Open access • Proceedings Article • DOI:10.1145/2756406.2756912

How Well Are Arabic Websites Archived — Source link

Lulwah M. Alkwai, Michael L. Nelson, Michele C. Weigle

Institutions: Old Dominion University

Published on: 21 Jun 2015 - ACM/IEEE Joint Conference on Digital Libraries

Topics: Web archiving and Web page

How Well Are Arabic Websites Archived?

Lulwah M. Alkwai*, Michael L. Nelson, and Michele C. Weigle
Department of Computer Science
Old Dominion University
Norfolk, Virginia 23529 USA
{lalkwai,mln,mweigle}@cs.odu.edu

ABSTRACT

It is has long been anecdotally known that web archives and search engines favor Western and English-language sites. In this paper we quantitatively explore how well indexed and archived are Arabic language web sites. 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 Al-Jazeera sites) and culled the collection to 7,976 definitely 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 discovered: 1) 46% are not archived and 31% are not indexed by Google (`www.google.com`), 2) only 14.84% of the URIs had an Arabic country code top-level domain (e.g., `.sa`) and only 10.53% had a GeoIP in an Arabic country, 3) having either only an Arabic GeoIP or only an Arabic top-level domain appears to negatively impact archiving, 4) most of the archived pages are near the top level of the site and deeper links into the site are not well-archived, 5) the presence in a directory positively impacts indexing and presence in the DMOZ directory, specifically, positively impacts archiving.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries and Archives

General Terms

Design, Experimentation, Measurement

Keywords

Web Archiving, Indexing, Digital Preservation, Arabic Web

*Department of Computer Science and Software Engineering, University of Hail, Hail, Saudi Arabia

JCDL'15, June 21–25, 2015, Knoxville, Tennessee, USA.

ACM 978-1-4503-3594-2/15/06 ...$15.00.
http://dx.doi.org/10.1145/2756406.2756912.

1. INTRODUCTION

Arabic is the fourth most popular language on the Internet, trailing only English, Chinese, and Spanish [12]. 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 [10], but by the end of 2013 that had increased to almost 36% (over 135 million), approaching the world average of 39% of the population using the Internet [11]. In 2010, the size of the indexed Arabic web was estimated to be about 2 billion pages [2]. 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 webpages is short (44-100 days) [6, 13], and that resources are disappearing from the live web [14, 19]. Thus, webpages need to be preserved for future cultural and historical data mining. Web archiving is becoming recognized as an important problem [15], and several institutions, most notably the Internet Archive, have created archives to preserve websites. There are even several country and language specific archives[1], such as the BnF Web Archives (.fr domain)[2], the National Archives of the UK government (.uk domain)[3], and the Icelandic Web Archive (.is domain)[4].

The lack of focused archiving of the Arabic web motivates our study of how well Arabic language webpages are being archived today. To investigate this, we obtained a sample of URIs from Arabic web directories. For those webpages 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 webpages to collect more URIs to investigate.

With this study, we have 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 webpages with Arabic language content use generic 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

[1]A list of prominent web archives is available at `http://netpreserve.org/resources/member-archives`.
[2]`http://www.bnf.fr/fr/collections_et_services/collections_departements.html`
[3]`http://www.nationalarchives.gov.uk`
[4]`http://vefsafn.is/`

Table 1: Countries with Arabic as the official language, their population, percentage of those who are Internet users, and ccTLD. Source: [11].

Country	Population (2014)	% are Internet Users	ccTLD	Note
Egypt	86,895,099	49.6%	.eg	
Algeria	38,813,722	16.5%	.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.

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.

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. [1] 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 [3] showed that the archival percentages had increased from 33% to 95%. However, these studies were not focused on content from specific countries or content in specific languages.

Thelwall and Vaughn [20] 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. [4] 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 country-code Top-Level Domain (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 [8] 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.

