Arabic crawled URIs

Figures 2 and 3 show the summary of collecting Arabic URIs for seed URIs and for crawled URIs, respectively. Combining the seed URIs and crawled URIs, we collected 300,646 Arabic URIs that we analyze in the remainder of the paper.

Table III: Most frequent domains in the Arabic dataset

Rank	Domain	URIs	GeoIP Location
1	alarab.net	284	US
2	aljarida.com	248	US
3	arabic.cnn.com	245	US
4	alarabiya.net	231	US
5	ar.wikipedia.org	230	US
6	aljazeera.net	213	US
7	moheet.com	142	US
8	facebook.com	133	US
9	al-sharq.com	132	US
10	lakii.com	123	US
17	kuwaitclub.com.kw	71	Kuwait

4. ARABIC DATASET ANALYSIS

In this section we examine the characteristics of our Arabic URI dataset. We investigate the number of unique domains, TLD and ccTLD, GeoIP location, and URI path depth. For the original Arabic seed URI dataset, we also investigate the GeoIP location.

4.1. Unique Domains

First, we investigate the number of unique domains in our dataset. Out of the 300,646 Arabic URIs, there are 17,536 unique domains. The most frequent domains are shown in Table III. We also tested the GeoIP location of the top-level Web page of each of these domains and found that the top 16 are all located in the US. The first domain we find located in an Arabic country is the 17th most frequent. We note that several of these top domains are popular Western sites, such as cnn.com and wikipedia.org. This indicates that the Arabic language community is already using services on Western sites that are likely to be archived.

4.2. Top Level Domains

We investigate the top-level domain (TLD) and country code TLD (ccTLD), together termed effective TLD, of the unique Arabic language domains. Generic TLDs such as .com, .net, and .org are open for any registrant. In addition to TLDs, many sites also use the two-letter ccTLD of their home country. Although a small percentage of the websites add the ccTLD, it may be a good indication of the source of the website. Table IV shows the distribution of the top 10 effective TLDs. Almost 58% of all URIs have a .com TLD, which is not unexpected since .com is a popular TLD and has an open registration policy.

Table V shows the top 5 ccTLDs from Arabic-speaking countries (listed in Table I). We found that Saudi Arabia was the most frequent Arabic ccTLD, followed by Egypt and Jordan. We note that the top Arabic ccTLD, .sa for Saudi Arabia, is used in fewer URIs than the generic TLDs .com, .net, and .org.

4.3. GeoIP Location

Here we want to look at the GeoIP location of the Arabic URI dataset. First, we obtained the IP addresses of the hostnames using `nslookup`, which uses DNS to convert the hostname to its IP address. Then we used the MaxMind GeoLite2[10] database to

[10]http://dev.maxmind.com/geoip/geoip2/geolite2/

Table IV: Top 10 effective TLDs in the Arabic dataset

TLD	Count	Percent
com	174,284	57.97%
net	45,307	15.07%
org	19,241	6.40%
gov.sa	5,832	1.94%
info	5,050	1.68%
edu.sa	3,818	1.27%
ws	3,487	1.16%
org.sa	2,916	0.97%
com.sa	2,405	0.80%
gov.eg	2,405	0.80%
other	35,897	11.94%

Table V: Top 5 Arabic ccTLDs

ccTLD	Country	Count	Percent
.sa	Saudi Arabia	16,024	5.33%
.eg	Egypt	6,012	2.00%
.jo	Jordan	6,012	2.00%
.ae	United Arab Emirates	3,186	1.06%
.kw	Kuwait	2,465	0.82%

determine location from the IP address, which tests at 99.8% accuracy at the country level[11].

We found that less than 11% of the URIs are hosted in Arabic countries. Table VI shows the top GeoIP locations, with Arabic countries grouped together. Table VIII shows the top 5 GeoIP locations from Arabic countries. Overall, almost 58% of the Arabic URIs are hosted at IP addresses in the US. Other Western countries, including Germany and the Netherlands, host more of the Arabic URIs than does Saudi Arabia, the highest contributor of the Arabic countries.

Table IX shows the number and percentage of Arabic seed URIs and total Arabic URIs that have Arabic ccTLD only, Arabic GeoIP only, both Arabic ccTLD and GeoIP, or neither. We find that a large percentage of Arabic URIs have neither Arabic ccTLD nor Arabic GeoIP, with 84.99% of the Arabic seeds, and 66.82% of the total. Only 7.81% of the total dataset has both Arabic ccTLD and GeoIP.

To investigate if the source of our seed collection affected our GeoIP results, we considered GeoIP location per source (Table VII). The percentage of US-based URIs is similar over all three of the sources, even though DMOZ is registered in the US, and both Raddadi and Star28 are registered in Arabic countries.

4.4. URI Path Depth

URI path depth is an important factor in archiving, as we assume that Web pages nearer to the top-level of a site will be better archived than pages deeper into the site (i.e., with higher path depth). Table X shows the breakdown of URI path depth for our Arabic URIs. Over half of the URIs have a path depth of 0 or 1, with barely 7% having a path depth greater than 3.

[11]http://dev.maxmind.com/faq/how-accurate-are-the-geoip-databases/

Table VI: Top GeoIP locations of Arabic URIs

Country	Count	Percent
US	174,284	57.97%
Arabic countries	31,658	10.53%
Germany	29,312	9.75%
Netherlands	15,904	5.29%
France	13,138	4.37%
Canada	9,951	3.31%
UK	9,229	3.07%
Others	17,166	5.71%

Table VII: Percent of Arabic seed URIs with GeoIP in US based on seed source

Name	US GeoIP Percent
DMOZ	55.41%
Raddadi	57.38%
Star28	55.49%

Table VIII: Top 5 Arabic GeoIP locations

Country	Count	Percent
Saudi Arabia	14,280	4.75%
Egypt	5,922	1.97%
Jordan	4,269	1.42%
Kuwait	2,134	0.71%
UAE	2,014	0.67%

Table IX: ccTLD and GeoIP of seed and total Arabic dataset

	Seed URIs	Percent	Total URIs	Percent
Arabic ccTLD	527	6.61%	44,609	14.84%
Arabic GeoIP	189	2.37%	31,671	10.53%
Arabic GeoIP and ccTLD	481	6.03%	23,479	7.81%
Neither	6,779	84.99%	200,887	66.82%

Table X: Path depth of the Arabic URIs

Path Depth	Example	Count	Percent
0	example.com	52,011	17.30%
1	example.com/a	121,521	40.42%
2	example.com/a/b	73,507	24.45%
3	example.com/a/b/c	32,499	10.81%
4+	example.com/a/b/c/d	21,108	7.02%

5. ARCHIVING AND INDEXING OF ARABIC WEB PAGES

In this section, we analyze the presence of the Arabic URIs in web archives and in search engine indexes. We consider archiving based on creation date, seed URI source, ccTLD, GeoIP location, and URI path depth. We look at indexing based on seed URI source, ccTLD, GeoIP location, and URI path depth.

5.1. Presence in the Archive

Between January-March 2015, we used the Memento Framework, through the ODU CS Memento Aggregator (mementoproxy.cs.odu.edu), to determine if the URIs in our Arabic dataset are archived[12]. For each URI, the aggregator returns a TimeMap that lists the number of mementos that exist in various archives. Overall, we found that 161,678 URIs (53.77% of our Arabic URIs) are archived (i.e., have one or more mementos). Figure 4 shows the number of mementos found for each archived URI, sorted by memento count, with a median of 16 mementos per URI-R. Table XI lists the top 10 archived URI-Rs. As expected, most of these are news websites.

Figure 5 shows the number of URI-Ms with Memento-Datetimes in each year. This reveals an increasing rate of archiving in recent years.

Since the TimeMap identifies mementos present in multiple archives, we want to compare the archival coverage in different archives. We use two measurements defined by Alsum et al. [AlSum et al. 2014].

(1) $Coverage_r(R, WA)$ is equal to one if the URI-*R* R has at least one memento at the web archive *WA*. Otherwise, it is equal to zero.

$$Coverage_r(R, WA_x) = \frac{\sum_{i=1}^{N} Coverage_r(R_i, WA_x)}{N} \times 100\% \quad (1)$$

(2) $Coverage_m(R, WA)$ is the total number of mementos of URI-*R* R discovered in a web archive *WA*; if the URI appears in the web archive, we count the number of mementos, otherwise 0.

$$Coverage_m(R, WA_x) = \frac{\sum_{i=1}^{N} Coverage_m(R_i, WA_x)}{\sum (R_i, WA_x)} \times 100\% \quad (2)$$

For example, if we have three web archives WA_1, WA_2, and WA_3; and R_1 has two mementos in WA_1, one memento in WA_2 and zero mementos in WA_3; and R_2 has

Table XI: Top 10 archived Arabic URI-Rs

URI-Rs	Memento Count	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

[12]Note that in early 2015, the CS Memento Aggregator did not include the Bibliotheca Alexandrina archive, as the archive had not always been reachable

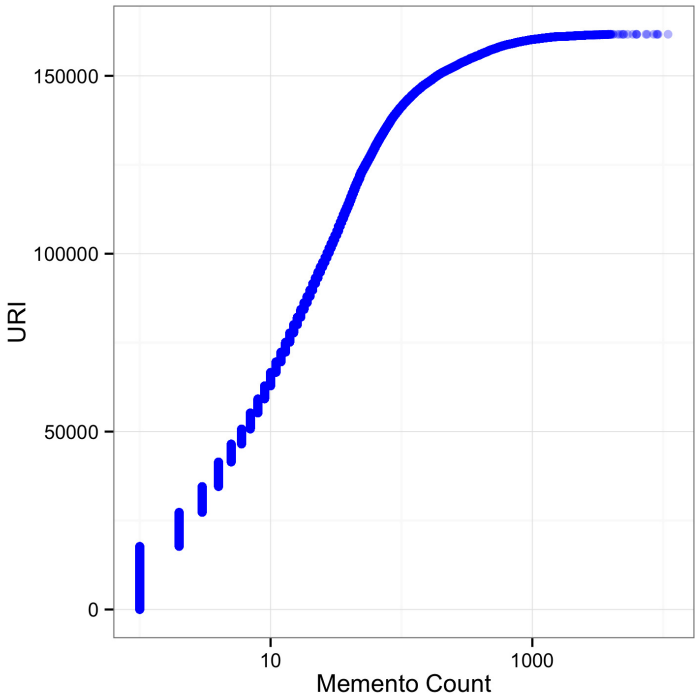


Fig. 4: Memento count frequency of Arabic URIs

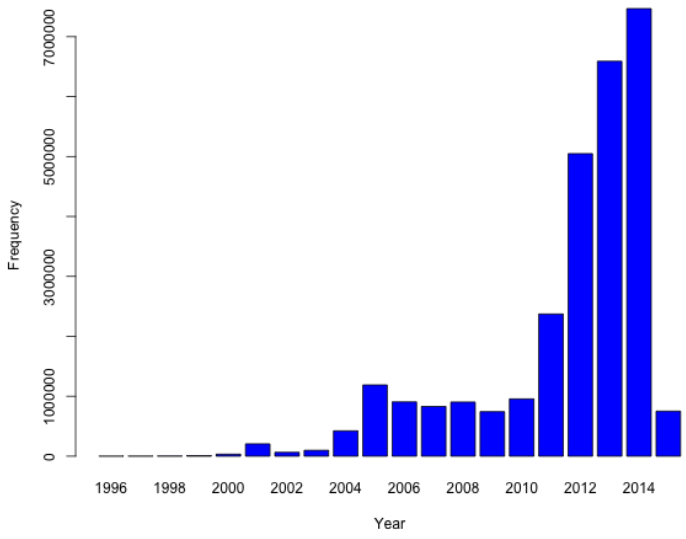


Fig. 5: Number of Arabic URI-Ms in each year

Table XII: $Coverage_r$ of archived Arabic URI-Rs present in each archive

Archive	Percent
Internet Archive	97.04%
Archive.today	6.58%
WebCite	6.00%
Archive-It	5.49%
British Library Archive	1.06%
UK Parliament Web Archive	0.88%
Icelandic Web Archive	0.87%
UK National Archives	0.62%
Proni	0.21%
Stanford	0.11%

three mementos in WA_1, zero mementos in WA_2 and zero mementos in WA_3; then we can compute the coverage measurements as follows:

