Global Proceedings Repository
American Research Foundation

ISSN 2476-017X

Available online at http://proceedings.sriweb.org

شبكة المؤتمرات العربية

http://arab.kmshare.net/

The 11th International Scientific Conference
Under the Title
“The role of humanities, social and natural sciences in supporting
sustainable development”

المؤتمر العلمي الدولي الحادي عشر

تحت عنوان "دور العلوم الانسانية والاجتماعية والطبيعية في دعم التنمية المستدامة"

9-10 ديسمبر 2020 - اسطنبول – تركيا

http://kmshare.net/isac2020/

Going Back in Time to Find What Existed on the Web and How much has been Preserved: How much of Palestinian Web has been Archived?

Thaer Sammar[1], Hadi Khalilia[1]

Palestine Technical University, Tulkarm, West Bank
thaer.sammar,h.khalilia@ptuk.edu.ps
http://www.ptuk.edu.ps

Abstract: The web is an important resource for publishing and sharing content. The main characteristic of the web is its volatility. Content is added, updated, and deleted all the time. Therefore, many national and international institutes started crawling and archiving the content of the web. The main focus of national institutes is to archive the web related to their country heritage, for example, the National Library of the Netherlands is focusing on archiving website that are of value to the Dutch heritage. However, there are still countries that haven’t taken the action to archive their web, which will result in loosing and having a gap in the knowledge. In this research, we focus on shedding the light on the Palestinian web. Precisely, how much of the Palestinian web has been archived. First, we create a list of Palestinian hosts that were on the web. For that we queried Google index exploiting the time range filter in order to get hosts overtime. We collected in 98 hosts in average in 5-years granularity from the year 1990 to 2019. We also obtained Palestinian hosts from the DMOZ directory. We collected 188 hosts. Second, we investigate the coverage of collected hosts in the Internet Archive and the Common-Crawl. We found that coverage of Google hosts in the Internet Archive ranges from 0% to 89% from oldest to newest time-granularity. The coverage of DMOZ hosts was 96%. The coverage of Google hosts in the Common-Crawl 57.1% to 74.3, while the coverage of DMOZ hosts in the Common-Crawl was in average 25% in all crawls. We found that even the host is covered in Internet Archive and Common-Crawl, the lifespan and the number of archived versions are low.

Keywords: Web Archive · Web Crawling · Digital Preservation.

1 Introduction

The World Wild Web (or simply the Web) is very dynamic. Content on the web is added, removed, and updated all the time. The life time of web pages is short. In 2004, Ntoulas et al. showed in a study of Web nature of web crawls that 80% of web pages is not accessible after one year [16]. Other Studies showed that the life time of web pages is short between 44 to 100 days [9]. The web is an important resource for sharing, publishing, and accessing information. The volatility characteristic of the web makes it risky to lose the knowledge on the web. Therefore, many national and international organizations and institutes realized the importance of archiving the web. The largest and the oldest web archive is the Internet Archive [3], the Internet Archive is archiving the web since 1996 from the entire web, and they are making their archive publicly accessible to the public mainly using URL search. In 2003, twelve institutes met at the National Library of France (BnF) to establish the International Internet Preservation Consortium (IIPC) [2]. The mission of IIPC is to develop tools for archiving the web, and promote the access and use of web archives. The awareness of archiving the web is increasingly getting attention from national institutes. Currently, there are around 55 members of IIPC[1]. However, with all the effort to archive the web, there is part of the web has not been archived, which means it has been lost forever. There are still part of the world not taking the action to preserve their web. In a study to evaluate how well Arabic web pages are archived and indexed [7], they found that 46% of the Arabic websites are not archived and that 31% were not indexed by Google. In Summary, there web pages that have never been archived; no action has been to taken to archive them. Even the part of web that has been archived is not complete. There are two main strategies for archiving followed by web crawlers; the bread-first strategy and the depth-first strategy. The bread-first focuses on archiving as much of the web as possible but bot in depth. While the depth-first strategy focuses on selected websites with goal to crawl and archive them as deep and complete as possible. Even the selection-based archive is not complete. A recent study of the Dutch web archive showed that there is high percentage of web pages that has references to them in the archive web pages have not been archived and in order to recover them they suggest to use links and anchor text in the archived web pages [13,17]. In this research, we investigate how much of the Palestinian web has been archived. There is no national institute taking the responsibility to archive the Palestinian web. Therefore, we rely on international archive that archive the entire web to investigate their coverage of the Palestinian web.

http://netpreserve.org/about-us/members/ [1]

In this research, we investigate the following research questions:

RQ1 Can we find websites that existed in the past?

