

Arabic News Archival Rate is Catching Up to English

2025 IEEE Big Data Conference (IEEE BigData 2025) CAS workshop
2025-12-09

Hussam Aldeen Hallak

hhallak@cs.odu.edu [@hussamhallak.bsky.social](https://bsky.social/hussamhallak)

Michael Nelson

mln@cs.odu.edu [@phonedudemln.bsky.social](https://bsky.social/@phonedudemln)

WS-DL Research Group at Old Dominion University
Norfolk, Virginia, United States

[@WebSciDL.bsky.social](https://bsky.social/WebSciDL)

Introduction

- The importance of web archiving
- We focus on archiving news web pages
- We studied news web pages in English and Arabic
- Contrary to expectations: Arabic news archival rate is growing faster than English.
- The gap between Arabic and English archival rates is shrinking.
- The IA is essential for both, but it is more critical for Arabic.
- Urgent need for decentralized Arabic-focused web archives.

Aljazeera English 2025-12-08

<https://web.archive.org/web/20251208091743/https://www.aljazeera.com/>

web.archive.org/web/20251208091743/https://www.aljazeera.com/

Wayback Machine 361,662 captures 27 Nov 1999 - 9 Dec 2025

ALJAZEERA News Middle East Explained Opinion Sport Video More

LIVE Sign up

Trending Syria: A year after al-Assad Russia-Ukraine war Tracking Israeli ceasefire violations Sudan war Donald Trump Thailand-Cambodia clashes



'We will rebuild Syria': Country marks one year since fall of al-Assad

LIVE UPDATES



SYRIA AFTER AL-ASSAD

How many Syrians have returned home one year since the fall of al-Assad?

Hope, flags, fireworks as Syria starts to celebrate a year without al-Assad

The fragile fight for justice in a post-Assad Syria

EXPLAINER

Are Israel, Hamas entering the second phase of the ceasefire?

Sudan group accuses RSF of raping 19 women who fled el-Fasher

Japan summons China envoy over 'fighter jet radar'



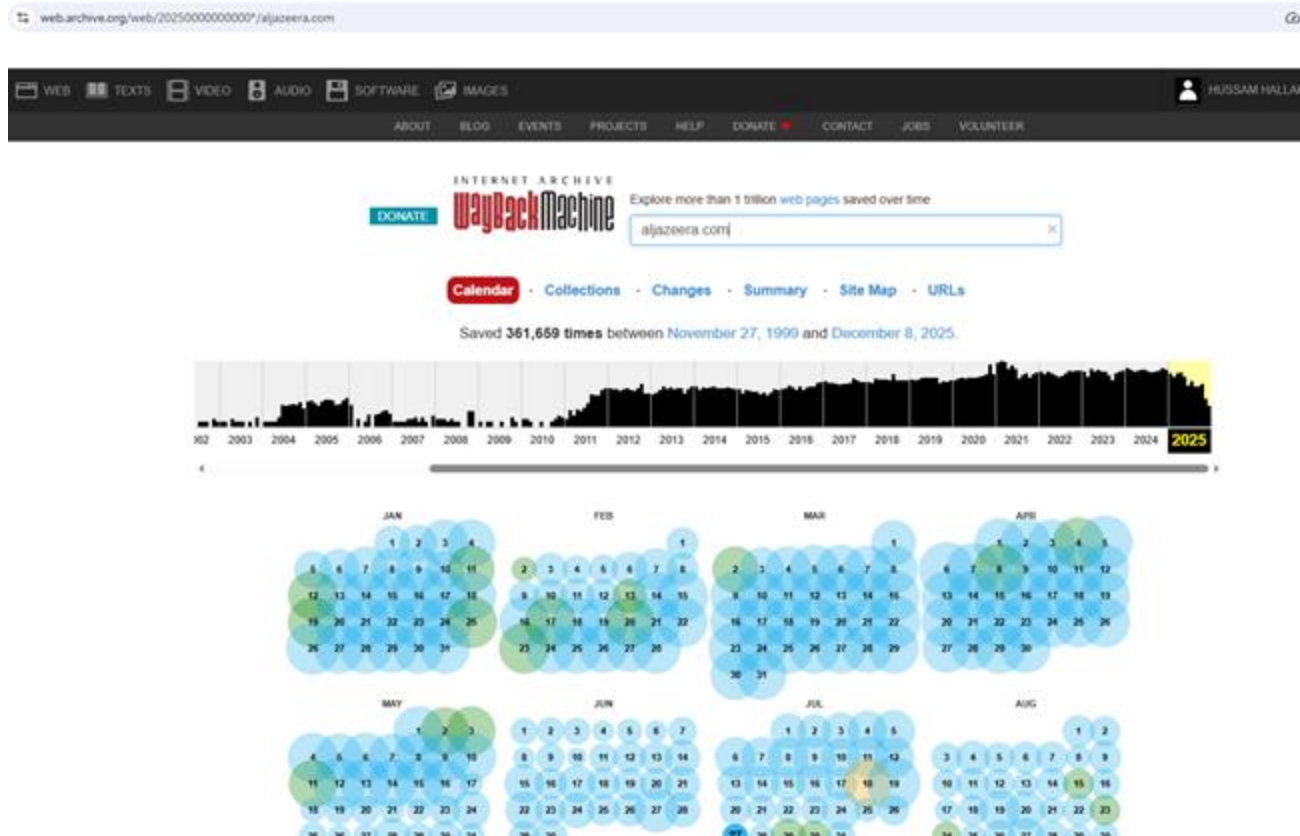
MORE HEADLINES

IndiGo chaos: Why is India's largest airline canceling hundreds of flights?

'Stop killing us': Huge crowds rally in Brazil, decrying rise in femicide

361,659 captures by the Wayback Machine

https://web.archive.org/web/20250000000000*/aljazeera.com



Domain Parked in 2002

<https://web.archive.org/web/20021201121245/http://www.aljazeera.com/>



web.archive.org/web/20021201121245/http://www.aljazeera.com/

Go NOV **DEC** JAN

361,662 captures

27 Nov 1999 - 9 Dec 2025

2001 **2002** 2003

About this capture

This domain is registered at DotRegistrar.com by a customer and **parked temporarily until the owner establishes a permanent site.** To contact the owner of the domain "aljazeera.com", please email . If you have any questions regarding our domain parking feature, please e-mail parking@dotregistrar.com.

EXPIRED DOMAIN

Try Searching the Internet for These Topics

[Arabic language](#)
[Middle east jobs](#)
[Arabic translation](#)

[Arabic](#)
[Television news](#)
[Cable tv](#)

Other Popular Categories

[Travel](#)
[Car Rental](#)
[Hotels](#)

[Health and Beauty](#)
[Skin Care Products](#)
[Exercise Equipment](#)

DotRegistrar
a division of iHoldings.com



Aljazeera.com was in Arabic (2003-07-18)

<https://web.archive.org/web/20030718162703/http://www.aljazeera.com/>



ADVERTISE HERE!