$Coverage_r$:

— (R_1, WA_1) = 1, (R_1, WA_2) = 1, and (R_1, WA_3) = 0.
— (R_2, WA_1) = 1, (R_2, WA_2) = 0, and (R_2, WA_3) = 0.

then the total percentage of $Coverage_r$:

— $Coverage_r$ (WA_1) = (2/2)$\times$ 100% = 100%
— $Coverage_r$ (WA_2) = (1/2)$\times$ 100% = 50%
— $Coverage_r$ (WA_3) = (0/2)$\times$ 100% = 0%

$Coverage_m$:

— (R_1, WA_1) = 2, (R_1, WA_2) = 1, and (R_1, WA_3) = 0.
— (R_2, WA_1) = 3, (R_2, WA_2) = 0, and (R_2, WA_3) = 0.

then the total percentage of $Coverage_m$:

— $Coverage_m$ (WA_1) = (5/6)$\times$ 100% = 83.33%
— $Coverage_m$ (WA_2) = (1/6)$\times$ 100% = 16.66%
— $Coverage_m$ (WA_3) = (0/6)$\times$ 100% = 0%

Here we present the breakdown of archives holding URI-Rs in our Arabic dataset, $Coverage_r(R, WA)$. Table XII shows the percentage of archived URI-Rs that each archive holds. We found that the Internet Archive has the highest percentage by far, followed by Archive.today[13] and WebCite. We note that the percentages sum to greater than 100% because multiple archives can have mementos from the same original resource (URI-R). For instance, 97% of archived URI-Rs have at least one memento at the Internet Archive, and 6% of URI-Rs have at least one memento at WebCite.

Next, we want to know the breakdown of the archives for all URI-Ms in our dataset, $Coverage_m(R, WA)$. Table XIII shows the percentage of mementos that each archive holds. We found that almost 73% were in the Internet Archive and 21% were in Archive-It. To determine how well a URI is archived, we can look at the timespan of the mementos (number of days between the datetimes of the first memento and last memento), but that does not indicate how often the URI was archived. These could be two endpoints with no other mementos in between, or the URI could be regularly archived over the timespan. Here, we exclude URIs that have only one memento (16,732 URIs). We calculate the average archiving period by dividing the timespan by

[13]Archive.today is now called Archive.is

Table XIII: $Coverage_m$ of archived Arabic URI-Ms present in each archive

Archive	Percent
Internet Archive	72.87%
Archive-It	21.26%
Archive.today	2.14%
WebCite	2.08%
Icelandic Web Archive	1.17%
British Library Archive	0.29%
UK Parliament Web Archive	0.10%
Proni	0.05%
UK National Archives	0.04%
Stanford	<0.01%

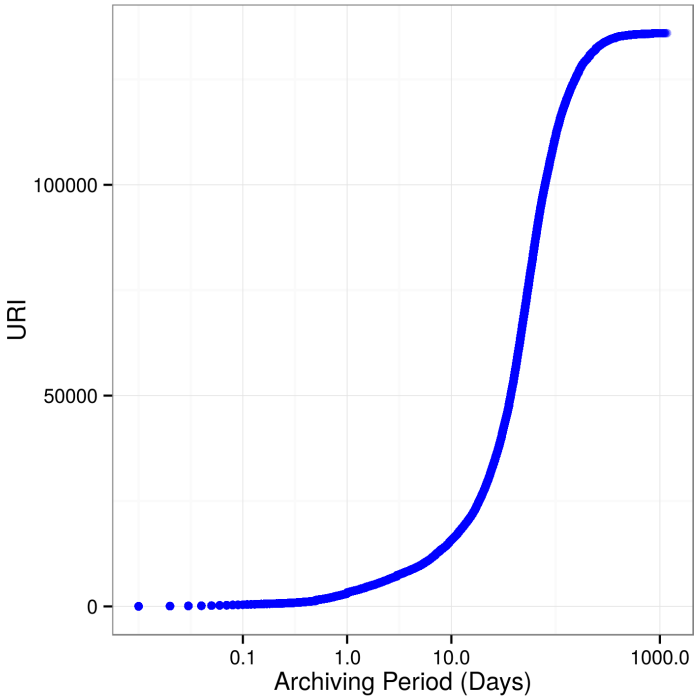


Fig. 6: Average archiving period (days) for each Arabic URI-R

the number of mementos for the URI. The smaller the period, the more regularly the URI was captured by the archives. In Figure 6, we show the average archiving period (in days) for each archived URI, where the URIs are sorted by archiving period, with a median of 1 memento every 48 days. Values less than 1 indicate that the URI is archived multiple times per day on average.

5.1.1. Creation Date.

Another interesting characteristic of a URI is its creation date. In terms of evaluating how well our Arabic URIs have been archived, we want to verify that we have URIs of various ages to ensure that they have been around long enough to be captured. For instance, if a Web page was created in 2000, we would expect to see several mementos

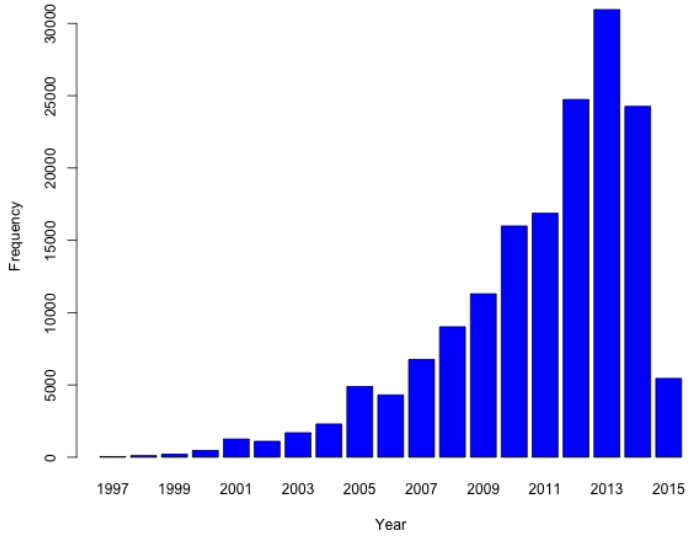


Fig. 7: Creation dates for archived Arabic URIs

in the archives. However, if the Web page was just created in 2015, we would not be surprised if it had not yet been archived or archived as much.

Usually we cannot definitively determine the creation date of a Web page, but there have been several methods proposed to estimate this. We use CarbonDate [SalahEldeen and Nelson 2013a], which looks to see when the URI was indexed in search engines, archived in public archives, and shared in social media. It then saves the oldest date found as the estimated creation date. We applied CarbonDate to our archived Arabic dataset. Figure 7 shows the frequency of estimated creation dates, with 2013 being the most frequent year. The figure also shows that our dataset contains a wide range of creation dates extending over the past 18 years.

5.1.2. Creation Date and First Memento.

Here we want to investigate the gap between the creation date of Arabic websites and when they were first archived. Figure 8 shows the URIs on the y-axis and the log of the delta (creation date - first memento) in days on the x-axis. We found that 19.48% of the URIs have an estimated creation date that is the same as first memento date and excluded those from the figure. For the remaining 130,184, almost 18% have creation dates over 1 year before the first memento was archived (solid vertical line).

5.1.3. Archiving Based on Seed URI Source.

Here we look at archiving based on seed URI source. As shown in Table XIV, we found that 96% of DMOZ seed URIs are archived and that 49% of those from Raddadi and 46% from Star28 are archived. This was expected because DMOZ URIs are more likely to be found and archived [Ainsworth et al. 2011; AlSum 2014]. DMOZ has historically been a source of seed URIs for indexing and archiving, at least as far back as 1999 [Cho 2001; Zerfos et al. 2005; Chakrabarti et al. 1999].

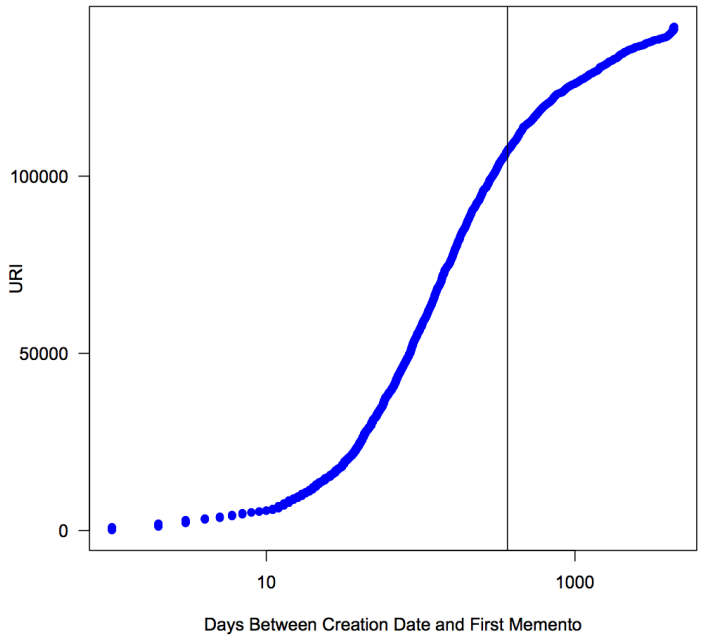


Fig. 8: Difference between creation date and first memento for archived Arabic URIs

Table XIV: Archiving and indexing based on Arabic seed source

Name	Arabic	Archived	Percent	Indexed	Percent
DMOZ	2,904	2,774	95.52%	2,385	82.13%
Raddadi	1,565	762	48.69%	1,104	70.54%
Star28	3,507	1,601	45.65%	2,514	71.68%
Total	7,976	5,137	64.40%	6,003	75.26%

5.1.4. Archiving Based on Location and ccTLD .

Based on our previous results, we want to look at how many archived URIs have an Arabic ccTLD, Arabic GeoIP, or both. Table XV shows the breakdown of the Arabic URIs that have both an Arabic ccTLD and an Arabic GeoIP, only an Arabic ccTLD, only an Arabic GeoIP, or neither Arabic ccTLD nor Arabic GeoIP. Only 33.18% of our set had evidence of location in an Arabic country (ccTLD or GeoIP), and these URIs were archived at a lower rate (34%) than URIs that had no evidence of location inside an Arabic country (65%). This finding goes with our intuition that sites hosted in Western countries would be more likely to be archived. Figure 9 shows the breakdown of GeoIP location, ccTLD, both, and neither GeoIP location nor ccTLD of the archived Arabic set.