The web is very dynamic, what we find on the Internet represent the current live web. We try to go back in time and find list websites that belong to the Palestinian web. To answer this question, we use two resources; the Google Index and the DMOZ open source web directory.

RQ2 How much of the Palestinian web has been archived?

To answer this research question we rely on the Internet Archive and the Common-Crawl to find their coverage of the Palestinian web. After creating a list of websites from the Palestinian web we examined how well these websites are represented in the Internet Archive and the Common-Crawl. The remainder of the paper is organized as follows. After discussing related work (Section 2). We discuss the experimental setup in detail in (Section 3) and answer research questions RQ1-2 in Sections 4. Finally, we discuss conclusions drawn from our findings (Section 5).

2 Related Works

Archiving the web is done with the help of web crawlers. Usually, there are two main strategies followed by the crawlers; the breadth-first strategy, and depth first strategy. The breadth-first strategy aims at archiving as much as possible (usually) from a specific top level domains. However, websites included in the process are not crawled deeply; the crawler does not follow subpages deeply. The other strategy is the depth-first strategy which focuses on specific selected websites (called crawler's seeds) with goal to crawl them as deep as possible. However, following this approach will exclude websites outside the crawler's seeds. Therefore, both strategies result in having incomplete archive. Different studies showed that web archives are incomplete. Web archiving crawlers often times fail to capture some contents from web pages such as Flash, JavaScript and other content [11,12,14]. The coverage and completeness of archives got the attention from many researchers. Web historian Brügger argued that almost every web archive is incomplete and it is hard to determine what is missing [10]. The missing information from Web Archive was classified into three levels as described by Brügger [15]. The first level is missing elements such as images, sounds and videos due to technical limitations at crawling time. The second level is missing entire web pages from web archive which might happen because of the crawling strategy. The third is missing information that were available when the archived content was online on the Web such as search engine results, queries, or open directories that provide statistics about the web, for example the Open Directory project of the web (DMOZ) [5]. In 2010, Ainsworth et al. [6] tried to answer the question how much of the web has been archived? They took a sample URLs from four differed resources; DMOZ, Delicious, Bitly, and indexes of search engine, and investigated their coverage in public web archives. They found that the archived percentage ranges from 16% to 79%. The same study was carried in 2013 [8], they found that the percentages of archival coverage percentage increases to range from 33% to 95%. The previous studies focused on the archive coverage of entire web.

More recent study focused on the archival percentage of Arabic web pages [7]. They found that 46% of the Arabic websites are not archived and that 31% were not indexed by Google.

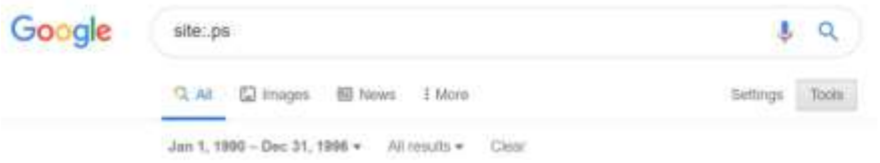
3 Setup

In this research we investigate how much of the Palestinian Web Archive has been archived. Based to our knowledge, no national institute nor organization is taking the action of archiving the Palestinian Web. Therefore, we shed the light on the Palestinian Web and how much of that has been archived. First, we need a list of websites that belong to the Palestinian domain. For that, we lookup into commercial Search Engines such as Google and from DMOZ. Second, we need to find what can be found in the international archives and crawls, for that we focus on the Internet Archive [3] and Common-Crawl [4]. The Internet Archive is the largest Web Archive world-wide, started archiving the Web since 1996, and the Common-Crawl collection contains crawls from the Web since 2009.

3.1 List of Websites

This task requires going back in time to find which websites where online which not possible. For this task, we used two resources; the first resource is the Google index and the second resource is DMOZ directory.