Over 12 Million Daily Hits - Promote your products globally.

sales@aljazeera.com

الرئيسية | مركز المساعدة | Help center | إلى الجزيرة | الصفحة الرئيسية
الأثنين 6/6/1424 هـ الموافق 4/8/2003 (آخر تحديث: الساعة 11:46 (مكة المكرمة) 8:46 (غرينتش))
كيف نقرأ العربية

آخر التطورات بالجويو شمالي زامبيا ... البرلمان في أذربيجان يعين



الأقصى: نتم الاحتفال باعتقال أحد كوادرها بعد هجوم القدس
اتهمت كتائب شهداء الأقصى إسرائيل باعتقال أحد كوادرها في مدينة



هل تعتقد أن اتهامات أميركا
للسعودية بتسويق جماعات
وصفتها بالتنظيم بالإرهابية
حقيقية؟



Totally archived page
(no missing resources)

Aljazeera.com is in English (2004-08-20)

<https://web.archive.org/web/20040820080353/http://www.aljazeera.com/>

INTERNET ARCHIVE
waybackmachine
http://www.aljazeera.com/
361,662 captures
27 Nov 1999 - 9 Dec 2025

Aljazeera.com™
International English Edition

SEARCH Google Aljazeera.com Go Make this my Homepage

Home Page
Middle East News
Let Talk - Dr.Kareem
Conspiracy Theories
Who's telling the truth
Poll of the Week
News Alerts
Email@Aljazeera.com

Login@Aljazeera.com
Email Address:
Password: Login
FREE Email Sign up
Password Reset

Global Poll
Do you think recent violence that stroke Najaf has dented the authority of interim Prime Minister Iyad Allawi?

Friday, August 20, 2004: 13:30 GMT Group threatens to kill Turkish hostage held in Iraq

LATEST HEADLINE

MIDDLE EAST BUSINESS

- Oil prices to reach \$49 U.S. as violence erupts in Iraq
- Oil prices hit all-time highs despite Iraqis peace deal
- Iranian economic reform falters
- Saudi prince wants oil prices at \$25-\$30

Middle East FOREX Markets per USD

Euro	1.1896	Egyptian Pound	6.1850
Pound Sterling	1.8289	Iraqi Dinar	1455.0
Bahraini Dinar	0.3770	Irani Riyal	8621.0
Kuwaiti Dinar	0.2947	Jordan Dinar	0.7080
Qatari Riyal	3.6403	Lebanese £	514.0
Saudi Riyal	3.4499	Sudan Dinar	248.80
UAE Dirham	3.6728	Syrian Pound	51450

U.S. renews attacks on Najaf
8/20/2004 7:49:00 AM
U.S. warplanes and tanks have been bombarding areas around the holy shrine of Imam Ali in the Iraqi city of Najaf.

Mostly archived page but missing a few images/icons

Documenting Events

<https://web.archive.org/web/20040829043439/http://www.aljazeera.com/>

The screenshot shows the Aljazeera.com website from the Internet Archive. The address bar displays the URL: <http://www.aljazeera.com/>. The page features the Aljazeera.com logo, a search bar, and a navigation menu. The main content area displays several news headlines. Red arrows point to specific headlines with labels indicating their archival status:

- Israel trial (not archived)**: Points to the headline "Israelis tried for murdering Palestinian civilian".
- Oil prices drop (archived)**: Points to the headline "Oil prices drop, Najaf peace deal struck".
- US elections (not archived)**: Points to the headline "China's military muscle stresses US interests in Asia".
- Iraq war (not archived)**: Points to the headline "New wave of violence hits Iraq, 5 killed".

Archived

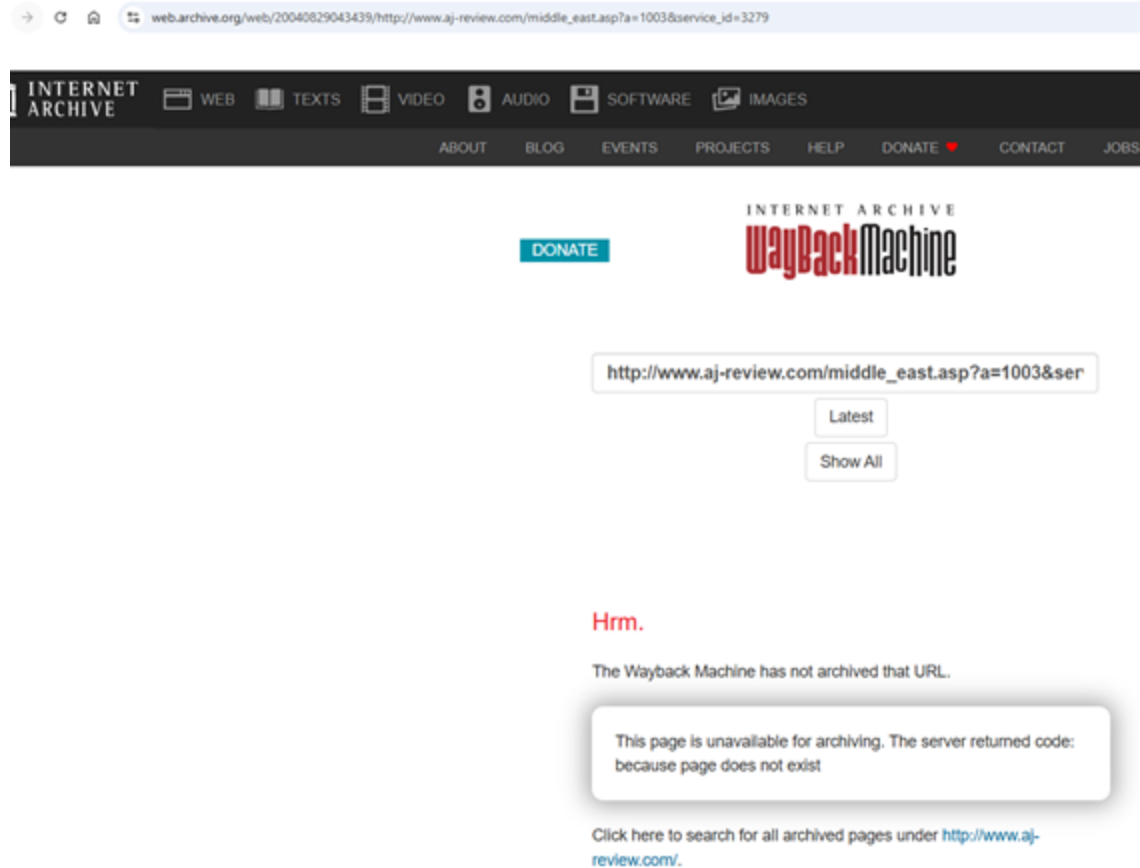
https://web.archive.org/web/20040909211824/http://www.aljazeera.com/cgi-bin/news_service/middle_east_full_story.asp?service_id=3224