Next we wanted to statistically analyze the archived Arabic dataset. Figure 10 shows the CDF of the Memento-Datetimes for the both Arabic ccTLD and Arabic GeoIP set. The CDFs for the other three sets (Arabic GeoIP, Arabic ccTLD, and neither), resulted in the same curve as observed visually. We found that half of mementos were archived after 2012, and less than 20% of the mementos were archived before 2010. To analyze

Table XV: Archiving based on location and ccTLD

	Total	Archived Count	Percent
Arabic ccTLD	44,609	12,532	28.09%
Arabic GeoIP	31,671	4,152	13.11%
Arabic GeoIP and ccTLD	23,479	13,969	59.50%
Neither	200,887	131,025	65.22%

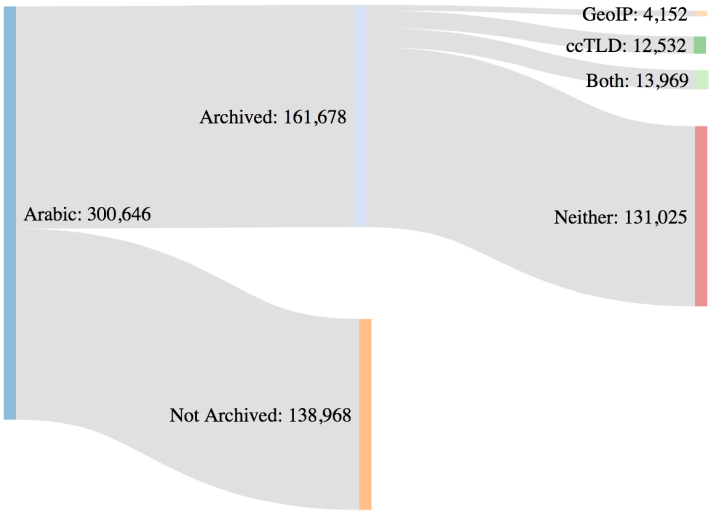


Fig. 9: Breakdown of Arabic GeoIP and ccTLD for Arabic URIs

these similarities further, we performed the Kolmogorov-Smirnov test to determine if the datasets are likely to be different. We compared the two sets Arabic GeoIP and Arabic ccTLD to the set with neither Arabic GeoIP nor Arabic TLD. We checked the p-value that gives us the probability of whether or not we can reject the null hypothesis, which is that two datasets have the same distribution. The D statistic is the absolute maximum distance between the CDFs of the two samples. The closer this number is to 0, the more likely it is that the two samples were drawn from the same distribution. The D value for comparing Arabic ccTLD and neither is 0.017 and for comparing Arabic GeoIP and neither it is 0.014. For both $p < 0.002$, meaning that the CDFs are statistically equivalent. Figure 11 shows the age of a URI (days since creation) vs. its number of mementos and the color of the dot represents both Arabic GeoIP and Arabic TLD, Arabic GeoIP only, Arabic TLD only, and neither. We found no correlation between location, ccTLD, and when a URI was archived. One might think that the older the resource, the more mementos it has. In the short term (less than 3 years), this is true (see Figure 12 for detail), but for URIs over 3 years old, this is not necessarily the case because of low historical archiving rates (as shown in Figure 5).

5.1.5. Archiving Based on URI Path Depth.

Next, we look at the effect of different URI path depths on archiving. As expected, we found that the shorter the URI path depth, the higher the rate of archiving. As shown

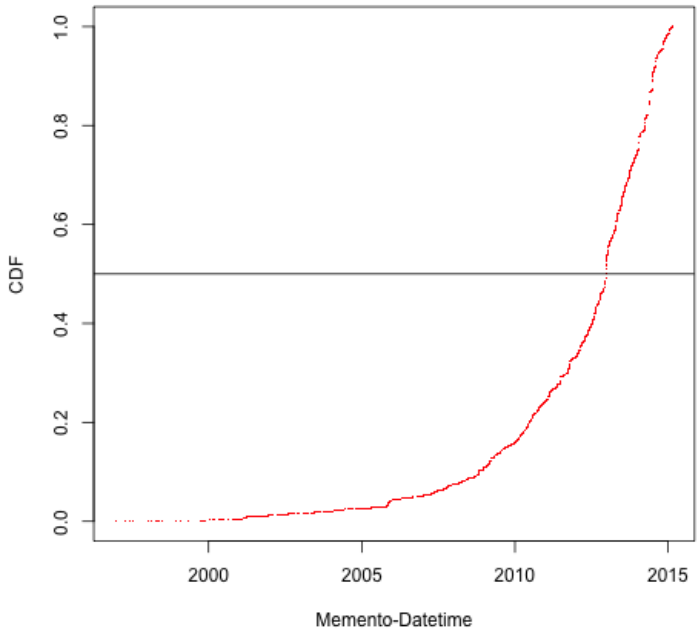


Fig. 10: CDF of Memento-Datetimes both Arabic GeoIP and Arabic ccTLD

Table XVI: Archiving of Arabic URIs based on URI path depth

Path Depth	Total	Archived Count	Percent
0	52,011	44,880	86.29%
1	121,521	65,001	53.49%
2	73,507	33,497	45.57%
3	32,499	11,585	35.65%
4+	21,108	6,715	31.82%

in Table XVI, we found that 86% of URIs with path depth 0 (i.e., top-level pages) were archived, with decreasing archiving rates as path depth increased. For those URIs with a path depth of greater than 3, only 32% were archived. This may be the result of the historical crawling strategy where search engines only sample sites instead of crawling the entire site, thus resulting in fewer deep links in a site being archived or indexed [Smith and Nelson 2008].

5.2. Search Engine Indexing

In addition to investigating if the Arabic URIs are archived, we are also interested to discover how well they are indexed in search engines such as Google. We used the Google Custom Search API to determine if the Arabic seed URIs are indexed by Google. We tested only the seed URIs because we were limited by the restriction of 1000 requests per day in the API. We found that only 36.2% of the Arabic seed URIs were indexed by Google. However, we note that the Google user Web interface may produce different results than the Custom Search API [McCown and Nelson 2007].

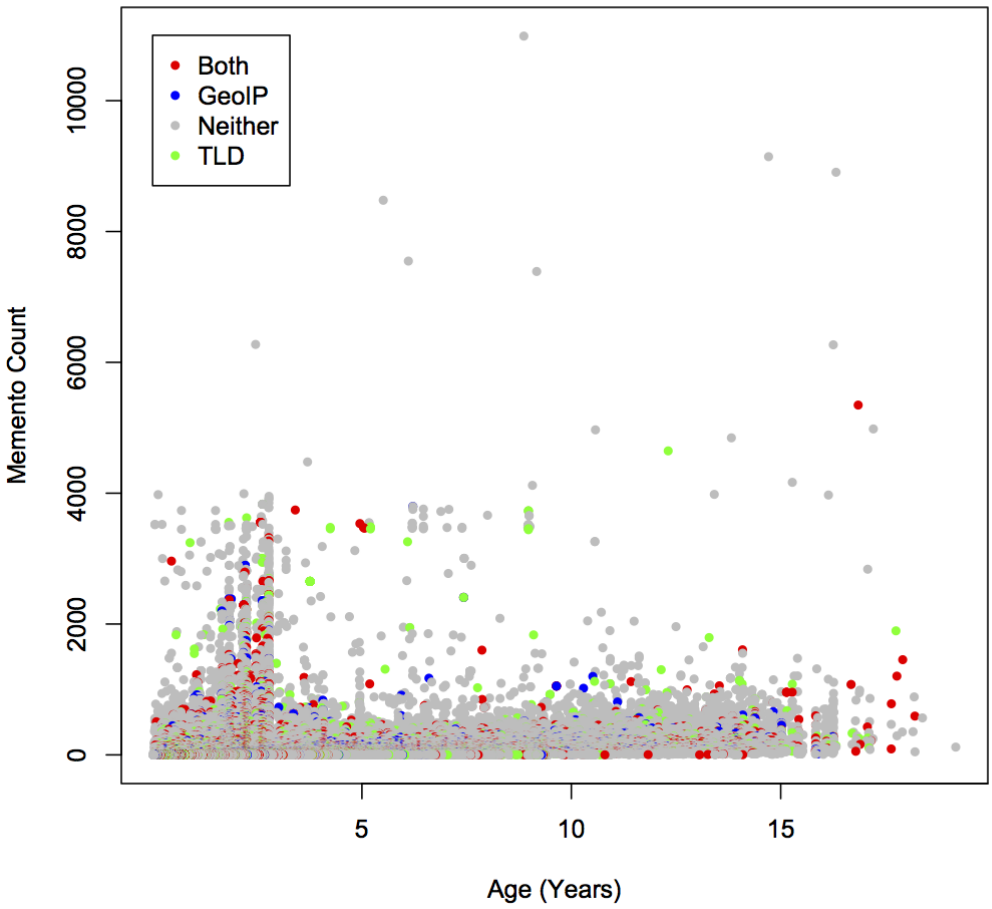


Fig. 11: URI age and memento count for archived Arabic URIs

For the Arabic seed URIs, we can indicate if they were present on the live Web, in the Google index, and present in an archive, creating a `(live, indexed, archived)` tuple. In Table XVII, we show the percentage of our Arabic seed URI dataset (7,976 URIs) that fell into each permutation of the tuple. We note that all of our Arabic seeds were present on the live Web at the time of our analysis. Almost 44% of the Arabic seed URIs were both indexed and archived, while only 15% were neither indexed nor archived.