A recent investigation into the unarchived web [9] has shown that the archived web can be a rich source of links to potentially unarchived content. In this 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 [21,22] 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 seed URIs, determining language, and crawling Arabic seed URIs.

3.1 Selecting Seed URIs

First, we searched for Arabic website directories and took the top three based on Alexa ranking[5]. Between March-May 2014, we collected all URIs from these three Arabic website directories: 1) the Arabic DMOZ listing, registered in 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 2 shows the number of collected URIs from these three sources. We collected 15,092 unique seed URIs. 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 2: 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 1, sorted by population, lists each country where Arabic is an official language, its population, the percentage of its population that are Internet users, its country code TLD, and if other languages are spoken. Although we gathered webpages 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 webpage, 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 629-2 code for Arabic, we considered the webpage to be written in Arabic.

- **HTML title tag** - The HTML title tag is often a good indicator of the language of a webpage's content [17]. We extracted the title tag of each webpage and used the guess-language Python library[6] to determine the language.

- **Trigram method** - The trigram technique uses letter trigrams, sequences of three letters, to determine language [5]. The identification is performed through basic trigram lookups paired with unicode character set recognition. We used the Python-Language-Detector tool[7], which implements the trigram method, on the extracted text from the HTML of each webpage.

- **Language detection API client** - The Language Detection API[8] is a web service that detects 106 different languages. We ran the test on the extracted text from the HTML of each webpage.

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. Figure 1 shows the intersection between the four language tests. We found 872 of the URIs tested as Arabic language in all four tests. We decided to consider the webpage 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 webpages, we found 7,976 that passed at least one of the language tests. We consider this set to be our Arabic seed URIs.

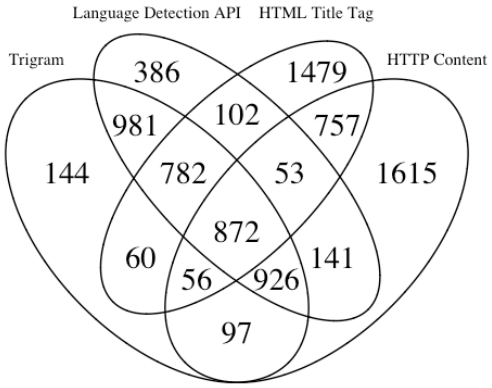

Figure 1: Language test intersection testing for Arabic language

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. 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 three sets (original URIs, 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.

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

4. RESULTS

In this section we examine the characteristics of our Arabic URI dataset. We investigate the number of unique domains, TLD and country-code TLD (ccTLD), URI path

[5] `http://www.alexa.com`
[6] `https://code.google.com/p/guess-language/`
[7] `https://github.com/decultured/Python-Language-Detector`
[8] `https://detectlanguage.com/`

depth, presence in the archive, and estimated creation date for our combined dataset. For the original Arabic seed URI dataset we also investigate the GeoIP location and presence in the Google search engine index.