The screenshot shows a web browser window with the address bar displaying the URL: http://www.aljazeera.com/cgi-bin/news_service/middle_east_full_story.asp?service_id=3224. The page is the International English Edition of Aljazeera.com, dated 9 Sep 2004. The main headline is "Oil prices drop, Najaf peace deal struck" with a sub-headline "8/27/2004 8:18:00 AM GMT" and "Source: Reuters". The article text states: "For the fifth day in a row, oil prices dropped on Thursday after word of a peace deal in Najaf raised hopes of more reliable oil shipments from the energy-rich country and encouraged a run of profit-taking by big money speculators." Below this, it mentions "U.S. crude futures settled down 37 cents at \$43.10 a barrel, bringing them down more than \$6 from a record high hit last week, when fears that world producers would not be". To the right of the article is a large advertisement for "Halo 2 LIMITED COLLECTOR'S EDITION" featuring a Spartan character and the text "FREE" and "click here". On the far right, there is a section titled "MIDDLE EAST" with a list of news items: "U.S. troops enter Samarra city in Iraq", "U.S. assaults in Falluja kill 38 Iraqis", "Israeli forces kill 4 Palestinians including one child", "UK gives Iran ultimatum to freeze nukes program", "Pakistan raids terrorists camp", and "Israel renews threats to remove Arafat". A red arrow points from the text "How many times?" to the "2 captures" link in the browser's address bar.

How many times?

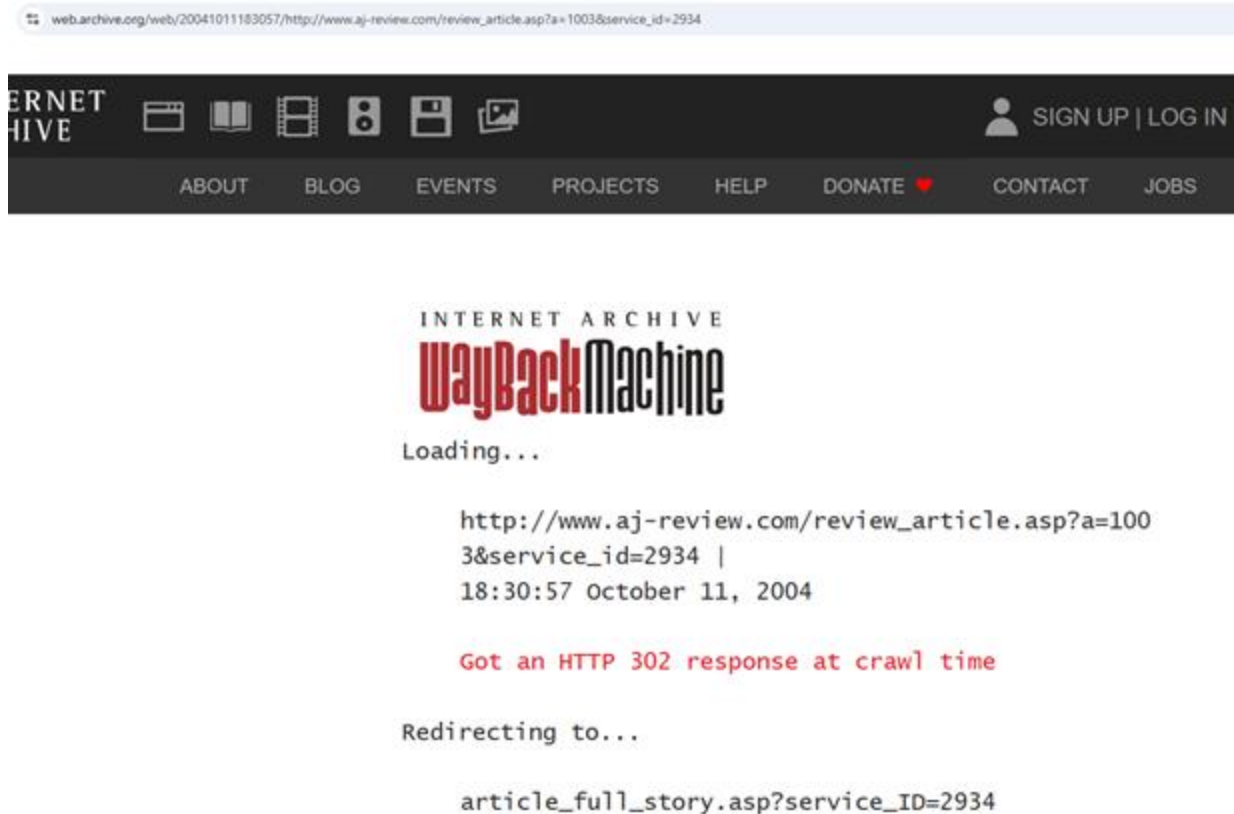
Not Archived

https://web.archive.org/web/20040829043439/http://www.aj-review.com/middle_east.asp?a=1003&service_id=3279



Crawl Time Redirects

https://web.archive.org/web/20041011183057/http://www.aj-review.com/review_article.asp?a=1003&service_id=2934



Languages Ranking by Speakers

<https://www.cia.gov/the-world-factbook/countries/world/#people-and-society>

The screenshot shows the CIA World Factbook website. The top navigation bar includes 'Go to CIA.gov', a search bar, and 'THE WORLD FACTBOOK' logo. The main navigation menu has 'Countries', 'Maps', and 'References'. The left sidebar lists various categories, with 'People and Society' highlighted. The main content area is titled 'People and Society' and includes sections for 'Population' and 'Languages'. The 'Languages' section lists the most-spoken and most-spoken first languages with their respective percentages. Two red arrows point from text annotations to the 'Languages' section.

Go to CIA.gov Search

THE WORLD FACTBOOK Countries Maps References

Contents

- Introduction
- Geography
- People and Society**
- Environment
- Government
- Economy
- Energy
- Communications
- Transportation
- Military and Security
- Transnational Issues

People and Society

Population

total: 8,057,236,243 (2024 est.)

male: 4,046,854,454

female: 4,010,381,789

Languages

most-spoken language: English 18.8%, Mandarin Chinese 13.8%, Hindi 7.5%, Spanish 6.9%, French 3.4%, Arabic 3.4%, Bengali 3.4%, Russian 3.2%, Portuguese 3.2%, Urdu 2.9% (2022 est.)