5.2.1. Indexing Based on Seed URI Source.
Here we look at indexing based on seed URI source. As shown in Table XIV, we found that 82% of DMOZ seed URIs are indexed by Google and that 66% of those from Raddadi and 65% from Star28 are indexed. This was expected because DMOZ URIs are more likely to be found and indexed [Ainsworth et al. 2011; AlSum 2014], and many

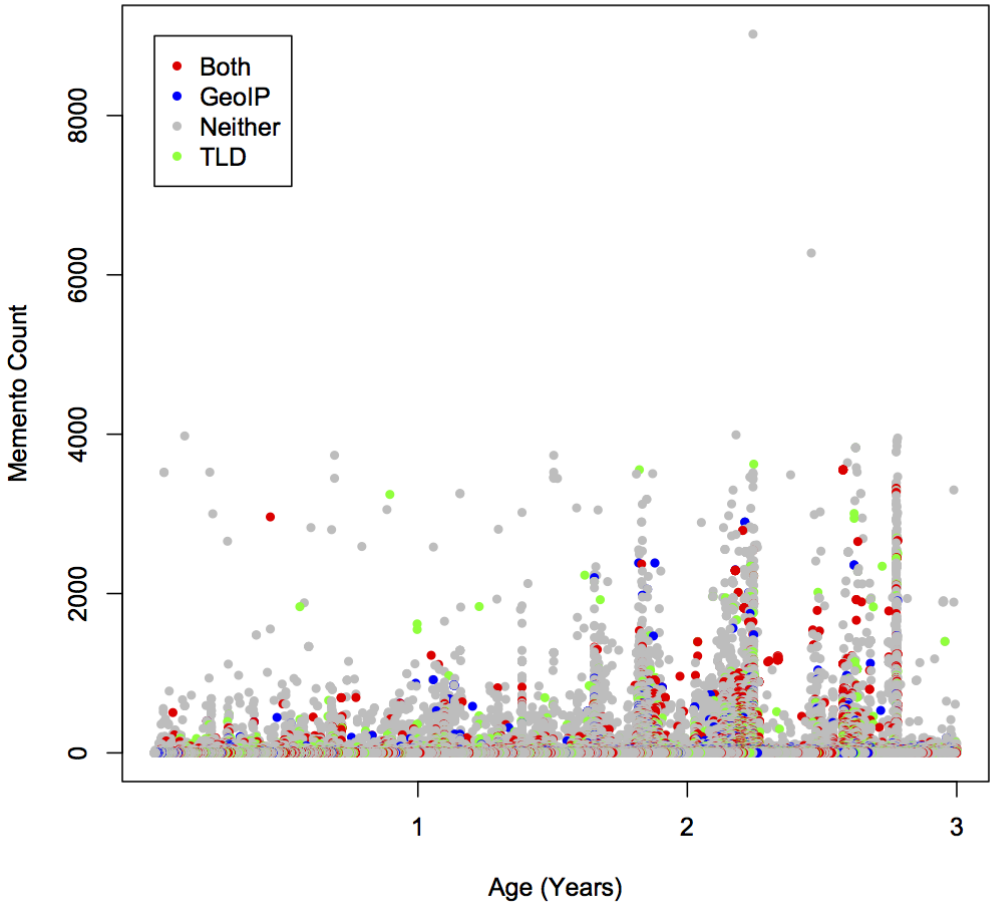


Fig. 12: URI age less than three years and memento count for archived Arabic URIs

Table XVII: Status of Arabic seed URIs

(Live, Indexed, Archived)	Count	Percent
(1, 1, 1)	3,457	43.34%
(1, 1, 0)	2,041	25.59%
(1, 0, 1)	1,218	15.27%
(1, 0, 0)	1,257	15.76%

crawling projects mine DMOZ for reference [Cho 2001; Zerfos et al. 2005; Chakrabarti et al. 1999].

Table XVIII: Indexing based on location and ccTLD of Arabic seed URIs

	Total	Indexed Count	Percent
Arabic GeoIP and ccTLD	481	410	85.24%
Arabic ccTLD only	527	401	76.09%
Arabic GeoIP only	189	139	73.54%
Neither	6,779	4,548	67.09%

Table XIX: Indexing based on URI path depth of Arabic seed URIs

Path Depth	Total	Indexed Count	Percent
0	6,863	5,120	74.60%
1	776	302	38.91%
2+	337	76	22.55%

5.2.2. Indexing Based on Location and ccTLD.

So far, we have looked at how archiving is affected by location and path depth. Next we look at how these factors affect search engine indexing. Similar to what we did with archiving, here we look at how location (Arabic GeoIP and Arabic ccTLD) affects indexing.

Table XVIII shows the breakdown of indexing based on location. For seed URIs with both Arabic GeoIP and Arabic ccTLD, we found that 85% are indexed by Google. For those with only Arabic ccTLD, 76% were indexed, and for those with only Arabic GeoIP, 74% were indexed. We found that seed URIs that had some Arabic location (GeoIP or ccTLD) had a higher indexing rate (79%) than URIs with no Arabic location evidence (67%).

5.2.3. Indexing Based on URI Path Depth.

Here we look at indexing based on URI path depth. As with archiving, we would expect that URIs with lower path depths would be more likely to be indexed. As shown in Table XIX, we found that 74.6% of URIs with path depth 0 are indexed, and only 22.5% of the URIs with path depth of 2 or more are indexed. As with archiving, URIs closer to the top-level are more likely to be indexed than those with higher path depths.

6. COMPARING ARABIC ARCHIVING RATES TO OTHER LANGUAGES

Our original work [Alkwai et al. 2015] focused on the Arabic language. In this work, we wanted to compare what we had found for Arabic language Web pages with those in other languages.

First, we chose English because it is the most popular language on the Internet. More than 65 countries around the world have English as an official language [Agency 2015]. In 2015, 62.4% of English speakers used the Internet, exceeding the world average of 45% of the population using the Internet [Internet World Stats 2015c].

Next, we wanted a non-English European language with a high Internet using population. We found that Danish meets this, as 96% of people in Denmark are Internet users. In addition, Denmark has a government initiative to archive Danish culture heritage on the Web [Zierau 2015; Christensen 2005].

Finally, we chose an Asian language, Korean. Among Asian countries, South Korea has the highest percentage of its population using the Internet (92%) [Internet World Stats 2015b]. Korean is an official language in both North Korea and South Korea, but since most people in North Korea have limited access to the Internet [Sparkes 2014], we only consider South Korea.

6.1. Selecting Seed URIs

The steps performed to collect a sample of URIs for each new language (English, Danish, Korean) were similar to the steps performed to collect the Arabic sample. For the English and Danish languages, the DMOZ directory has a large sample so we randomly collected 10,000 unique seeds for each. For the Korean language there is only a small number of URIs in the DMOZ directory. After removing duplicates we collected all 2,347 URIs found in DMOZ. Using cs.odu.edu machines we tested the existence of each seed URI on the live Web and collected only those that returned HTTP 200 OK status code (some collected after redirection), as shown in Table XX.

6.2. Determining Language

To check the languages of these Web pages we used the HTTP Content-Language, HTML title tag, and Trigram method, as described in Section 3.2. However, instead of the language detection API client method that was performed on Arabic Web pages, we used langID [Lui 2011] since the previous method is costly in terms of money when tested on a large dataset [Alkwai 2016b]. We found that both tools determined the same language on 99.4% of the Arabic seeds. As done previously, we considered the Web page to be part of the language if it passes any one of the language tests. We identified 8,576 English language seed URIs [Alkwai 2017c], 6,331 Danish language seed URIs [Alkwai 2017b], and 1,517 Korean language seed URIs [Alkwai 2017d]. The results for the seed URIs are shown in Table XX.

6.3. Crawling Seed URIs

To increase the size of our dataset, we crawled the seed URIs of the new languages, between December 2015 - March 2016. For the seed URIs that had at least one memento, we crawled the oldest memento and gathered more URIs as shown in Table XX. In the Arabic dataset we used the latest memento, however that resulted in many duplicates with the live dataset.

Then we crawled the unique dataset two levels deep. If a link leads to a page we have not discovered before, then we follow all of the links on that page as well. We performed this approach twice.

From the resulting crawled dataset, only some were available on the live Web. We ran the live Web pages through the language tests. Table XX shows the total crawled seed URIs that passed the language tests for each language.

At the bottom of Table XX, we group both seed and crawled URIs. In total, we collected 146,526 English URIs, 105,350 Danish URIs, and 9,482 Korean URIs, which we analyze and compare to the Arabic results in the remainder of the paper. We found that only 60.87% of live Arabic URIs passed the language test, meaning many pages linked by Arabic Web pages are not in Arabic, which is lower than the other three languages. We found that 77.67% of live English URIs passed the language test, followed by 74.86% of live Danish URIs, and 72.00% of live Korean URIs.

6.4. Unique Domains

First, we investigate the number of unique domains in our new datasets. For English URIs there are 27,857 unique domains, for Danish URIs there are 11,664 unique domains, and for Korean URIs there are 2,925 unique domains.

The most frequent domains for English, Danish, and Korean are shown in Tables XXI, XXII, and XXIII, respectively. We consider popularity anything that has a rank less than 9000 in either the global or local Alexa ranking. Notice that some websites are not ranked by Alexa. Note that some of the top unique domains were in the seed

Table XX: Collected datasets for Arabic, English, Danish, and Korean

Total	Arabic	English	Danish	Korean
Seeds	15,742	10,000	10,000	2,347
Live	11,014	9,384	9,245	2,070
Passed language test	7,976	8,576	6,331	1,517
Crawled	663,443	224,249	174,369	16,016
Live	482,905	176,261	131,484	11,099
Passed language test	292,670	137,950	99,019	7,965
Total passed language test	300,646	146,526	105,350	9,482
Total percentage of live passed language	60.87%	77.67%	74.86%	72.00%

DMOZ list, such as kuwaitiful.com although they are not popular based on Alexa global and local rankings in April 2016. In some cases unpopular domains (e.g., kuwaitiful.com, iconspedia.com, and shopper.com) are overrepresented because they are densely linked and function as "crawler traps" for naive search strategies; this is the basis for historic behavior of search engines only sampling dense sites instead of crawling the entire site [Smith and Nelson 2008].

Based on our crawl approach mentioned in Section 6.3, we consider the URIs collected from DMOZ as depth 0, and the URIs collected from depth 0 in the archive as depth 1, and the URIs collected from depth 1 in the live Web as depth 2, and the last URIs collected from depth 2 in the live Web as depth 3. For example, in Table XXI, we found 7 different URIs with same domain, espn.go.com. Then we crawled the 7 Web pages from the archive to get 422 different Web pages. Then we crawled the live Web of those Web pages to get 4,250 different URIs. Finally we crawled those URIs and got 7,403 URIs.

In general, we found six kinds of results in the most frequent domains:

(1) in DMOZ seed list and popular, such as espn.go.com in Table XXI,
(2) in DMOZ seed list and not popular, such as rito.dk in Table XXII,
(3) in DMOZ seed list and no local or global ranking, such as kuwaitiful.com in Table XXI,
(4) not in DMOZ seed list and popular, such as twitter.com in Table XXI,
(5) not in DMOZ seed list and not popular, such as shopper.com in Table XXI,
(6) not in DMOZ seed list and no local or global ranking, such as modspil.dk in Table XXI,

In Table XXI, we found that 7 out of the top 10 English language domains were popular. In Table XXII, we found that 7 out of the top 10 Danish language domains were popular. In Table XXIII, we found that 3 out of the top 10 Korean language domains were popular. However, 3 domains were only neither ranked globally nor locally.