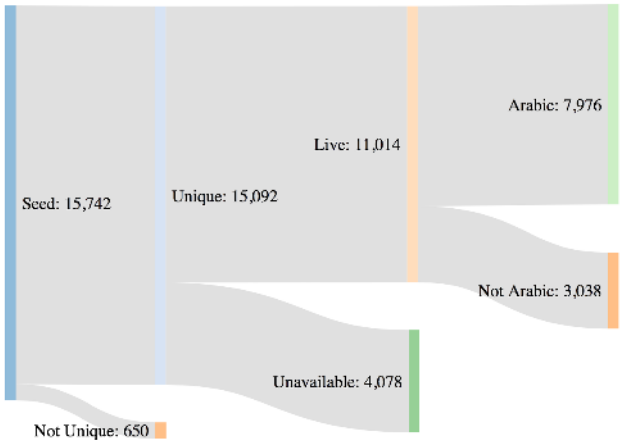

Figure 2: Seed URIs count detail

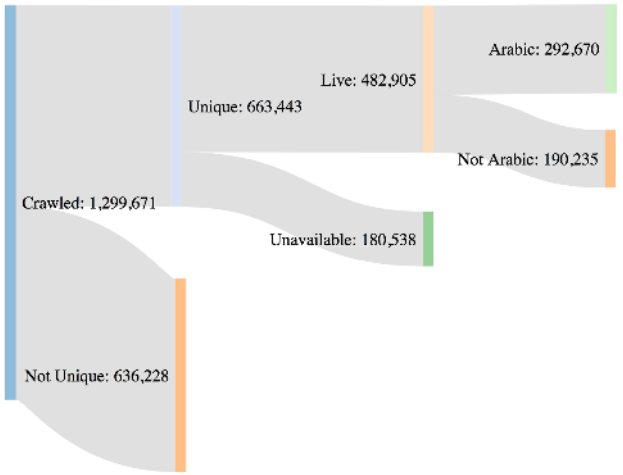

Figure 3: Crawled URIs count detail

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 3. We also tested the GeoIP location of the top-level webpage 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.

Table 3: Most frequent domains

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

4.2 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 4 shows the distribution of the top 10 effective TLDs. We also checked if the ccTLD was from a country where Arabic is an official language (listed in Table 1). 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. We note that the top Arabic ccTLD, .sa for Saudi Arabia, is used in fewer URIs than the generic TLDs .com, .net, and .org.

Table 5 shows the top 5 ccTLDs from Arabic-speaking countries. We found that Saudi Arabia was the most frequent Arabic ccTLD, followed by Egypt and Jordan.

Table 4: Top 10 effective TLDs

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%

Table 5: Top 5 Arabic ccTLDs

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%

4.3 URI Path Depth

URI path depth is an important factor in archiving, as we assume that webpages nearer to the top-level of a site are better archived than pages deeper into the site (i.e., with higher path depth). Table 6 shows the breakdown of URI path depth for our Arabic URIs. As expected, over half of the URIs have a path depth of 0 or 1, with barely 7% having a path depth greater than 3.

Table 6: Path depth of the Arabic URIs

Path Depth	Example	Percent
0	example.com	17.30%
1	example.com/a	40.42%
2	example.com/a/b	24.45%
3	example.com/a/b/c	10.81%
4+	example.com/a/b/c/d	7.02%

4.4 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 dataset are archived. 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.

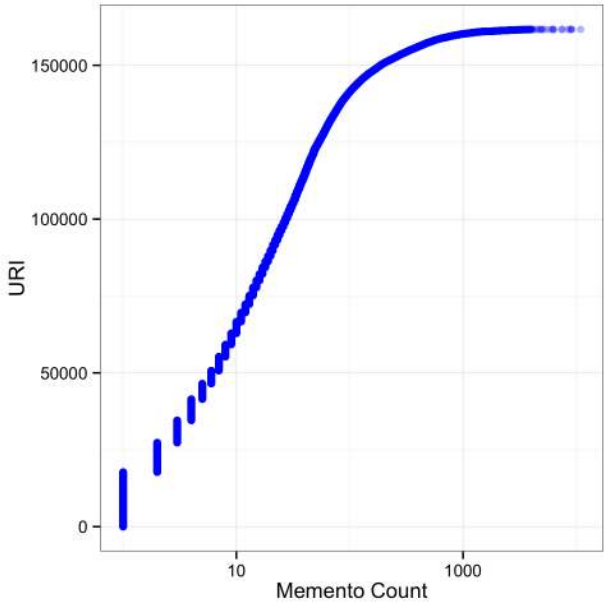

Figure 4: Memento count frequency

Table 7 lists the top 10 archived URI-Rs with the most mementos. 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, especially by the Internet Archive.