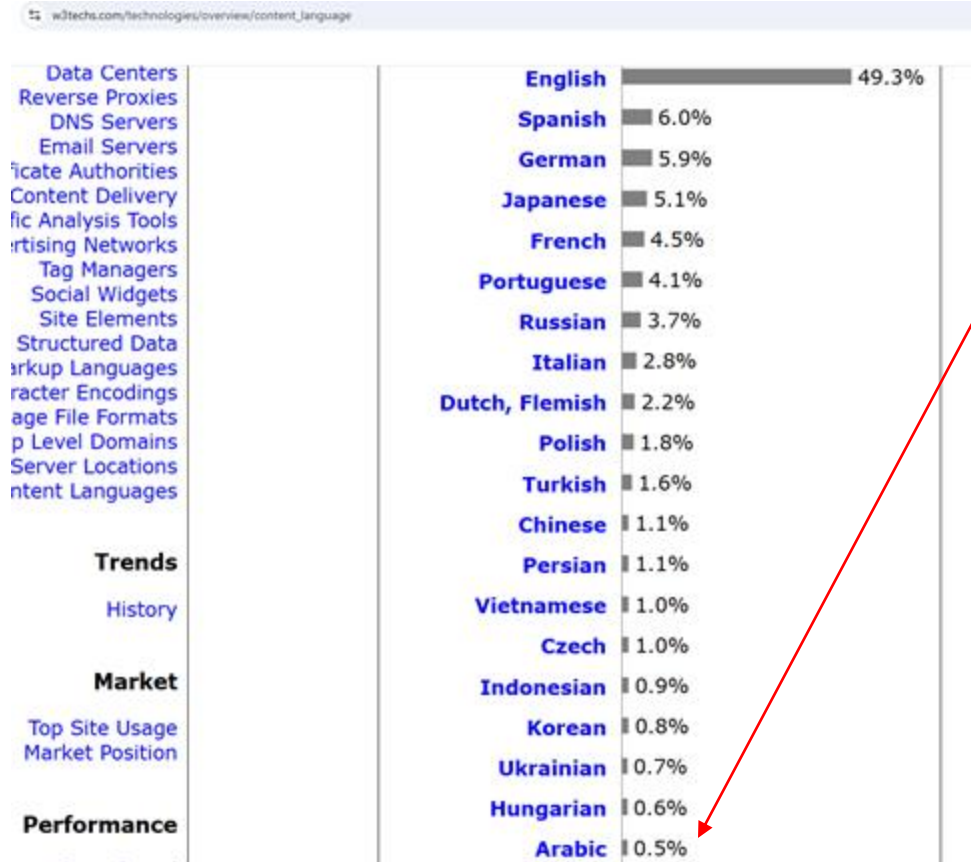
most-spoken first language: Mandarin Chinese 12.3%, Spanish 6%, English 5.1%, Arabic 5.1%, Hindi 3.5%, Bengali 3.3%, Portuguese 3%, Russian 2.1%, Japanese 1.7%,

Arabic is in the sixth place

Arabic is in fifth place

Arabic Digital Footprint is Small

https://w3techs.com/technologies/overview/content_language



Only 0.5% of web pages are in Arabic

Archiving the Rare Arabic Web Pages

- We collected 1.5M live URIs from four news outlets (JANA1.5M)
- Aljazeera (ar, en), Alarabiyah (ar), Arab News (en)
- We compare the archival rate of Arabic vs. English news pages between 1999 and 2022.
- 1.5M web pages will be sent to the Internet Archive (IA) to be archived

Collection Methods

- X (Twitter)
- RSS feeds
- GDELT
- SiteMaps

Not all News is Posted on X

Aljazeera Arabic has
113.2k posts on X



aljazeera_arabic_sitemap_links_set.txt - Notepad

File Edit Format View Help

https://www.aljazeera.net/videos/2012/4/23/الإسلامية-في-البوسنة
https://www.aljazeera.net/2004/11/01/الشرعية-الإسلامية-في-نيجيريا
https://www.aljazeera.net/sport/football/المنتخب-معسكر-حضور
https://www.aljazeera.net/news/arabic/2002/5/22/جاء-الحوار-مع
https://www.aljazeera.net/news/presstour/2014/7/1/ومحاسبة
https://www.aljazeera.net/ebusiness/2004/11/28/لإدراجها-داخل
https://www.aljazeera.net/news/arabic/2005/3/10/راقية-واثنين-من
https://www.aljazeera.net/news/arabic/2009/3/26/ستانفان-بالقاهرة
https://www.aljazeera.net/news/international/2008/3/6/اركية
https://www.aljazeera.net/news/politics/2020/4/17/الخرية
https://www.aljazeera.net/news/miscellaneous/2012/5/1/ة-من
https://www.aljazeera.net/ebusiness/2009/10/10/صفقة-خصخصة
https://www.aljazeera.net/news/cultureandart/2016/6/20/ري
https://www.aljazeera.net/news/reportsandinterviews/2012/
https://www.aljazeera.net/news/politics/2019/3/18/سطنبول-يوم
https://www.aljazeera.net/news/scienceandtechnology/2014/
https://www.aljazeera.net/news/international/2014/1/2/بانغي

816,892 live news stories from
Aljazeera Arabic SiteMap

RSS (Outdated, Non-existent, Too Many Feeds)

```
This XML file does not appear to have any style information associated with it. The document root is:

<?xml version="1.0" encoding="UTF-8" ?>
<feed xmlns="http://www.w3.org/2005/Atom" xmlns:dc="http://purl.org/dc/terms/" xmlns:media="http://search.yahoo.com/m"
  <id>http://www.bbc.co.uk/arabic/middleeast/index.xml</id>
  <title xml:lang="ar">BBCArabic.com | الشرق الأوسط</title>
  <updated>2014-09-25T10:31:19+00:00</updated>
  <link rel="self" href="http://www.bbc.co.uk/arabic/middleeast/index.xml"/>
  <author>
    <name>BBC Arabic</name>
    <email>arabic@bbc.co.uk</email>
    <uri>http://www.bbc.co.uk/arabic/</uri>
  </author>
  <generator>http://www.bbc.co.uk/arabic/</generator>
  <category xml:lang="ar" term="arabic_1024" label="arabic_1024"/>
  <category xml:lang="ar" term="middleeast" label="الشرق الأوسط"/>
  <logo>http://www.bbc.co.uk/arabic/images/gel/rss_logo.gif</logo>
  <rights xml:lang="ar">2014 حقوق النشر محفوظة لهيئة الإذاعة البريطانية</rights>
  <entry>
    <id>tag:www.bbc.co.uk,2014-09-25:34934377</id>
    <dc:identifier>34934377</dc:identifier>
    <updated>2014-09-25T07:51:40+00:00</updated>
    <published>2014-09-25T06:13:44+00:00</published>
    <category xml:lang="ar" term="arabic_1024" label="arabic_1024"/>
    <category xml:lang="ar" term="middleeast" label="الشرق الأوسط"/>
    <rights>not restricted</rights>
    <title xml:lang="ar">غازات لتتحالف على مصالي النفط التابعة لتنظيم "الدولة" في سوريا</title>
    <summary xml:lang="ar">تقودها الولايات المتحدة المالية واستهدفت مصافي النفط التي يسيطر عليها التنظيم في سوريا</summary>
  </entry>
```