Google: We rely on Google index to find websites that were available on the internet using the time range filter of Google and as a query we run ***site:.ps***, this will give us the Web pages that are from the *.ps* domain. With this approach we miss websites that belong to the Palestinian web but not from .ps domain such as .com, .org, .edu, etc. We ran the query site:.ps with different time range using 5 years range starting from 1990. Figure 1 shows how to query Google for web pages from the .ps domain in the time range from first of January 1990 to end of December 1994.


site:.ps

Jan 1, 1990 – Dec 31, 1996

All results

Fig. 1: Querying Google using time range filter

DMOZ: The second resource that we used to collect our websites is the open directory project (ODP) also known as DMOZ [5]. DMOZ is an open-content directory for classification and categorization of web pages world-wide into categories based on topics. DMOZ provides an on-line service provides for browsing Web resources categories in hierarchically organized directory. We collected Palestinian web sites from DMOZ in two approaches; first manually from the DMOZ websites, and from the dataset of domains obtained from 2016 DMOZ dump [18]. In total we obtained 188 unique hosts.

Top Level Domains (TLD): URLs collected from the Google index are all from the .ps top level domain. Of course there are other web pages that are from the Palestinian web but use other Top Level Domains such .org, .com. Here, we investigate the distribution of TLDs in the Palestinian web hosts collected from DMOZ. Around 30% of hosts are from the .com TLD, which is the most popular TLD. Second top TLD is the .ps TLD. Table shows the distribution of TLD of hosts collected from DMOZ directory.

Table 1: Distribution of TLDs of Palestinian hosts from DMOZ.

TLD	Count	Percentage
Com	57	30.32%
ps	56	29.79%
org	43	22.87%
edu	12	6.38%
net	12	6.38%
uk	2	1.06%
eu	2	1.06%
gov	1	0.53%
ca	1	0.53%
il	1	0.53%
int	1	0.53%
Total	188	

3.2 Web Archive/Crawls Collections

We are not aware of any Palestinian Web Archive collection. Therefore, we relied on two large-scale Web Archive collections; namely the Internet Archive and the Common-Crawl.

Internet Archive is a non-profit foundation. The Internet Archive has been archiving web pages world-wide. The Internet Archive started since 1996, which makes it the largest and oldest Web Archive. The goal of the Internet Archive is to build a library of the web and make it accessible for the public. The Internet Archive makes their collection available via the Wayback Machine [4]. Given the URL, the Internet Archive Wayback machine returns a calendar of all crawls for that URL overtime (see Figure 2).

Explore more than 418 billion web pages saved over time

Fig. 2: The Internet Archive Wayback Machine URL-based search, screenshot March 6 2020 at 06:07 AM

Common Crawl: is a non-profit organization that crawls the web and makes it freely available to the public. We used the Common-Crawl URL CDX index [1] of the crawls in the period between 2009 and 2020 (excluding 2011 crawls (the index is not available)). Given the query, the Common-Crawl index API returns pages that are found for that query in the format (”urlkey”: , ”timestamp”:, ”digest”:, ”offset”:, ”length”:, ”filename”:, ”url”:, ”status”:, ”mime”:). Table 2 summarizes number of web pages from the .ps domain that have been archived by Common-Crawl.

Table 2: Number of crawls per year in Common-Crawl index.

Year	Number of crawls
2009	1
2010	1
2012	1
2013	2
2014	8
2015	10
2016	9
2017	12
2018	12
2019	12
2020	1 (by the time of this study)

4 Results & Discussion

4.1 Find Websites from the Past

In this section, we summarize how we collected the list of websites. In Section 3.1, we described how to query Google to obtain web pages that are from the .ps domain. We exploited the time range query filter of Google search to retrieve results in the 5−years granularity, we extracted the URLs from Google search results and for these URLs, we got the host websites. Table 3 presents number of unique hosts in the corresponding time ranges. As we can see from the table, number of web hosts increases over the years. In the period until end of 1994, Google index has only 20 unique hosts. Currently, we found 168 unique hosts from the .ps domain. Google search page says that there are 8,820,000 hits. However, after result number 172, Google shows a message indicating that the remaining results were omitted because they are similar to the first 172 hits.

Table 3: Websites From Google Index Overtime.