Table 7: Top 10 archived 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

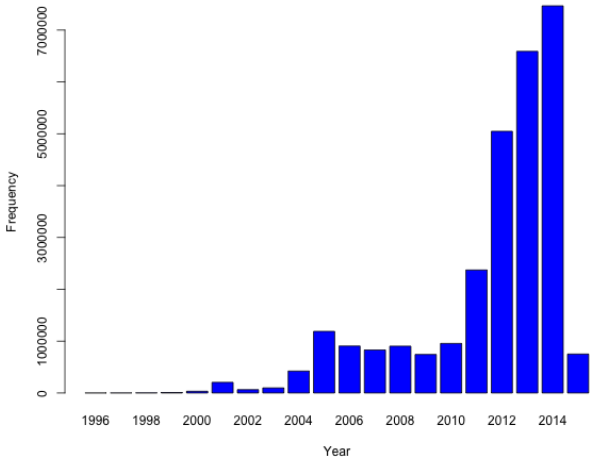

Figure 5: Number of URI-Ms in each year

Since the TimeMap identifies mementos present in multiple archives, here we present the breakdown of archives holding URIs in our Arabic dataset. Table 8 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 and Webcitation. We note that the percentages sum to greater than 100% because multiple archives can have mementos from the same original resource (URI-R).

Table 8: Archived URI-Rs present in all archives

Archive	Percent
Internet Archive	97.04%
Archive.today	6.58%
Webcitation	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%
Total	118.86%

Next, we want to know the breakdown of the archives for all mementos (URI-M) in our data set. Table 9 shows the percentage of archived mementos that each archive holds. We found that almost 73% were in the Internet Archive and 21% were in Archive-It.

Table 9: Archived URI-Ms present in all archives

Archive	Percent
Internet Archive	72.87%
Archive-It	21.26%
Archive.today	2.14%
Webcitation	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%
Total	100.00%

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 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 48 days. Values less than 1 indicate that the URI is archived multiple times per day on average.

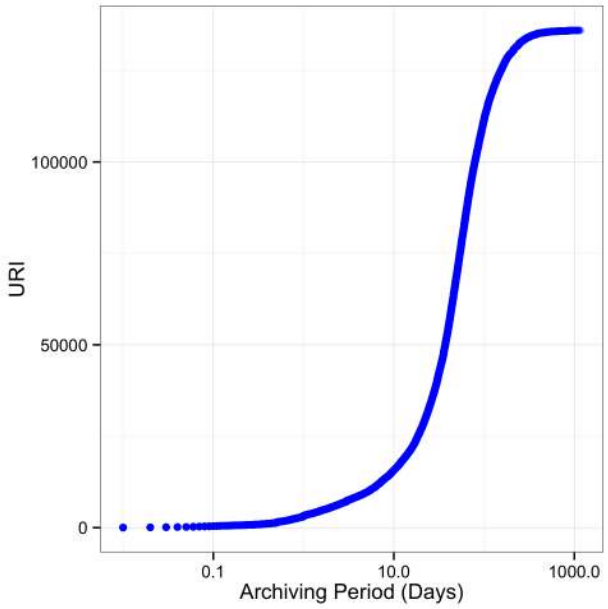

Figure 6: Average archiving period (days)

4.5 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 webpage was created in 2000, we would expect to see several mementos in the archives. However, if the webpage 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 webpage, but there have been several methods proposed to estimate this. We use CarbonDate [18], 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 data set. 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.