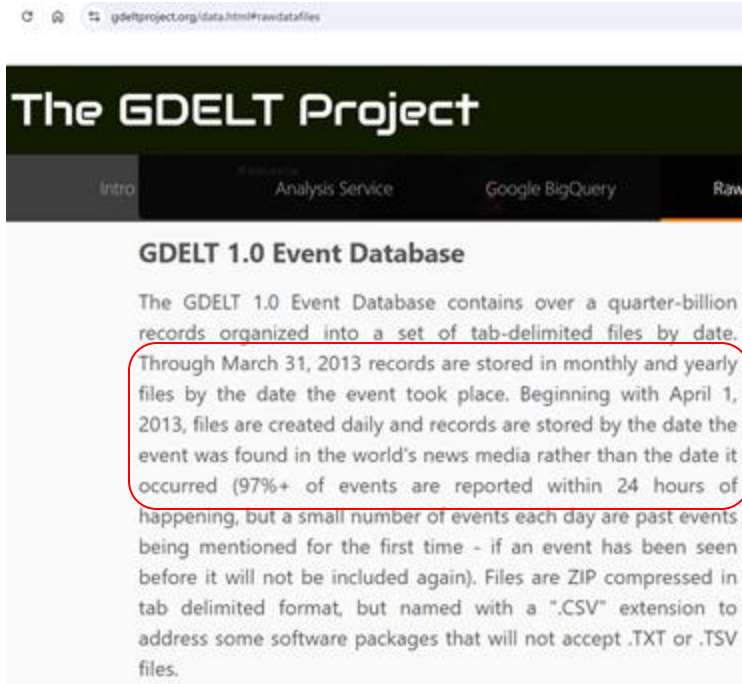
Middle East RSS feed from BBC Arabic has not been updated since 2014

The screenshot shows the Alarabiya.net website. In the footer, there is a link for the RSS feed: <http://www.alarabiya.net/tools/mrss>. The website is in Arabic and features various news categories and a search bar.

العنوان	رابط RSS
أخبار الصفحة الرئيسية	رابط RSS
أخبار العرب والعالم	رابط RSS
أخبار السعودية	رابط RSS
أخبار اقتصاد (أسواق)	رابط RSS
أخبار الرياضة	رابط RSS
أخبار مؤونة	رابط RSS
المقالات	رابط RSS
أخبار تكنولوجيا	رابط RSS
أخبار الأندلس	رابط RSS

Nine RSS feeds to be monitored to collect news URLs from Alarabiya
<https://www.alarabiya.net/tools/mrss>

GDELT did not offer URLs for news stories before 2013-04-01



The GDELT Project

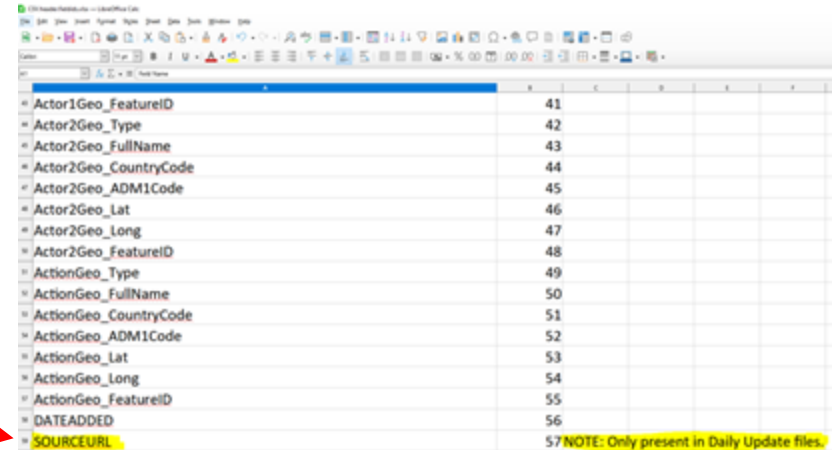
Intro Analysis Service Google BigQuery Raw

GDELT 1.0 Event Database

The GDELT 1.0 Event Database contains over a quarter-billion records organized into a set of tab-delimited files by date. Through March 31, 2013 records are stored in monthly and yearly files by the date the event took place. Beginning with April 1, 2013, files are created daily and records are stored by the date the event was found in the world's news media rather than the date it occurred (97%+ of events are reported within 24 hours of happening, but a small number of events each day are past events being mentioned for the first time - if an event has been seen before it will not be included again). Files are ZIP compressed in tab delimited format, but named with a ".CSV" extension to address some software packages that will not accept .TXT or .TSV files.

Daily updated files started on April 1st, 2013

URLs are only offered in daily updated files starting April 1st, 2013



Actor1Geo_FeatureID	41
Actor2Geo_Type	42
Actor2Geo_FullName	43
Actor2Geo_CountryCode	44
Actor2Geo_ADM1Code	45
Actor2Geo_Lat	46
Actor2Geo_Long	47
Actor2Geo_FeatureID	48
ActionGeo_Type	49
ActionGeo_FullName	50
ActionGeo_CountryCode	51
ActionGeo_ADM1Code	52
ActionGeo_Lat	53
ActionGeo_Long	54
ActionGeo_FeatureID	55
DATEADDED	56
SOURCEURL	57
NOTE: Only present in Daily Update files.	

GDELT 1.0 GKG Begins 2013-04-01

<https://www.gdeltproject.org/data.html#rawdatafiles>



GDELT 1.0 Global Knowledge Graph (GKG)

The GDELT 1.0 Global Knowledge Graph begins April 1, 2013 and consists of two parallel data streams, one encoding the entire

GDELT 2.0 has Arabic News URLs (2015-present)

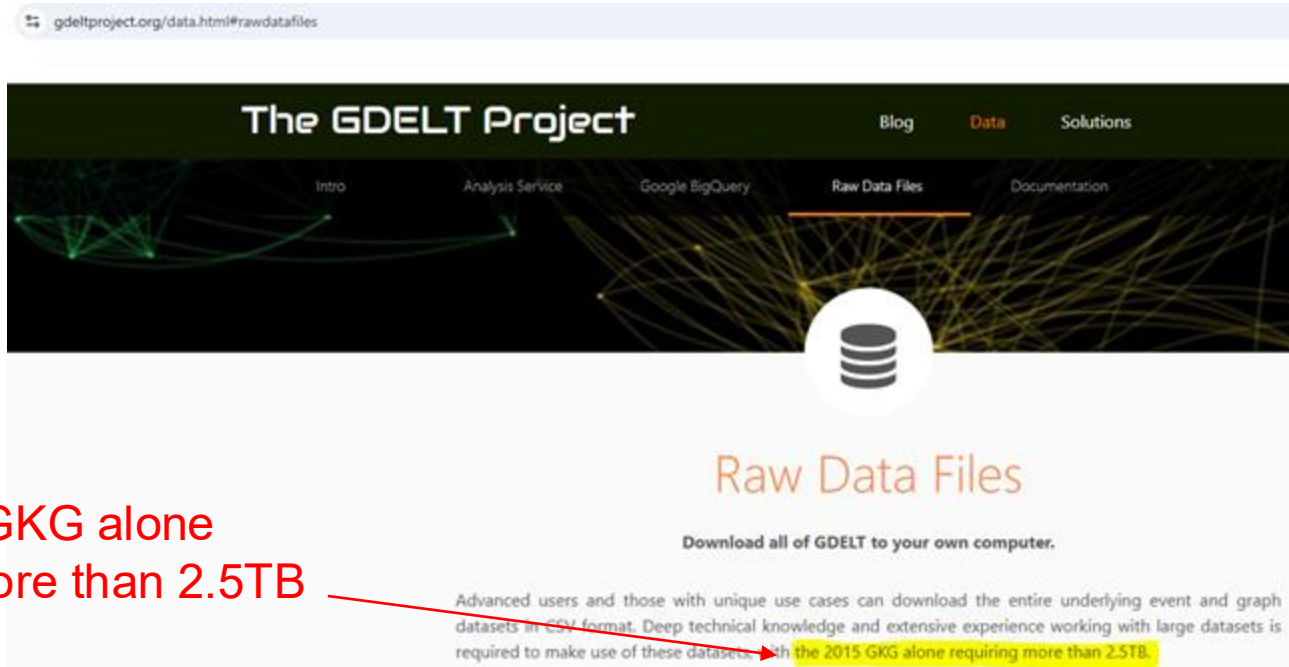
<http://data.gdeltproject.org/gdeltv2/masterfilelist.txt>