Time Range	Number of Hosts
current	168
01/01/1990 - 31/12/1994	20
01/01/1990 - 31/12/1999	60
01/01/1990 - 31/12/2004	73
01/01/1990 - 31/12/2009	140
01/01/1990 - 31/12/2014	148
01/01/1990 - 31/12/2019	147

4.2 Coverage in the Common Crawl

In this section, we explore the coverage of the Palestinian web in the Common-Crawl. We queried the Common-Crawl URL index of each crawl to get all URLs that are included in the Common-Crawl from the .ps domain using the query (*.ps). The Common-Crawl dataset has several crawls per year. Therefore, we aggregated founded pages by year. Table 4 summarizes number of web pages and number of Hosts in the Common-Crawl datasets in the years from 2009 to 2020 (excluding 2011). From all years, there are 2,997 unique hosts that belong to the .ps domain. In total, there 240,071 archived web pages from all hosts, in average there are 80 pages per host. Table 5 shows number of pages for the top hosts. All these hosts are still alive except http://6ollap.ps/.

Table 4: Number of Palestinian's Web pages and Web hosts found in the Common-Crawl overtime.

Year	Number of Unique Web Pages	Number of Versions	Number of Unique Hosts
2009	2,453	2,468	141
2010	6,167	6,787	257
2012	13,797	14,462	534
2013	13, 797	14,462	534
2014	9,580	84,506	1,476
2015	7,709	61,190	1,351
2016	18,021	78,395	1,996
2017	55,188	159,898	1,166
2018	12,105	12,410	79
2019	97,575	160,829	215
2020	3,679	3,760	49

Table 5: Top Palestinian hosts based on number of archived versions overtime.

Web Host	Number of crawls (years)
aten.com.ps	1
clark.ps	1
doorste.ps	1
transfermarkt.ps	1
aten.com	2
yam.ps	2
hcc.plo.ps	2
share.sni.ps	2
siap.ps	2
soco.ps	2

Number of Versions: In total, there are 240,071 unique web pages, and 599,167 snapshots from these Web pages. On average, there is 2.5 versions for each web page in the whole Common-Crawl collection of 11-years span. In Table 2, we showed number of crawls per year. Therefore, some pages are crawled several times in the same year. This is clear if we look at the average number of versions for each year. Therefore, we look next into the life span of a web page in terms of years.

Oldest Hosts: The oldest crawl is 2009. We focus on the list of hosts that have been crawled in 2009. Out of the 2,468 found in 2009, only 9 hosts were crawled in 2010. Out of these 9 hosts only 4 were crawled in 2012, 2013; {computerland.ps, lamasat.com.ps, softech.com.ps, yellowpages.com.ps}. Only one host of them was crawled in years after that until 2020.

Life Span per web page: As mentioned previously, there are 240,071 unique web pages (URLs) in the crawls from 2009 to 2020. By examining the number of years (crawls) per web page, we found that in average a web page has 1.25 versions. Recall that we aggregated result obtained from the Common-Crawl index by year granularity. The maximum number of version was 9; only one page has 9 versions (exist in 9 years of Common-Crawl crawls). Table 6 shows frequency of each version count, the majority of web pages have one version in period from 2009 to 2020, precisely 80.5% of web pages has one version.

Table 6: Frequency of number of years.

Number of Versions	Frequency
1	154,479
2	29,075
3	6,632
4	1,466
5	142
6	113
7	35
8	5
9	1

Coverage of Google URLs in Common Crawl: Now, we investigate the coverage of websites that were found in Google index in the Common-Crawl. In Section 3.1, we described how we collected websites from the Palestinian domain using Google index using the query ***site:.ps*** with time range filter. By increasing the time window filter when searching, the number of hosts increases, see the column (Number of Hosts) in Table 7. Here, we examine how much of these websites were crawled by Common-Crawl. First, we look at how much of the websites exist in the Common-Crawl index on yearly basis; the website is considered exist if there is at least one web page from that website in one of the yearly crawls. The highest percentage of websites was for websites collected from Google index in the period until end of 2009. Table 7 describes found web hosts in terms of the number of crawled hosts, percentage of coverage ***(Number of crawled hosts / total number of hosts)***, and number of crawled web pages from all crawled web hosts.

Table 7: Summary of Common-Crawl coverage of hosts collected from Google.