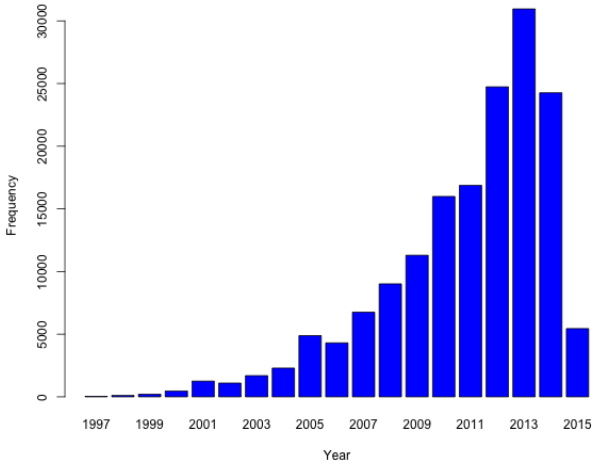

Figure 7: Creation dates for archived Arabic URIs

4.6 GeoIP Location

Earlier we looked at the ccTLD of the URIs to help determine where the hosts of the webpages might be located. Now we want to look at the GeoIP location of the IP address of the unique hostnames. 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[9] database to determine location from the IP address. Which tests at 99.8% accuracy at the country level[10].

We used this method to determine GeoIP for the Arabic URI dataset (300,646 URIs). We found that less than 11% of the URIs are hosted in Arabic countries. Table 10 shows the top GeoIP locations, with Arabic countries grouped together. Table 11 shows the top 5 GeoIP locations from Arabic countries. Overall, almost 58% of the Arabic seed URIs are hosted at IP addresses in the US. Other Western coun-

[9] `http://dev.maxmind.com/geoip/geoip2/geolite2/`
[10] `http://dev.maxmind.com/faq/how-accurate-are-the-geoip-databases/`

tries, including Germany and the Netherlands, host more of the Arabic seed URIs than does Saudi Arabia, the highest contributor of the Arabic countries.

Table 10: Top GeoIP locations

Country	Percent
US	57.97%
Arabic countries	10.53%
Germany	9.75%
Netherlands	5.29%
France	4.37%
Canada	3.31%
UK	3.07%
Others	5.71%

Table 11: Top 5 Arabic GeoIP locations

Country	Percent
Saudi Arabia	4.75%
Egypt	1.97%
Jordan	1.42%
Kuwait	0.71%
UAE	0.67%

4.7 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 [16].

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 12, 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.

Table 12: 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%

5. ANALYSIS

5.1 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. We used the creation date obtained in Section 4.5 and the date of the first memento obtained in Section 4.4.

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).

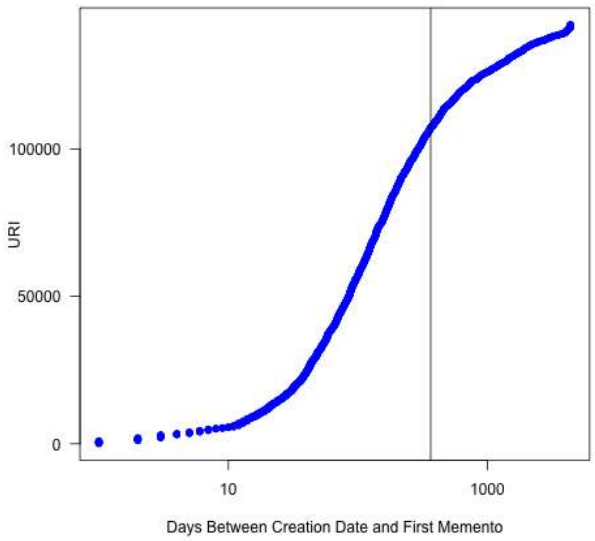

Figure 8: Difference between creation date and first memento

5.2 Archiving Based on Seed URI Source

Here we look at archiving based on seed URI source. As shown in Table 13, we found that 96% of DMOZ seed URIs are archived and that 45% of those from Raddadi and 42% from Star28 are archived. This was expected because DMOZ URIs are more likely to be found and archived [1,3]. DMOZ has historically been a source of seed URIs for indexing and archiving, at least as far back as 1999 [7].

5.3 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 14 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 detail count of GeoIP location, ccTLD, both, and neither of the archived Arabic set.