← → ↻ 🏠 ⚠ Not secure data.gdeltproject.org/gdeltv2/masterfilelist.txt

```
150383 297a16b493de7cf6ca809a7cc31d0b93 http://data.gdeltproject.org/gdeltv2/20150218230000.export.CSV.zip
318084 bb27f78ba45f69a17ea6ed7755e9f8ff http://data.gdeltproject.org/gdeltv2/20150218230000.mentions.CSV.zip
10768507 ea8dde0beb0ba98810a92db068c0ce99 http://data.gdeltproject.org/gdeltv2/20150218230000.gkg.csv.zip
149211 2a91041d7e72b0fc6a629e2ff867b240 http://data.gdeltproject.org/gdeltv2/20150218231500.export.CSV.zip
339037 dec3f427076b716a8112b9086c342523 http://data.gdeltproject.org/gdeltv2/20150218231500.mentions.CSV.zip
10269336 2f1a504a3c4558694ade0442e9a5ae6f http://data.gdeltproject.org/gdeltv2/20150218231500.gkg.csv.zip
149723 12268e821823aae2da90882621feda18 http://data.gdeltproject.org/gdeltv2/20150218233000.export.CSV.zip
357229 744acad14559f2781a8db67715d63872 http://data.gdeltproject.org/gdeltv2/20150218233000.mentions.CSV.zip
11279827 66b03e2efd7d51dabf916b1666910053 http://data.gdeltproject.org/gdeltv2/20150218233000.gkg.csv.zip
158842 a5298ce3c6df1a8a759c61b5c0b6f8bb http://data.gdeltproject.org/gdeltv2/20150218234500.export.CSV.zip
374528 dd322c888f28311aca2c735468405551 http://data.gdeltproject.org/gdeltv2/20150218234500.mentions.CSV.zip
11212939 cd20f295649b214dd16666ca451b9994 http://data.gdeltproject.org/gdeltv2/20150218234500.gkg.csv.zip
362610 c4268d558bb22c02b3c132c17818c68b http://data.gdeltproject.org/gdeltv2/20150219000000.export.CSV.zip
```

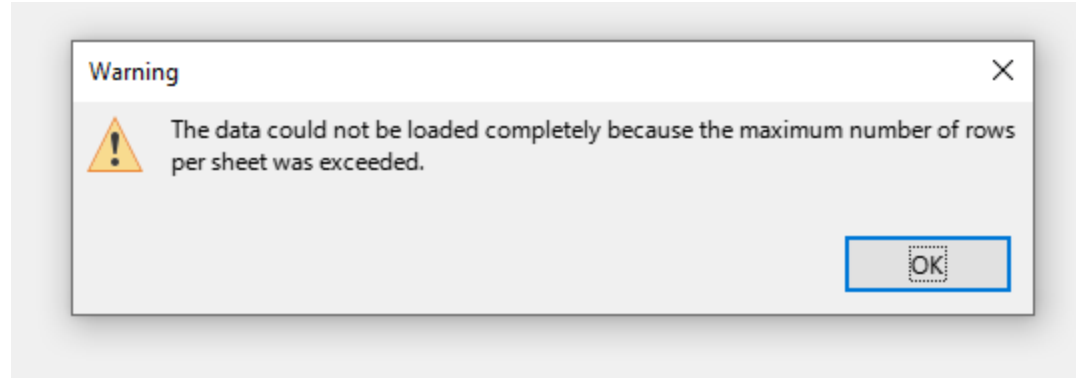
GDELT is Too Big to Manage Easily with Limited Resources

<https://www.gdelproject.org/data.html#rawdatafiles>



The 2015 GKG alone
requires more than 2.5TB

GDELT 1.0 Events Dataset (Year: 2000)



More recent years are much larger

GDELT is Large but Far from Complete

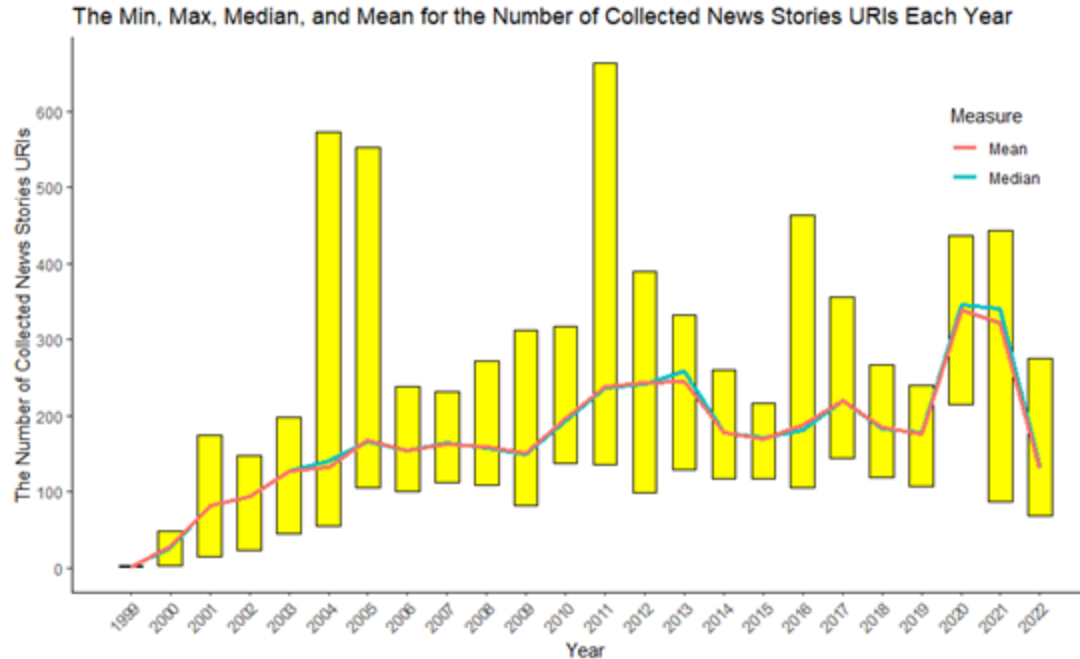
- Per news outlet, JANA1.5M is more complete than GDELT.
- This is true for Arabic and English news URLs.