6.5. Top Level Domains

Here we investigated the TLD and ccTLD of the unique domains. In Table XXIV and in Figure 13 we show Arabic, English, Danish, and Korean common TLDs. Here we use common TLDs without ccTLD (e.g., .com.sa and .com.kw are merged). We found that the most common TLD for both Arabic language Web pages and English language Web pages was .com. Unlike Arabic, the most common TLD for Danish language websites was .dk, which is the ccTLD for Denmark. The most common TLDs for Korean language websites was .kr, which is the ccTLD for South Korea.

In Table XXV we show the top 10 TLD for Arabic, English, Danish, and Korean. For Arabic the top TLD is .com, same as English and Korean. For Danish the top TLD is .dk. In English, trailing .com is .org, .gov, and .edu which are TLDs commonly

Table XXI: Most frequent English domains, showing Alexa global and local rankings in April 2016 and the number of URLs collected at each depth (* indicates that domain was *not* in the DMOZ seed list)

Rank	English Domain	Alexa Global Rank	Alexa Local Rank	Depth 0	Depth 1	Depth 2	Depth 3	Total	Total 200 status English
1	espn.go.com	116	21	7	422	4,250	2,731	7,403	6,119
2	iconspedia.com*	74,173	87,090	0	1	222	6,174	7,824	5,572
3	congress.gov	22,031	4,752	6	637	2,664	2,197	5,504	5,254
4	twitter.com*	10	8	0	470	1,699	4,122	6,291	4,614
5	github.com	70	76	1	37	142	4,116	4,296	4,247
6	itunes.apple.com	48	35	3	187	3,729	39	3,958	3,683
7	chowhound.com*	3,859	1,047	0	0	288	1,867	2,155	2,130
8	abcnews.go.com	563	160	1	43	261	1,754	2,059	1,983
9	shopper.com*	2,041,101	1,048,598	0	5	179	1,740	1,924	1,892
10	kuwaitiful.com	136,281	No rank	1	116	2,203	0	2,319	1,746

Table XXII: Most frequent Danish domains, showing Alexa global and local rankings in April 2016 and the number of URLs collected at each depth (* indicates that domain was *not* in the DMOZ seed list)

Rank	Danish Domain	Alexa Global Rank	Alexa Local Rank	Depth 0	Depth 1	Depth 2	Depth 3	Total	Total 200 status Danish
1	wattoo.dk*	644,606	2,422	0	1	435	6,484	6,919	6,647
2	ekstrabladet.dk	3,513	6	1	164	1,425	1,172	2,762	2,625
3	rito.dk	857,702	9,186	1	119	1,829	0	1,948	1,945
4	politiken.dk	12,609	30	1	26	489	1,312	1,828	1,683
5	billigsport24.dk	810,093	3,261	1	35	490	1,416	1,942	1,633
6	bakker.dk	1,240,909	3,791	1	362	1,161	2	1,526	1,205
7	modspil.dk*	8,105,019	No rank	0	1	160	696	857	821
8	batteribyen.dk	348,866	1,193	1	2	756	0	759	751
9	dr.dk	3,453	8	14	156	244	512	926	662
10	lohse.dk	8,625,930	No rank	1	224	442	0	666	590

Table XXIII: Most frequent Korean domains, showing Alexa global and local rankings in April 2016 and the number of URLs collected at each depth. All of these domains were in the DMOZ seed list.

Rank	Danish Domain	Alexa Global Rank	Alexa Local Rank	Depth 0	Depth 1	Depth 2	Depth 3	Total	Total 200 status Korean
1	iffice.com	1,433,858	No rank	1	363	342	0	706	683
2	plaync.com	30,125	519	2	89	639	31	761	508
3	bok.or.kr	163,570	2,474	1	257	0	2	260	248
4	jtv.co.kr	15,865,831	No rank	1	247	0	2	250	246
5	keimyung.ac.kr	No rank	No rank	1	113	106	0	220	220
6	bomul.com	21,493,149	No rank	1	168	51	0	221	189
7	seongju.go.kr	No rank	No rank	1	173	1	0	175	174
8	cheongju.go.kr	599,739	9,254	2	21	116	11	150	168
9	doctor.co.kr	No rank	No rank	4	111	71	30	216	168
10	ui4u.net	405,866	No rank	1	162	2	4	169	147

(but not always, e.g. web.cs.toronto.edu) limited to the United States. Next we find the United Kingdom commercial TLD, .co.uk. For Danish TLDs, we found that the Denmark ccTLD was the most frequent, trailing that was .org and .com. For the Korean TLDs, we found .com is the most common, trailed by the Korean ccTLDs co.kr, go.kr, or.kr, and ac.kr.

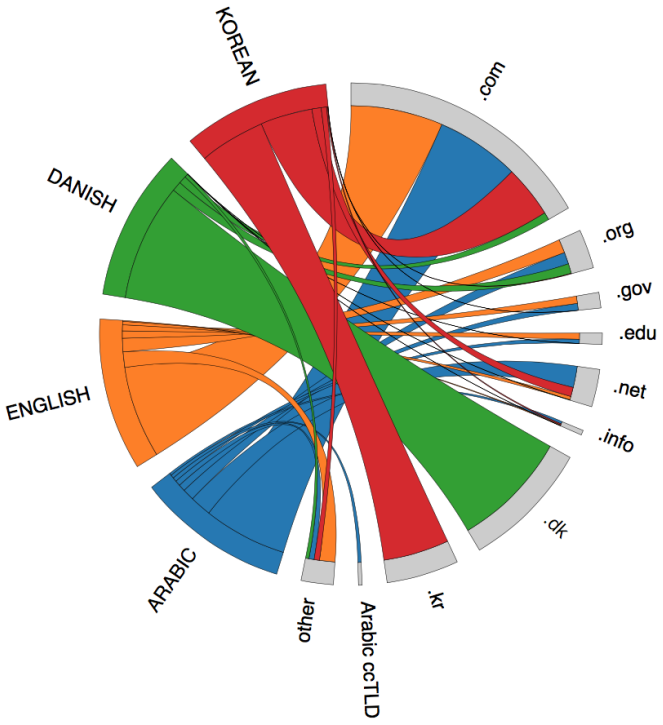


Fig. 13: Arabic, English, Danish, and Korean language and associated TLDs. Languages are on the left and labeled in all caps, and TLDs are on the right

Table XXIV: Percentage of top effective TLDs for Arabic, English, Danish, and Korean

TLD	Arabic	English	Danish	Korean
com	**60.28%**	**66.34%**	5.54%	37.50%
net	15.32%	2.47%	0.26%	6.63%
org	8.49%	9.78%	7.04%	0.38%
gov	4.91%	5.23%	<0.01%	0.09%
edu	3.05%	4.34%	0.03%	0%
Arabic ccTLD	2.38%	0%	0%	0%
info	1.73%	0.66%	0.04%	0.75%
dk	0%	0%	**84.64%**	0%
kr	0%	0%	0%	**47.37%**
other	3.74%	11.12%	2.44%	3.82%

6.6. GeoIP Location

In this section we looked at the GeoIP address location of the IP address of the unique host names. Using the same steps as in Section 4.6, we obtained the GeoIP location of English URIs, Danish URIs, Korean URIs and we compared it to Arabic GeoIP results.

In Table XXVI and in Figure 14, we show the GeoIP location of English, Danish, and Korean Web pages. We found that 59.97% of Arabic URIs are located in USA, and that 83.12% of English URIs are also located in USA. For Danish URIs we found that 42.91% of the URIs are located in Denmark. Finally, for Korean URIs 89.54% of the URIs are located in South Korea.

Table XXV: Percentage of top 10 TLDs for Arabic, English, Danish, and Korean

Arabic TLD	Count	English TLD	Count	Danish TLD	Count	Korean TLD	Count
com	58.15%	com	65.63%	dk	84.64%	com	37.50%
net	15.08%	org	9.27%	org	7.03%	co.kr	15.56%
org	6.41%	gov	4.87%	com	5.49%	go.kr	12.14%
gov.sa	1.95%	edu	4.14%	it	0.40%	or.kr	9.21%
info	1.67%	co.uk	2.68%	eu	0.34%	ac.kr	8.02%
edu.sa	1.27%	net	2.44%	no	0.34%	net	6.63%
ws	1.16%	de	1.56%	nu	0.32%	org	3.84%
org.sa	0.97%	co	1.35%	net	0.25%	is	3.24%
com.sa	0.80%	ca	1.03%	ag	0.13%	kr	1.38%
gov.eg	0.80%	info	0.67%	se	0.13%	info	0.75%

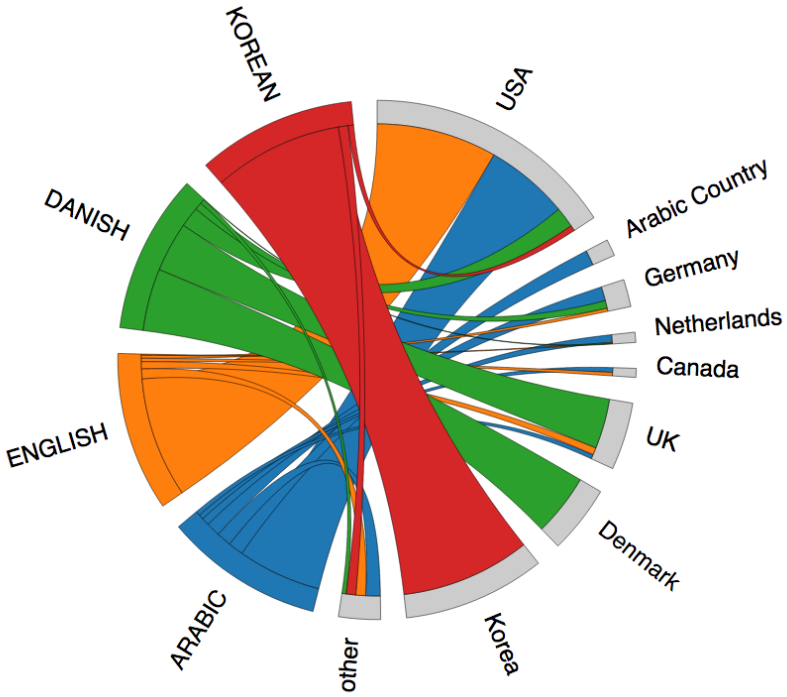


Fig. 14: Arabic, English, Danish, and Korean language and associated GeoIP location. Languages are on the left and labeled in all caps, and GeoIP locations are on the right

6.7. URI Path Depth

In Table XXVII, we investigate the depth of URIs of all languages. We found that the most frequent depth for all four language is depth one.