Table 13: Archiving and Indexing based on Arabic seed source

Name	Total	Arabic	Percent	Archived Count	Percent	Indexed Count	Percent
DMOZ	4,086	2,904	34.43%	2,774	95.52%	2,385	82.13%
Raddadi	3,271	1,677	19.88%	762	45.44%	1,104	65.83%
Star28	8,386	3,854	45.69%	1,601	41.54%	2,514	65.23%
Total	15,743	8,435					

Table 14: Archiving based on location and ccTLD

	Total	Percent	Archived Count	Percent
Arabic ccTLD	44,609	14.84%	12,532	28.09%
Arabic GeoIP	31,671	10.53%	4,152	13.11%
Arabic GeoIP and ccTLD	23,479	7.81%	13,969	59.50%
Neither	200,887	66.82%	131,025	65.22%

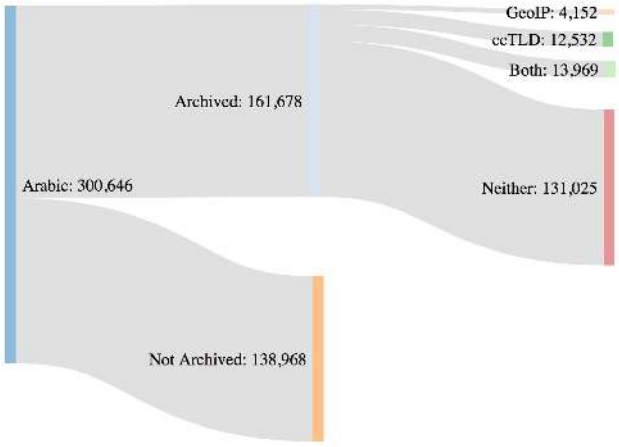

Figure 9: Arabic URIs count detail for Arabic GeoIP and ccTLD

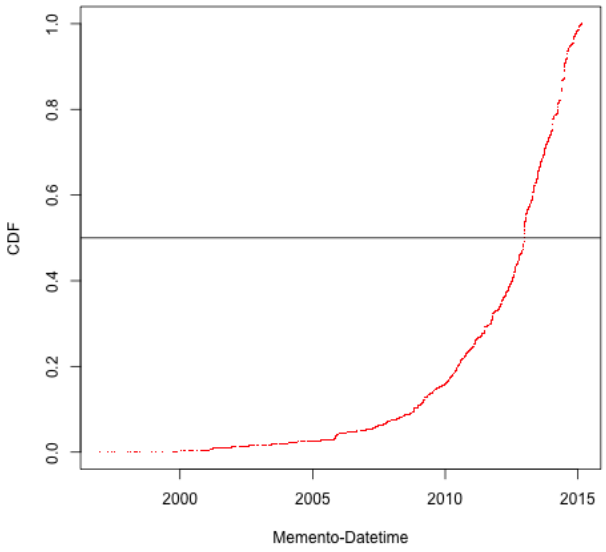

Figure 10: CDF of Memento-Datetimes both Arabic GeoIP and Arabic ccTLD

Next we wanted to statistically analyze the archived Arabic data set. 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. The means for the four groups (both Arabic ccTLD and Arabic GeoIP, Arabic GeoIP, Arabic ccTLD, and neither) were respectively, 0.5016, 0.5010, 0.5013 and 0.5005. To analyze these similarities further, we performed the Kolmogorov-Smirnov test to determine if the data sets 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 and for comparing Arabic GeoIP and neither is 0.017 and 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. 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).

5.4 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 in Table 15, 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.

Table 15: Archiving based on URI path depth

Path Depth	Total	Percent	Archived Count	Percent
0	52,011	17.30%	44,880	86.29%
1	121,521	40.42%	65,001	53.49%
2	73,507	24.45%	33,497	45.57%
3	32,499	10.81%	11,585	35.65%
4+	21,108	7.02%	6,715	31.82%

Table 16: Indexing based on location and ccTLD of Arabic seed URIs

	Total	Percent	Indexed Count	Percent
Arabic ccTLD	527	6.61%	401	76.09%
Arabic GeoIP	189	2.37%	139	73.54%
Arabic GeoIP and ccTLD	481	6.03%	410	85.24%
Neither	6,779	84.99%	4,548	67.09%