News Outlet	GDELT	Our dataset	Shared	GDELT only	Our dataset only
Aljazeera English	19	44	18	1	25
Arab News	2	84	0	2	82
Total URLs collected	21	128	18	3	107

The number of URLs for one day's worth of data (2013-04-01) studied for GDELT's GKG and JANA1.5 for aljazeera.com (English) and arabnews.com (English).

Sample from Median day per year

- Sample: 4116 URIs (2684 Arabic, 1432 English).
- JANA1.5M has symmetrical distribution.



Query All Public Web Archives (21 Archives)

- Used [MemGator](#)
- Found archival status for each URI.
- Found archived URIs' exclusivity by web archive.
- Compared Arabic news vs. English archival rate
- Archival rate periods:
1999-2022, 1999-2013, and 2013-2022

Web Archives Contribution to Archiving News Pages in our Dataset

Archive	Archived Arabic	Archived English	Unique Arabic	Unique English
The Internet Archive (web.archive.org)	1435	828	1429	732
Archive-It: powered by IA (wayback.archive-it.org)	0	156	0	0
Stanford Web Archive (swap.stanford.edu)	0	46	0	5
Arquivo.pt	3	34	0	0
Archive Today (archive.today)	4	27	0	0
Perma.cc	0	2	0	0
BAnQ (waext.banq.qc.ca)	0	0	0	0
Icelandic Web Archive (wayback.vefsafn.is)	0	0	0	0
Australian Web Archive (web.archive.org.au)	0	0	0	0
UK National Archives Web Archive (webarchive.nationalarchives.gov)	0	0	0	0
National Records of Scotland (webarchive.nrscotland.gov.uk)	0	0	0	0
UK Web Archive (webarchive.org.uk)	0	0	0	0
UK Parliament Web Archive (webarchive.parliament.uk)	0	0	0	0
National Library of Israel (wayback.nli.org.il)	0	0	0	0
Library of Congress (loc.gov/lcwa)	0	0	0	0
Library & Archives Canada (collectionscanada.gc.ca)	0	0	0	0
Web Archive of Catalonia (padi.cat)	0	0	0	0
Croatian Web Archive (haw.nsk.hr)	0	0	0	0
Archive of the Czech Web (webarchiv.cz)	0	0	0	0
National Taiwan University (webarchive.lib.ntu.edu.tw)	0	0	0	0
National Diet Library, Japan (warp.da.ndl.go.jp)	0	0	0	0
PRONI Web Archive (webarchive.proni.gov.uk)	Moved to archive-it.org			
PastPages (pastpages.org)	Defunct			
Bibliotheca Alexandrina Web Archive (web.archive.bibalex.org)	Defunct			

Archival Rate is Increasing, but for Arabic News, it is Faster

Percentage of Archived Arabic News Stories' URIs by Year

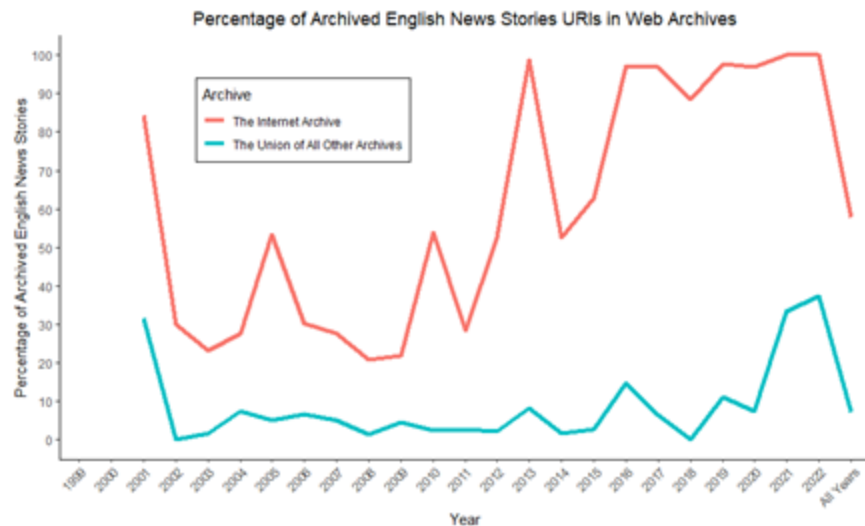
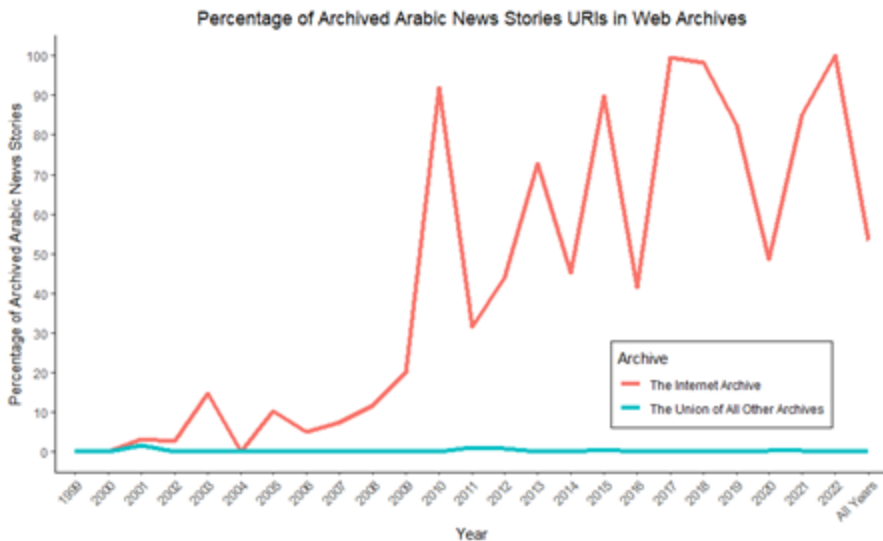