Note that 17.30% of Arabic URIs have a depth zero since two of the three sources for collecting seeds were Arabic website listings that are updated with URIs that had their own domain. However, when we sample English, Danish, and Korean from DMOZ only, most websites did not have their own domain, since only recently has it been easy to have your own domain.

Table XXVI: GeoIP location for Arabic, English, Danish, and Korean

Country	Arabic Percent	English Percent	Danish Percent	Korean Percent
USA	**57.97%**	**83.12%**	14.06%	3.87%
Arabic Country	10.53%	-	-	-
Germany	9.75%	2.23%	5.24%	-
Netherlands	5.29%	0.64%	0.46%	-
Canada	3.31%	2.31%	-	-
UK	3.07%	4.99%	34.37%	-
Denmark	-	-	**42.91%**	-
South Korea	-	-	-	**89.54%**
other	10.08%	6.71%	2.94%	6.58%

Table XXVII: Path depth of URI-Rs

Depth	Arabic	English	Danish	Korean
Depth 0	17.30%	2.29%	1.13%	1.6%
Depth 1	**40.42%**	**47.52%**	**34.42%**	**45.24%**
Depth 2	24.45%	19.46%	11.69%	29.54%
Depth 3	10.81%	11.06%	16.23%	16.79%
Depth 4+	7.02%	19.67%	36.53%	6.78%

6.8. Presence in the Archive

Between March-April 2016, we used Memgator [Alam and Nelson 2016], which is a different Memento aggregator than used previously, to determine if the URIs in our dataset are archived. In this tool we specified the archives that we wanted to check. We also added the Bibliotheca Alexandrina[14] since it has mirror site of the Internet Archive and in some cases has additional mementos.

We have had private communications with Youssef Eldakar from Bibliotheca Alexandrina (BA). He has indicated that recent BA crawls have been focused on Egyptian content, though there may be some non-Arabic content included. However, the selection process has so far been manual. No language identification methods are used at the present time. The BA planned to expand its crawling activities to cover Web content related to all of Arab world starting in 2013. This means that currently there is a lacuna in the Arabic Web coverage in the archive. To further check if the Bibliotheca Alexandrina have some effect on the percentage of the Arabic URIs that were archived, we tested the 2,904 seed URIs from DMOZ. We found that only 66 URIs were archived now that were not previously found in an archive in our 2015 analysis (which did not include the BA). We also found that none of them were archived only by the BA. To further investigate, we crawled a new Arabic dataset, with a total of 87,228 URIs. Over all archives, 43,481 URIs (49.84%) were archived, but only 98 of these were archived only at the BA. The result of this analysis convinced us that we would not find a significant number of URIs from our original Arabic dataset only in the BA, so we did not repeat the analysis for the entire set.

Shown in Table XXVIII, we found that 73.30% of English language Web pages are archived (i.e., have one or more mementos), but only 39.59% and 41.89% of Danish language Web pages and Korean language Web pages are archived, respectively. As expected, we found that English language Web pages are archived more than the other three groups. Although the Danish language is not that well archived in public archives, there is an effort by the Danish government to archive Danish content

[14]http://archive.org/about/bibalex_p_r.php

Table XXVIII: Archiving results for Arabic, English, Danish, and Korean seed and crawled datasets

	Arabic	English	Danish	Korean
Total DMOZ 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%
Total DMOZ	2,904	8,576	6,331	1,517
Archived	2,774	8,014	6,164	1,358
Percent	95.52%	93.44%	97.36%	89.52%
Total Crawled	297,742	137,950	99,019	7,965
Archived	158,904	99,384	35,539	2,614
Percent	53.36%	72.04%	35.89%	32.81%

in their private web archive institute NetArkive [Christensen 2005] that cannot be accessed by the general public.

By comparing the archiving results of DMOZ seeds vs. crawled URIs shown in Table XXVIII, we found that the archiving rates are higher for DMOZ in all four languages. In the seed sample we found Danish has most URIs archived with 97.36% followed by Arabic, English, and Korean. The sample size may have a role in the result, since for example we only found 1,517 Korean URIs in DMOZ. In the crawled sample we found English was most archived with 72.04% followed by Arabic, Danish, and Korean.

Next, we check the top 10 archived URI-Rs for English, Danish, and Korean. In Table XXIX, we list the top ten archived English URI-Rs and found a very high memento count such as google.com with 726,680 mementos, compared to the Danish top most archived jubii.dk with 23,815 URIs shown in Table XXX, and the Korean most memento count doumi.hosting.bora.net/infomail/error/message04.html with 54,339 mementos shown in Table XXXI.

In Table XXXI, we found no Korean TLDs in the top ten most archived URIs. We also found that 5 out of the 10 most archived Korean URIs are websites with custom error pages.

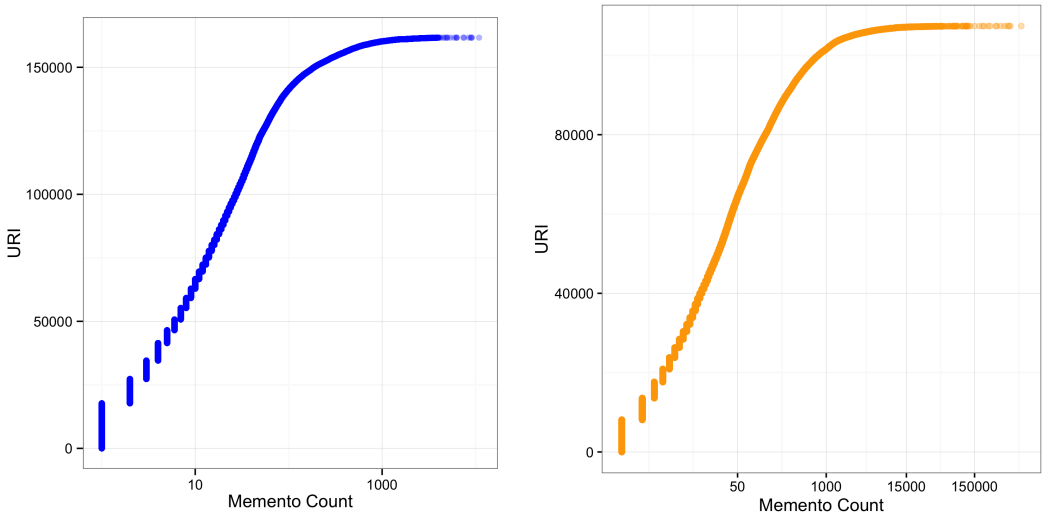
This means that missing Web pages that were to be archived were redirected to the custom error page, which was archived several times. Table XXXII shows the top 5 archived Korean URIs with a .kr TLD and their memento counts.

We noticed that some of the memento counts were higher than expected. When we investigated the TimeMaps, we found that some included a large number of redirects, since TimeMaps report HTTP events (and may include HTTP 404, 3xx, etc.) [Kelly et al. 2017]. This caused the reported number of mementos to be higher than we had expected.

Figure 15 shows the number of mementos found for each archived URI, sorted by memento count for each of the four different languages. The figures have different y-axis due to the difference in maximum memento count.

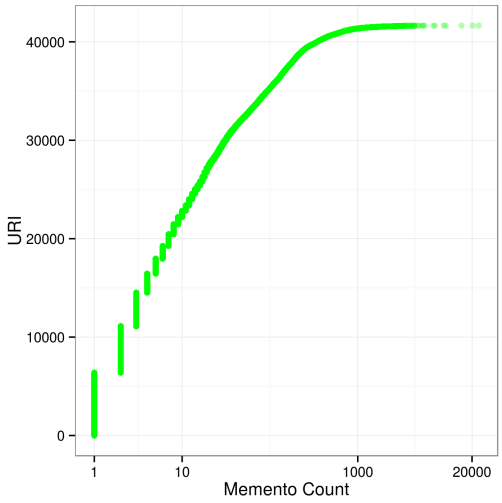
Table XXXIII shows the URI $Coverage_m(R, WA)$ of each archive. We used the same calculation as described in Section 4.4. The results show that $Coverage_m(IA)$ has the best coverage for all samples. The Arabic $Coverage_m(IA)$ is 72.87%, and the highest $Coverage_m(IA)$ was the Korean with 86.15%.

Table XXXIV shows $Coverage_m(R, WA)$. The results show that IA has the best coverage for all samples. $Coverage_m(IA)$ ranges between 91.84% and 99.38%. The Arabic IA percentage was 97.04%.

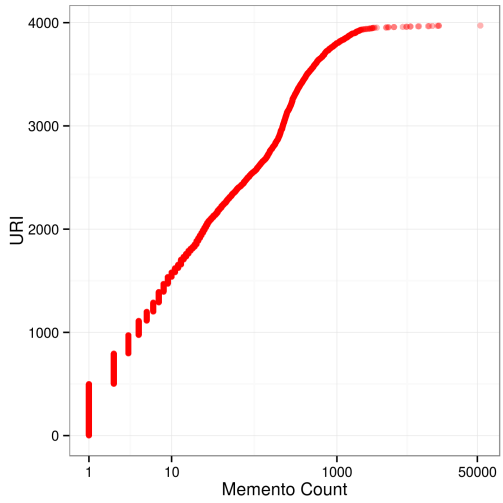


(a) Arabic

(b) English



(c) Danish



(d) Korean

Fig. 15: Memento count frequency of Arabic, English, Danish, and Korean language URIs

6.9. Creation Date

Here we want to compare the creation date of the URIs. In our previous method to predict creation date we used CarbonDate, mentioned in Section 4.5. However, this tool used a service named Topsy[15] as one of the methods to gather creation date information, which is is no longer available [Roberts 2015]. For the other language analysis, we instead used the datetime of the first memento as the creation date.