Table 17: Indexing based on URI path depth of Arabic seed URIs

Path Depth	Total	Percent	Indexed Count	Percent
0	6,863	86.05%	5,120	74.60%
1	776	9.77%	302	38.91%
2	297	3.72%	53	17.85%
3+	40	0.50%	23	57.5%

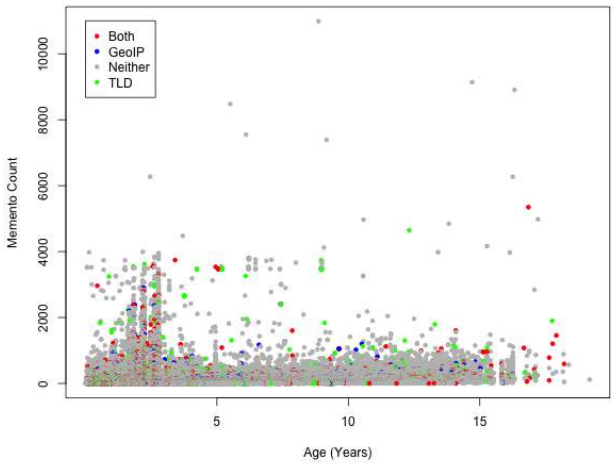

Figure 11: URI age and memento count

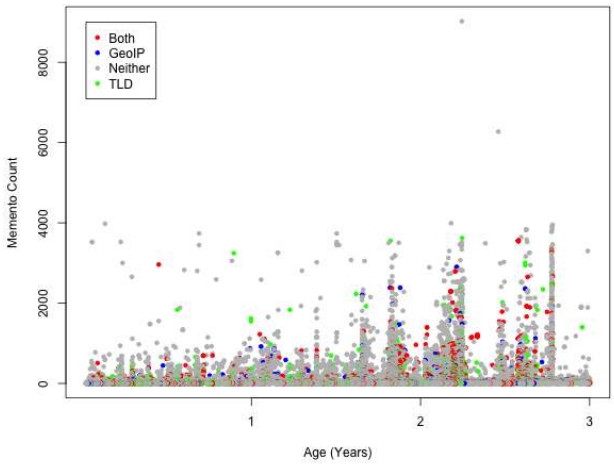

Figure 12: URI age less than three years and memento count

5.5 Indexing Based on Seed URI Source

Here we look at indexing based on seed URI source. As shown in Table 13, 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 [1, 3].

5.6 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. We note that, as in section 4.7, we only look at indexing for the Arabic seed URI set.

Table 16 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%).

5.7 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 17, we found that 74.6% of URIs with path depth 0 are indexed, and only 22.5% of the URIs with path depth of 3 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. CONCLUSIONS

In this study, we evaluated how well Arabic webpages are archived and indexed. First we collected webpages from Arabic directories, then determined if these webpages are written in the Arabic language. After that, we crawled the seed URIs to enlarge the dataset. Then we analyze those Arabic webpages. We used four different language tests to check the webpages language, then we performed some basic data analysis, such as checking the presence on the live Web, TLD, GeoIP, URI path depth, and creation date. Then we checked if these webpages are archived and measured the archiving frequency and the gap between creation date and the first archived version. Finally, we investigated if archiving and indexing were affected by Arabic ccTLD, Arabic GeoIP, URI path depth, or creation date.

We found that 46% of the Arabic websites are not archived and that 31% were not indexed by Google. We also found that archiving and indexing appear to be affected by the TLD, 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. URIs that are present in a directory are more likely to be indexed, and those present in the DMOZ directory are more likely to be archived. We also found that only 14.84% of the URIs had an Arabic ccTLD and only 10.53% had a GeoIP location in an Arabic country. Popular Western sites (such as facebook.com, wikipedia.org, and google.com) were in the top 10 domains found in our sample of Arabic language URIs. This seems to indicate 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.