Time Range	Number of Hosts	Number of crawled hosts	Percentage of crawled hosts	Number of crawled pages
current	168	102	63.8%	138,554
01/01/1990 - 31/12/1994	20	14	70%	8,881
01/01/1990 - 31/12/1999	60	44	70.3%	23,156
01/01/1990 - 31/12/2004	73	48	65.8%	35,401
01/01/1990 - 31/12/2009	140	104	74.3%	58,370
01/01/1990 - 31/12/2014	148	85	57.4%	57,152
01/01/1990 - 31/12/2019	147	84	57.1%	57,150

From the list of the past websites that we managed to collect from Google index, the percentage of crawled hosts ranges between 57.1% to 74.3%. The number of crawled web pages from all hosts in the corresponding time window is low. For example, for websites that existed in Google index by end of 2009 was 393 from all hosts (140). If we look at the oldest hosts that were on the web by end of 1994, out of the 20 hosts, 14 hosts have web pages in the Common-Crawl. Total number of web pages from these hosts is 54 pages. These 14 web hosts can be categorized as political parties' websites, and national organization. Two hosts are not alive; they don't exist on the web. Even if the host exist in the Common-Crawl, in most cases it exists in few crawls (year-granularity).

Coverage of DMOZ URLs in Common Crawl: Now, we investigate how much of the collected hosts from DMOZ have been crawled by Common-Crawl over the years. At most 23.4% of Palestinian hosts were crawled in 2016. In total, from all years of Common-Crawl, only 25% of DMOZ hosts were found in Common-Crawl. Table summarizes the coverage of DMOZ hosts in Common-Crawl over the years. The oldest crawl from Common-Crawl dates back to 2009, around 5.32% of DMOZ Palestinian hosts were covered in this year. In total, there are 279 crawled pages from all covered hosts. The question is how frequently does Common-Crawl crawled these hosts overtime. In different words, are there crawled pages from hosts covered in 2009 in later years. In order to investigate this, we checked the overlap between hosts covered in 2009 and other years. Table summarizes the percentage of overlap. The overlap ranges between 0% to 100%. For example, the overlap with 2011,2018,2019,2020 is 0%, while the overlap with 2010 is 10%; this means that 10% of hosts crawled in 2009 are also crawled in 2010.

Table 8: Common-Crawl coverage of DMOZ hosts over the years.

Year	Count	Percentage
2009	10	5.32%
2010	7	3.72%
2012	11	5.85%
2013	11	5.85%
2014	37	19.68%
2015	38	20.21%
2016	44	23.40%
2017	20	10.64%
2018	4	2.13%
2019	7	3.72%
2020	1	0.53%
All Years	47	25.00%

4.3 Coverage in the Internet Archive

In the previous section, we explored the coverage of Palestinian web in the Common-Crawl. In this section, we explore the coverage of the Palestinian web in the Internet Archive which is the oldest and the largest web archive targeting the entire Web. The Internet Archive makes its archive publicly available via URL-search of the Wayback machine. Given the URL, the Internet Archive return the history of that URL, all snapshots available for that URL in a calendar view. We used the Internet Archive API to get the archival history of each website in the list that we created for Palestinian Web using Google index and DMOZ (as described in Section 3.1).

Coverage of Google URLs in the Internet Archive: We queried the Internet Archive Wayback machine of each host to get all archived pages that are included in the Internet Archive for the Palestinian hosts that are collected from Google engine. Recall that we collected websites from Google index overtime. For each, period of time; 5-years granularity from 1990 to 2020, we investigate the coverage of websites in these time granularity in the Internet Archive. The coverage ranges between 0% to 93%, from old to new. The coverage of old websites is very low. For example, the coverage of websites that are younger than 2000 is 0%. Table 9 shows the archival percentage coverage of Palestinian websites that were collected from Google index in the Internet Archive over time.

Table 9: Summary of Internet Archive coverage of hosts collected from Google.

Period	Number of Palestinian Hosts from Google	Number of Palestinian Hosts Archived in Internet Archive	Percentage
1990-01-01 to 1994-12-31	20	0	0%
1990-01-01 to 1999-12-31	60	0	0%
1990-01-01 to 2004-12-31	73	39	53%
1990-01-01 to 2009-12-31	140	101	72%
1990-01-01 to 2014-12-31	149	106	71%
1990-01-01 to 2019-12-31	147	131	89%
ALL	398	373	93%

In average, the coverage of all websites collected from Google index in all time granularities is 57%. For websites found in Google index in recent years, the coverage is high. For example, as you see in Table 9, the coverage of websites found in Google index in the period until end of 2019 is 89%. In addition to websites collected using the time range filter from Google index, we also collected websites without any time filter to get what google has currently, the coverage of these websites is 93%, which you can see in the row (ALL). We considered the host covered in the Internet Archive if there is at least one archived version of any page from that host. However, in terms of number archived versions, the most archived host has 193 archived pages. Table 10. The average number of archived pages per hosts is 57.1%.