[15]http://topsy.com

Table XXIX: Top 10 archived English URI-Rs

English URI-Rs	Memento Count	Category
google.com	726,680	Search Engine
creativecommons.org/licenses/by-sa/3.0	500,903	Tool Sharing
twitter.com	480,187	Social Media
labs.reliefweb.int/	455,544	Information Digital Service
geocities.jp/server-errors/not_found.html	409,829	Web hosting
pcmag.com	361,012	PC Magazine
youtube.com/html5	317,398	Videos
myspace.com	305,304	Social Media
facebook.com	253,246	Social Media
google.com/reader/about	237,319	Search Engine

Table XXX: Top 10 archived Danish URI-Rs

Danish URI-Rs	Memento Count	Category
jubii.dk	23,815	News
espanol.yahoo.com	20,223	Search Engine
google.dk	15,160	Search Engine
twitter.com/search?q=%23ENS	10,036	Social media
spray.se	9,535	News
voila.fr/pages/arret-pages-perso.html	7,533	News
deezer.com/soon	7,269	Music
tripadvisor.dk	5,713	Travel
borsen.dk	5,588	News
www2000179.thinkquest.dk	5,396	News

Table XXXI: Top 10 archived Korean URI-Rs

Korean URI-Rs	Memento Count	Category
doumi.hosting.bora.net/infomail/error/message04.html	54,339	Error Page
joins.com	17,096	News
html.giantsoft.co.kr/404.html	17,046	Error Page
errdoc.gabia.net/403.html	16,414	Error Page
daum.net	14,305	Search Engine
img.kbs.co.kr/pageerror	13,042	Error Page
hani.co.kr/oops.html	12,676	Error Page
chosun.com	9,839	Newspaper
donga.com	9,587	News
hankooki.com	7,762	Search Engine

Table XXXII: Top 5 archived Korean URI-Rs with a .kr TLD

Korean URI-Rs	Memento Count	Category
wisecart.co.kr/pr	6,922	Shopping
auction.co.kr	4,883	Shopping
hani.co.kr	4,175	Newspaper
postech.ac.kr	2,702	Education
whois.co.kr	2,605	Domain Register

Figure 16 shows the creation date of Arabic URIs and the first memento of English, Danish, and Korean URIs. We note that the number of mementos for all four languages has increased in recent years. We note that the creation of Arabic Web pages has increased over the years, with 2013 as the most frequent creation date. For both English

Table XXXIII: $Coverage_m$ of archived URI-Rs present in each archive

Archive	Arabic	English	Danish	Korean
Internet Archive	72.87%	79.27%	74.74%	86.15%
Archive-It	21.26%	9.00%	5.23%	0.15%
Archive.today	2.14%	1.09%	0.7%	0.39%
WebCite	2.08%	0.40%	0.06%	0.47%
Iceland Web Archive	1.17%	1.63%	1.98%	0.07%
British Library Archive	0.29%	0.09%	0.05%	<0.01%
UK Parliament Web Archive	0.10%	0.05%	0.01%	0.01%
Proni	0.05%	0.02%	0.01%	<0.01%
UK National Archives	0.04%	0.02%	<0.01%	0.01%
Stanford	<0.01%	1.20%	0.72%	<0.01%
Bibliotheca Alexandrina	-	7.24%	16.39%	12.75%

Table XXXIV: $Coverage_r$ of archived URI-Rs present in each archive

Archive	Arabic	English	Danish	Korean
Internet Archive	97.04%	98.14%	99.38%	91.84%
Archive.today	6.58%	17.94%	12.15%	31.50%
WebCite	6.00%	2.26%	0.43%	0.96%
Archive-It	5.49%	24.48%	3.83%	5.61%
British Library Archive	1.06%	1.97%	0.30%	0.15%
UK Parliament Web Archive	0.88%	2.70%	0.35%	2.19%
Iceland Web Archive	0.87%	4.16%	2.29%	2.82%
UK National Archives	0.62%	1.04%	0.15%	1.96%
Proni	0.21%	0.44%	0.07%	0.08%
Stanford	0.11%	2.48%	0.38%	0.20%
Bibliotheca Alexandrina	-	21.50%	19.89%	36.25%

and Danish the most frequent year for the first Memento-Datetime is 2015, and the most frequent first Memento-Datetime year for Korean is 2002.

In general, we can not know when a resource was created, as this information is not stored anywhere [Nelson 2010]. The Creation-Datetime, Memento-Datetime, and Last-Modified Datetime all have different semantics. The Memento-Datetime is stored in the archive. However, it is not the Creation-Datetime or the Last-Modified Datetime. The URIs that are archived could have existed for a longer time than the first Memento-Datetime indicates, because it was not archived as soon as it was created. Thus, there may be a large gap between the actual Creation-Datetime and the Memento-Datetime.

Over the last 20 years the archiving rate has increased as the capacity of web archives increased. In the early days, a Web page might have only a few mementos, but more now as the archiving rate has increased. Also, the archival sampling process before was slow, compared to the sampling process used now. So even if the first Memento-Datetime was in 2010, the Web page may have been actually created in 2000, but was not archived until 2010. Historically, the web archiving rate has been much slower than that of indexing by search engines, and entire web sites could be born and die before being visited by web archives [Smith et al. 2006]. As a result, as we go further back in time, the web archives become less reliable as a source of evidence as to when pages were first created.

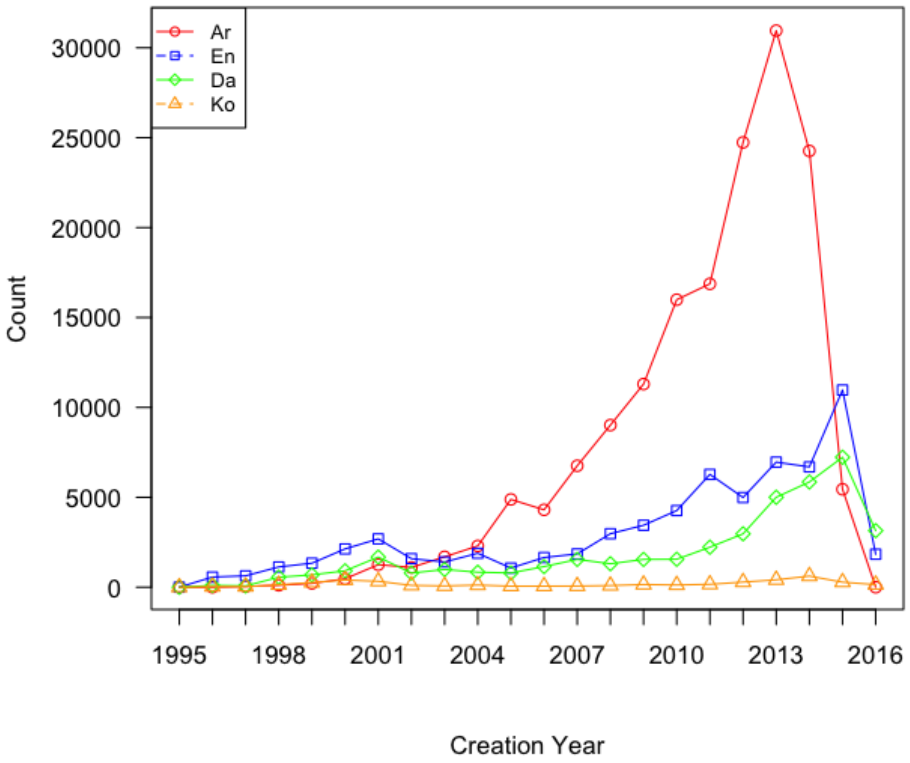


Fig. 16: Arabic creation date and English, Danish, and Korean first Memento-Datetime

7. DISCUSSION

In this work we have to note that different language tests can be affected by factors that we did not consider in this work. One such factor is having multilingual Web pages. Many non-English Web pages have common English terms, such as Facebook, Twitter, or other terms in another language that can affect the results. Also, the length of the text may affect the results, because some tools work better with longer text.

When we started measuring how well the Arabic Web is archived, it turned out that characterizing what is the Arabic Web was more challenging then we initially thought. Previous work used the TLD as identifying the location of the domain [Žabička 2003]. However, we found that in this work other factors must be considered such as the language of the page, ccTLD, and the GeoIP location. If we had restricted our research to a specific ccTLD or location, the results may have been different for archiving and indexing. For example, in our Arabic dataset 65% of what was archived had neither Arabic ccTLD nor Arabic location.

We found that most English Web pages had a generic TLD, Korean Web pages mostly had .kr ccTLD, and Danish Web pages were mostly .dk ccTLD. But this was not the case for Arabic domains, where the majority did not have an Arabic ccTLD. We think

this is because, similar to English, the Arabic language maps to many countries within a large region instead of one or a small number of countries. In this work we found 84.64% of Danish websites had .dk ccTLD, and 47.37% of Korean websites had .kr ccTLD, on the other hand only 22.65% of Arabic websites had Arabic ccTLD.

It is also noted that the language of some websites may shift over time, without a change in the domain. An example of this is the Libyan newspaper website www.azzahfalakhder.com that changed over time to be a German sports blog written in the German language [Aturban 2016]. In another example, the language stayed the same but the content was different such as www.alfajraljadeed.com, where it was a Libyan newspaper that changed to an Egyptian news Website.

In addition, in this work we did not explore the visibility of the Web pages from one country versus another. In some cases some websites such as www.6arab.com are currently restricted from being accessing in Saudi Arabia but not in other countries [Alkwai 2016a]. In this work, we only focused on what is seen can be from Norfolk, VA (location of Old Dominion University) and from San Francisco, CA (location of the Internet Archive). We may have different results if this was tested in different locations. Also, some countries such as Russia blocked the Internet Archive in June 2015 [Moody 2016]. This would also affect the results if the test was performed there.

8. CONCLUSIONS

In this study, we evaluated how well Arabic Web pages are archived and indexed. First we collected Web pages from Arabic directories, then determined if these Web pages are written in the Arabic language. After that, we crawled the seed URIs to enlarge the dataset. Then we analyzed those Arabic Web pages. We used four different language tests to check the Web pages language, then we performed basic data analysis, such as checking the presence on the live Web, top-level domain, GeoIP, URI path depth, and creation date. Then we checked if these Web pages are archived and measured the archiving frequency and the gap between creation date and the first archived version. Also, we investigated if archiving and indexing were affected by Arabic country code top-level domain, Arabic GeoIP, URI path depth, or creation date. We also compared the archiving rate of Arabic language Web pages with that of English, Danish, and Korean language Web pages.

We found that English has a higher archiving rate than Arabic, with 72.04% archived. However, Arabic has a higher archiving rate than Danish and Korean, with 53.36% of Arabic URIs archived and 69% indexed by Google, followed by Danish and Korean with 35.89% and 32.81% archived respectively. Arabic still has a lacuna in its archive, and there is a need to preserve its pages better. Danish has a private Web archive that cannot be accessed by public (i.e., a dark archive), however, if the Danish archive goes from dark to live, we expect that the archive percentage of Danish language pages will be high. On the other hand, South Korea appears self-contained and will need to archive its pages in order to preserve them for the future.

We also found that Arabic archiving and indexing appear to be affected by the top-level domain, GeoIP location, URI path depth, creation date, and presence in a directory. Arabic language sites having either only an Arabic GeoIP or only an Arabic top-level domain are less likely to be archived than others. Most Arabic and English language pages are located in the US, with only 14.84% of the Arabic URIs having an Arabic country code top-level domain (e.g., .sa) and only 10.53% having a GeoIP in an Arabic country, Danish language pages are mostly located in Denmark, and Korean language pages in South Korea. Having either only an Arabic GeoIP or only an Arabic top-level domain appears to negatively impact archiving. We also found that most of the archived pages are near the top-level of the site and deeper links into the site are not well archived. The presence in a directory positively impacts indexing and presence

in the DMOZ directory, specifically, positively impacts archiving in all four languages. Popular Western sites based on Alexa ranking (such as facebook.com, wikipedia.org, and google.com) were in the top 10 domains found in our Arabic, English, Danish, and Korean dataset. We also notice that the Arabic language community is using services hosted on Western sites and their cultural discourse is occurring on Western sites where archiving is likely to be already taking place. Finally, we found that although the number of English, Arabic, and Danish Web pages archived over time was increasing, the number of newly archived Korean Web pages has remained fairly steady over time.