Future work will include study of comparing archiving for other languages, such as Chinese, English, and other languages. In future work, we will check if the characteristics of the language, culture, and technology have an influence the archiving results.

7. REFERENCES

[1] S. G. Ainsworth, A. Alsum, H. SalahEldeen, M. C. Weigle, and M. L. Nelson. How Much of the Web is Archived? In *Proceedings of the IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, pages 133–136. ACM, 2011.

[2] A. Alarifi, M. Alghamdi, M. Zarour, B. Aloqail, H. Lraqibah, K. Alsadhan, and L. Alkwai. Estimating the Size of Arabic Indexed Web Content. *Scientific Research and Essays*, 7(28):2472–2483, 2012.

[3] A. AlSum. *Web Archive Services Framework for Tighter Integration Between the Past and Present Web*. PhD thesis, Old Dominion University, 2014.

[4] R. Baeza-Yates, C. Castillo, and E. N. Efthimiadis. Characterization of National Web Domains. *ACM Transactions on Internet Technology (TOIT)*, 7(2):9, 2007.

[5] K. R. Beesley. 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*, volume 47, page 54, 1988.

[6] B. E. Brewington and G. Cybenko. Keeping up with the changing web. *Computer*, 33(5):52–58, 2000.

[7] J. Cho and H. Garcia-Molina. Parallel crawlers. In *Proceedings of the 11th International Conference on World Wide Web*, pages 124–135, 2002.

[8] D. Gomes and M. J. Silva. Characterizing a National Community Web. *ACM Transactions on Internet Technology (TOIT)*, 5(3):508–531, 2005.

[9] H. C. Huurdeman, A. Ben-David, J. Kamps, T. Samar, and A. P. de Vries. Finding Pages on the Unarchived Web. In *Proceedings of the IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, pages 331–340. IEEE, 2014.

[10] Internet World Stats. Arabic Speaking Internet Users Statistics. `https://web.archive.org/web/20100515122707/http://www.internetworldstats.com/stats19.htm`, 2009.

[11] Internet World Stats. Arabic Speaking Internet Users Statistics. `https://web.archive.org/web/20141002202223/http://www.internetworldstats.com/stats19.htm`, 2014.

[12] Internet World Stats. Internet World Users By Language. `https://web.archive.org/web/20141213103739/http://www.internetworldstats.com/stats7.htm`, 2014.

[13] B. Kahle. Preserving The Internet. *Scientific American*, 276(3):82–83, 1997.

[14] M. Klein, H. Van de Sompel, R. Sanderson, H. Shankar, L. Balakireva, K. Zhou, and R. Tobin. Scholarly context not found: One in five articles suffers from reference rot. *PLoS ONE*, 9(12):e115253, 2014.

[15] J. Lepore. The Cobweb: Can the Internet be Archived? *The New Yorker*, January 2015.

[16] F. McCown and M. L. Nelson. Agreeing To Disagree: Search Engines And Their Public Interfaces. In *Proceedings of the IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, pages 309–318. ACM, 2007.

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

[18] H. M. SalahEldeen and M. L. Nelson. Carbon Dating the Web: Estimating the Age of Web Resources. In *Proceedings of the Temporal Web Analytics Workshop (TempWeb)*, pages 1075–1082, 2013.

[19] H. M. Salaheldeen and M. L. Nelson. Resurrecting my revolution. In *Research and Advanced Technology for Digital Libraries*, pages 333–345. Springer, 2013.

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

[21] H. Van de Sompel, M. L. Nelson, and R. Sanderson. HTTP framework for time-based access to resource states – Memento, Internet RFC 7089. 2013.

[22] H. Van de Sompel, M. L. Nelson, R. Sanderson, L. L. Balakireva, S. Ainsworth, and H. Shankar. Memento: Time Travel for the Web. Technical Report arXiv:0911.1112, 2009.