Table 10: Top 10 hosts from Google archived in the Internet Archive.

Hosts	Number of archived Versions
http://pchrgaza.org/	193
http://aten.com/	184
http://english.wafa.ps/	176
http://jawwal.ps/	162
http://p p.ps/	158
http://mosa.gov.ps/	151
http://pcbs.gov.ps/	150
http://hadara.ps/	148
http://popchips.com/	147
http://paltoday.ps/	146

Life span of old websites: We showed the coverage of websites collected from Google in the Internet Archive, the coverage means that there are archived pages from that website in the Internet Archive. Here, we are interested in the lifespan of the archived websites. We focus on the oldest websites, we take websites found in Google index by end of 2004, because the coverage of the period until end of 1994 and the period until end of 1999 was 0%. We measure this by counting how many years do these websites exist in the Internet Archive. Table 11 shows the top hosts in terms of their lifespan existence in the Internet Archive. For example, the pcbs.gov.ps has archived pages that span 18 years. In average, we found that the lifespan of hosts archived by the Internet Archive of the Palestinian hosts is 10.4; such that in average archived versions were found in 10 years.

Table 11: Top 10 hosts that exist in Google index in the period from 01-01-1990 to 31-12-2004 in terms of their lifespan in the Internet Archive.

Host Name	Lifespan(number of years)
nbprs.ps	14
palestinecabinet.gov.ps	14
palpolice.ps	14
tirawi.ps	14
mnofal.ps	15
saraya.ps	15
p p.ps	16
english.wafa.ps	17
shabab.ps	17
pcbs.gov.ps	18

Coverage of DMOZ hosts in the Internet Archive: Now, we explore the coverage of Palestinian hosts collected from the DMOZ directory. These hosts were collected manually from the online DMOZ website. Therefore, we checked their coverage without time range split. The coverage was high for DMOZ hosts in the Internet Archive, 96% of hosts has archived pages in the Internet Archive. We consider the host covered in the Internet Archive if it has at least one archived page. Looking at the lifespan of hosts as we did for hosts from Google index. We find that there are some hosts from DMOZ that span longer years. For example, the birzeit.edu.ps which is a Palestinian university website exist in Internet Archive for 25 years. Table 12 shows the top hosts in terms of number of years of their lifespan. In average the lifespan of DMOZ Palestinian hosts archived by Internet Archive is 15.9. Also number of versions of archived hosts in the Internet Archive was higher for Palestinian hosts from DMOZ, see Table 13. In average, the number of archived versions is 111.4.

Table 12: Top hosts from DMOZ in terms of their lifespan in Internet Archive.

Host Name	Lifespan(number of year)
http://nablus.org/	23
http://najah.edu/	23
http://arij.org/	24
http://baraka.org/	24
http://padico.com/	24
http://zaytona.com/	24
http://amin.org/	25
http://bethlehem.edu/	25
http://birzeit.edu/	25
http://paltrade.org/	25

Table 13: Top 10 hosts from DMOZ covered in the Internet Archive.

Host	Number of Archived Versions
http://mof.gov.ps/	275
http://angelfire.com/	249
http://worldbank.org/	246
http://who.int/	240
http://birzeit.edu/	227
http://abyznewslinks.com/	222
http://amin.org/	220
http://dur.ac.uk/	220
http://bethlehem.edu/	220
http://arij.org/	213

5 Conclusions

In this research, we explore the coverage of Palestinian web in the international archives. Knowing what exist in the past on the web is not an easy task. The archiving of the national Palestinian web is not taking by any national institute.