ACKNOWLEDGMENTS

This work is supported in part by University of Hail, Hail, Saudi Arabia. We would like to thank the anonymous reviewers and the associate editors for their feedback for improving the presentation of our findings.

REFERENCES

Central Intelligence Agency (Ed.). 2015. *The World Factbook 2014-15*. Government Printing Office.

Scott G Ainsworth, Ahmed Alsum, Hany M. SalahEldeen, Michele C. Weigle, and Michael L. Nelson. 2011. How Much of the Web is Archived?. In *Proceedings of the 11th IEEE/ACM Joint Conference on Digital Libraries (JCDL)*. ACM, 133–136.

Sawood Alam and Michael L. Nelson. 2016. MemGator - A Portable Concurrent Memento Aggregator. In *Proceedings of the 16th IEEE/ACM Joint Conference on Digital Libraries (JCDL)*. ACM, 234–244.

Abdulrahman Alarifi, Mansour Alghamdi, Mohammad Zarour, Batoul Aloqail, Heelah Alraqibah, Kholood Alsadhan, and Lamia Alkwai. 2012. Estimating the Size of Arabic Indexed Web Content. *Scientific Research and Essays* 7, 28 (2012), 2472–2483.

Lulwah M. Alkwai. 2016a. Are My Favorite Arabic Websites Archived? http://ws-dl.blogspot.com/2016/10/2016-10-24-are-my-favorite-arabic.html. (2016).

Lulwah M. Alkwai. 2016b. Language Detection: Where to start? http://ws-dl.blogspot.com/2016/03/2016-03-22-language-detection-where-to.html. (2016).

Lulwah M. Alkwai. 2017a. Arabic language Web pages dataset. (1 2017). DOI:http://dx.doi.org/10.6084/m9.figshare.4588702.v2

Lulwah M. Alkwai. 2017b. Danish language Web pages dataset. (1 2017). DOI:http://dx.doi.org/10.6084/m9.figshare.4588732.v1

Lulwah M. Alkwai. 2017c. English language Web pages dataset. (1 2017). DOI:http://dx.doi.org/10.6084/m9.figshare.4588729.v2

Lulwah M. Alkwai. 2017d. Korean language Web pages dataset. (1 2017). DOI:http://dx.doi.org/10.6084/m9.figshare.4588735.v1

Lulwah M. Alkwai, Michael L. Nelson, and Michele C. Weigle. 2015. How Well Are Arabic Websites Archived?. In *Proceedings of the 15th IEEE/ACM Joint Conference on Digital Libraries (JCDL)*. ACM, 223–232.

Ahmed AlSum. 2014. *Web Archive Services Framework for Tighter Integration Between the Past and Present Web*. Ph.D. Dissertation. Old Dominion University.

Ahmed AlSum, Michele C. Weigle, Michael L. Nelson, and Herbert Van de Sompel. 2014. Profiling web archive coverage for top-level domain and content language. *International Journal on Digital Libraries* 14, 3-4 (2014), 149–166.

Mohamed Aturban. 2016. Pro-Gaddafi Digital Newspapers Disappeared from the Live Web! http://ws-dl.blogspot.com/2016/11/2016-11-05-pro-gaddafi-digital.html. (2016).

Ricardo Baeza-Yates, Carlos Castillo, and Efthimis N Efthimiadis. 2007. Characterization of National Web Domains. *ACM Transactions on Internet Technology (TOIT)* 7, 2 (2007), 9.

Kenneth R. Beesley. 1988. Language Identifier: A Computer Program for Automatic Natural-Language Identification of On-line Text. In *Proceedings of the 29th Annual Conference of the American Translators Association*, Vol. 47. 54.

Brian E. Brewington and George Cybenko. 2000. Keeping up with the changing web. *Computer* 33, 5 (2000), 52–58.

Jamie Callan, Mark Hoy, Changkuk Yoo, and Le Zhao. 2009. The ClueWeb09 Dataset, 2009. *http://boston.lti.cs.cmu.edu/Data/clueweb09* (2009).

Soumen Chakrabarti, Martin Van den Berg, and Byron Dom. 1999. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks* 31, 11 (1999), 1623–1640.

Junghoo Cho. 2001. *Crawling the web: discovery and maintenance of large-scale web data*. Ph.D. Dissertation. Stanford University.

Niels Christensen. 2005. Preserving the bits of the Danish Internet. In *5th International Web Archiving Workshop (IWAW05), Viena,Austria*.

Facebook. 2016. Company Info / Facebook Newsroom. https://web.archive.org/web/20161110081856/https://newsroom.fb.com/company-info/. (2016).

Daniel Gomes and Mário J. Silva. 2005. Characterizing a National Community Web. *ACM Transactions on Internet Technology (TOIT)* 5, 3 (2005), 508–531.

Jessie Graves. 2012. Python Language Detector. https://github.com/decultured/Python-Language-Detector. (2012).

Hugo C. Huurdeman, Anat Ben-David, Jaap Kamps, Thaer Samar, and Arjen P. de Vries. 2014. Finding Pages on the Unarchived Web. In *Proceedings of the 14th IEEE/ACM Joint Conference on Digital Libraries (JCDL)*. IEEE, 331–340.

Internet World Stats. 2009. Arabic Speaking Internet Users Statistics. https://web.archive.org/web/20100515122707/http://www.internetworldstats.com/stats19.htm. (2009).

Internet World Stats. 2015a. Arabic Speaking Internet Users Statistics. https://web.archive.org/web/20160229163031/http://www.internetworldstats.com/stats19.htm. (2015).

Internet World Stats. 2015b. Internet Users in Asia November 2015. https://web.archive.org/web/20160422031013/http://www.internetworldstats.com/stats3.htm. (2015).

Internet World Stats. 2015c. Internet World Users By Language. https://web.archive.org/web/20160424042315/http://www.internetworldstats.com/stats7.htm. (2015).

Kent Johnson. 2010. Guess Language. https://github.com/kent37/guess-language. (2010).

Brewster Kahle. 1997. Preserving The Internet. *Scientific American* 276, 3 (1997), 82–83.

Andreas M. Kaplan and Michael Haenlein. 2010. Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons* 53, 1 (2010), 59–68.

Mat Kelly, Lulwah M. Alkwai, Michael L. Nelson, Michele C. Weigle, and Herbert Van de Sompel. 2017. *Impact of URI Canonicalization on Memento Count*. Technical Report submitted for publication. Old Dominion University.

Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. 2014. Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLoS ONE* 9, 12 (2014), e115253.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a Social Network or a News Media?. In *Proceedings of the 19th international conference on World wide web*. ACM, 591–600.

Jill Lepore. 2015. The Cobweb: Can the Internet be Archived? http://www.newyorker.com/magazine/2015/01/26/cobweb, *The New Yorker* (January 2015).

Marco Lui. 2011. Language Identifier. https://github.com/saffsd/langid.py. (2011).

Frank McCown and Michael L. Nelson. 2007. Agreeing To Disagree: Search Engines And Their Public Interfaces. In *Proceedings of the 7th IEEE/ACM Joint Conference on Digital Libraries (JCDL)*. ACM, 309–318.

Glyn Moody. 2016. Wayback Machine's 485 billion web pages blocked by Russian government order. http://arstechnica.com/tech-policy/2015/06/wayback-machines-485-billion-web-pages-blocked-by-russian-government-order/. (2016).

Michael L. Nelson. 2010. Memento-Datetime is not Last-Modified. http://ws-dl.blogspot.com/2010/11/2010-11-05-memento-datetime-is-not-last.html. (2010).

Alireza Noruzi. 2007. A Study of HTML Title Tag Creation Behavior of Academic Web Sites. *The Journal of Academic Librarianship* 33, 4 (2007), 501–506.

Leonard Richardson. 2013. Beautiful soup. *Crummy: The Site* (2013).

Daniel Roberts. 2015. Apple Just Shut Down the Best Way to Search Twitter. http://fortune.com/2015/12/16/apple-shuts-down-topsy/. (2015).

Hany M. SalahEldeen and Michael L. Nelson. 2013a. Carbon Dating the Web: Estimating the Age of Web Resources. In *Proceedings of the Temporal Web Analytics Workshop (TempWeb)*. 1075–1082.

Hany M. SalahEldeen and Michael L. Nelson. 2013b. Resurrecting My Revolution. In *International Conference on Theory and Practice of Digital Libraries: Research and Advanced Technology for Digital Libraries*. Springer, 333–345.

Joan A. Smith, Frank McCown, and Michael L. Nelson. 2006. Observed web robot behavior on decaying web subsites. *D-Lib Magazine* 12, 2 (2006).

Joan A. Smith and Michael L. Nelson. 2008. Site design impact on robots: An examination of search engine crawler behavior at deep and wide websites. *D-Lib Magazine* 14, 3 (2008).

Matthew Sparkes. 2014. Internet in North Korea: everything you need to know. http://www.telegraph.co.uk/technology/11309882/Internet-in-North-Korea-everything-you-need-to-know.html, *The Telegraph* (2014).

Reem Suwaileh, Mucahid Kutlu, Nihal Fathima, Tamer Elsayed, and Matthew Lease. 2016. ArabicWeb16: A New Crawl for Todays Arabic Web. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR*. ACM, 673–676.

Mike Thelwall and Liwen Vaughan. 2004. A Fair History of the Web? Examining Country Balance in the Internet Archive. *Library & Information Science Research* 26, 2 (2004), 162–176.

Twitter. 2016. Twitter usage / Company facts. https://web.archive.org/web/20161111202246/https://about.twitter.com/company. (2016).

Herbert Van de Sompel, Michael L. Nelson, and Robert Sanderson. 2013. HTTP framework for time-based access to resource states – Memento, Internet RFC 7089. http://tools.ietf.org/html/rfc7089, (2013).

Herbert Van de Sompel, Michael L. Nelson, Robert Sanderson, Lyudmila L. Balakireva, Scott Ainsworth, and Harihar Shankar. 2009. *Memento: Time Travel for the Web*. Technical Report arXiv:0911.1112.

Petr Žabička. 2003. Archiving the Czech Web: Issues and Challenges. In *3rd Workshop on Web Archives, Trondheim, Norway, August 21st*. 111–117.

Petros Zerfos, Junghoo Cho, and Alexandros Ntoulas. 2005. Downloading textual hidden web content through keyword queries. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*. IEEE, 100–109.

Eld Zierau. 2015. Automatic Collection of Danish online Cultural Heritage Outside the. dk Top Level Domain. In *12th International Conference on Preservation of Digital Objects*. http://hdl.handle.net/109.1.5/2ed8684e-62a3-4ba3-bdc5-7c90869fc616