Therefore, our work in this paper consists of two parts. First find hosts from the Palestinian web overtime, and second check their coverage in the Internet Archive and Common-Crawl. First, finding Palestinian websites that were on the web in addition to what exist currently. For that, we queried Google index to retrieve websites from the .ps TLD using the time range filter. We run the same query several times by changing the time range based on 5-year granularity starting from 1990 to 2020. For example, in the period from 1990 to end of 1994, we found 20 hosts in Google index. This number increases as we expand the time range filter. For example, number of hosts found by end of 1999 was 60. The second source that we used for collecting Palestinian hosts was DMOZ, we collected 188 hosts from DMOZ directory. These Palestinian hosts are from different Top Level Domains, almost 30.32% from the .com TLD, and 29.79% from .ps TLD. Second, explore the coverage of collected hosts in the Internet Archive and Common-Crawl. We found that the coverage of Internet Archive of Google hosts ranges from 0% to 89%; from oldest to newest hosts. While the coverage of Google hosts in Common-Crawl ranges from 57% to 74.3%. The coverage of DMOZ hosts in Common-Crawl ranges from 0.53% to 23.4% over time from 2009 to 2020, while the coverage of DMOZ hosts is around 96%. We found the lifespan (age of archived versions in terms of years in the archive) of archived hosts vary among the two sources; Google and DMOZ and is different between Internet Archive and Common-Crawl. The lifespan of hosts in the Internet Archive was 10.4, while the lifespan of DMOZ hosts in the Internet Archive is 15.9. We found that number of archived versions is low. For example, in average the number of versions of Google hosts in the Internet Archive is 57.1.

References

1. Common crawl url index. url (http://index.commoncrawl.org/).

2. International internet preservation consortium (iipc).http://www.netpreserve.org.

3. Internet archive (https://archive.org/).

4. Internet archive wayback machine. url (https://archive.org/web/).

5. The open directory project (https://dmoz-odp.org).

6. S. Ainsworth, A. Alsum, H. SalahEldeen, M. C. Weigle, and M. L. Nelson. How much of the web is archived? CoRR, abs/1212.6177, 2012.

7. L. Alkwai, M. Nelson, and M. Weigle. How well are arabic websites archived? pages 223–232, 06 2015. 8. A. AlSum. Web Archive Services Framework for Tighter Integration Between the Past and Present Web. PhD thesis, Old Dominion University, 2014.

9. B. Brewington and G. Cybenko. Keeping up with the changing web. Computer, 33:52 – 58, 06 2000.

10. N. Bru¨gger. Web history and the web as a historical source. Zeithistorische Forschungen - Studies in Contemporary History, 9(2):316 – 325, 2012.

11. M. J. Day. Preserving the fabric of our lives: A survey of web. In T. Koch and I. Sølvberg, editors, Research and Advanced Technology for Digital Libraries, 7th European Conference, ECDL, Trondheim, Norway, August 17-22, 2003.

12. H. Hockx-Yu. The past issue of the web. In D. D. Roure and M. S. Poole, editors, Web Science 2011, WebSci '11, Koblenz, Germany - June 15 - 17, 2011, pages 12:1–12:8. ACM, 2011.

13. H. C. Huurdeman, J. Kamps, T. Samar, A. P. de Vries, A. Ben-David, and R. A. Rogers. Lost but not forgotten: finding pages on the unarchived web. Int. J. on Digital Libraries, 16(3-4):247–265, 2015.

14. J. Masan`es. Web archiving. Springer, 2006.

15. F. Nanni, A. Ben-David, N. Bru¨gger, M. Dougherty, I. Milligan, and J. Winters. Web historiography - A new challenge for digital humanities? In M. Eder and J. Rybicki, editors, Digital Humanities 2016, DH 2016, Conference Abstracts, Jagiellonian University & Pedagogical University, Krakow, Poland, July 11-16, 2016, pages 74–76. Alliance of Digital Humanities Organizations (ADHO), 2016.

16. A. Ntoulas, J. Cho, and C. Olston. What's new on the web?: the evolution of the web from a search engine perspective. In S. I. Feldman, M. Uretsky, M. Najork, and C. E. Wills, editors, Proceedings of the 13th international conference on World Wide Web, WWW 2004, New York, NY, USA, May 17-20, 2004, pages 1–12. ACM.

17. T. Samar, H. C. Huurdeman, A. Ben-David, J. Kamps, and A. P. de Vries. Uncovering the unarchived web. In S. Geva, A. Trotman, P. Bruza, C. L. A. Clarke, and K. Ja¨rvelin, editors, The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014, pages 1199–1202. ACM, 2014.

18. G. Sood. parsed-domain.csv.7z. In Parsed DMOZ data. Harvard Dataverse, 2016.