Percentage of Archived English News Stories' URIs by Year

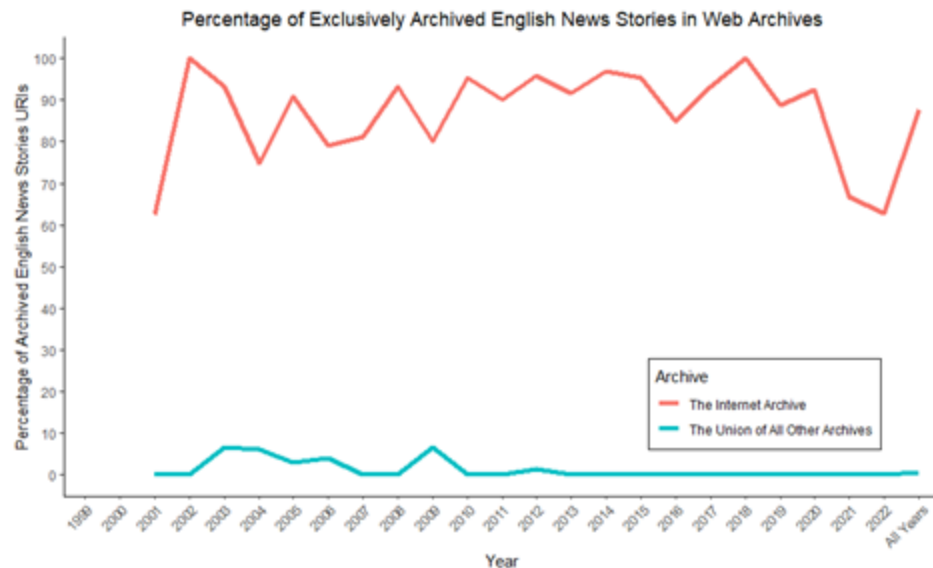
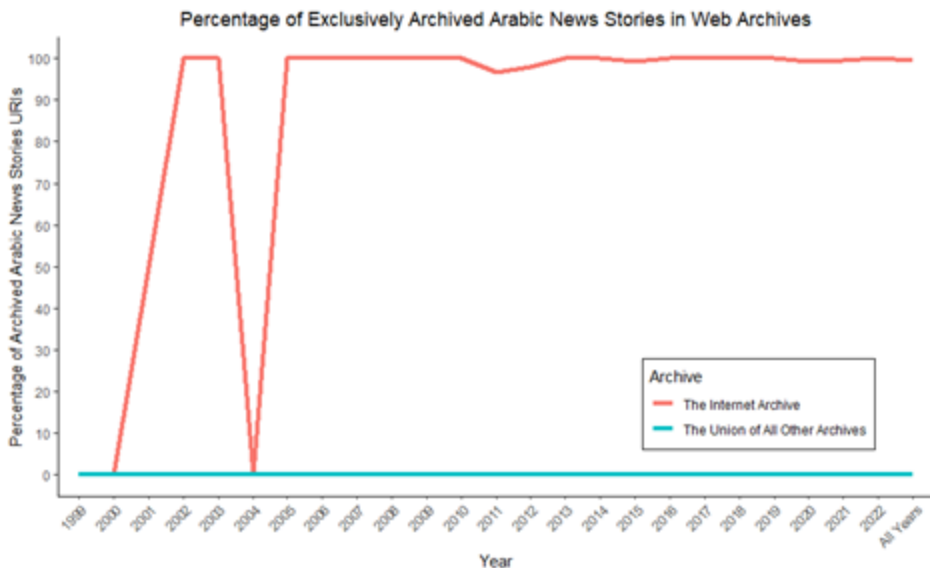


- From 1999-2022: Arabic news archival rate is 53%. English is 58%
- Arabic news archival rate increased from 24% to 53% from 2013 to 2022.
- English news archival rate increased from 47% to 58% from 2013 to 2022.
- 99.6% of Arabic news archived pages only exist in the IA.
- 87.7% of English news archived pages only exist in the IA.

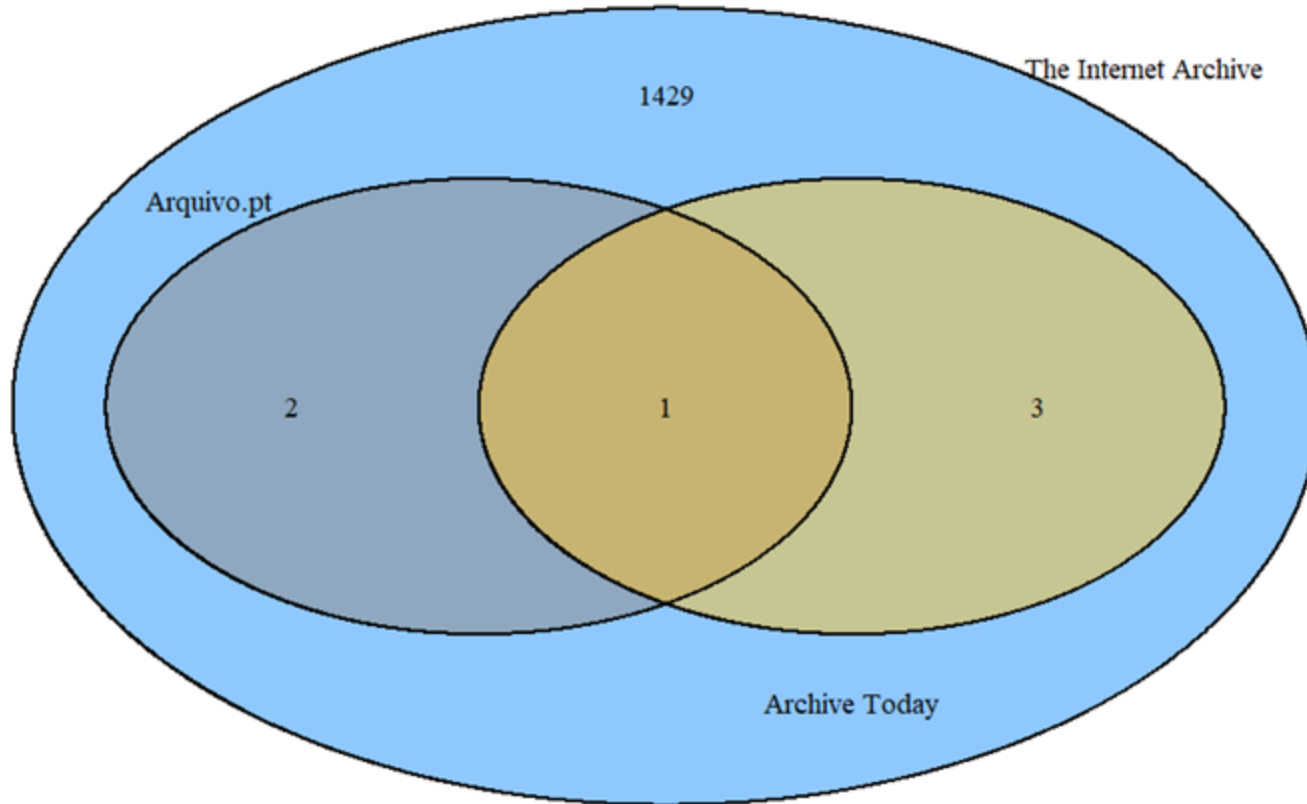
The Internet Archive Holds More URIs than the Sum of All Other Web Archives



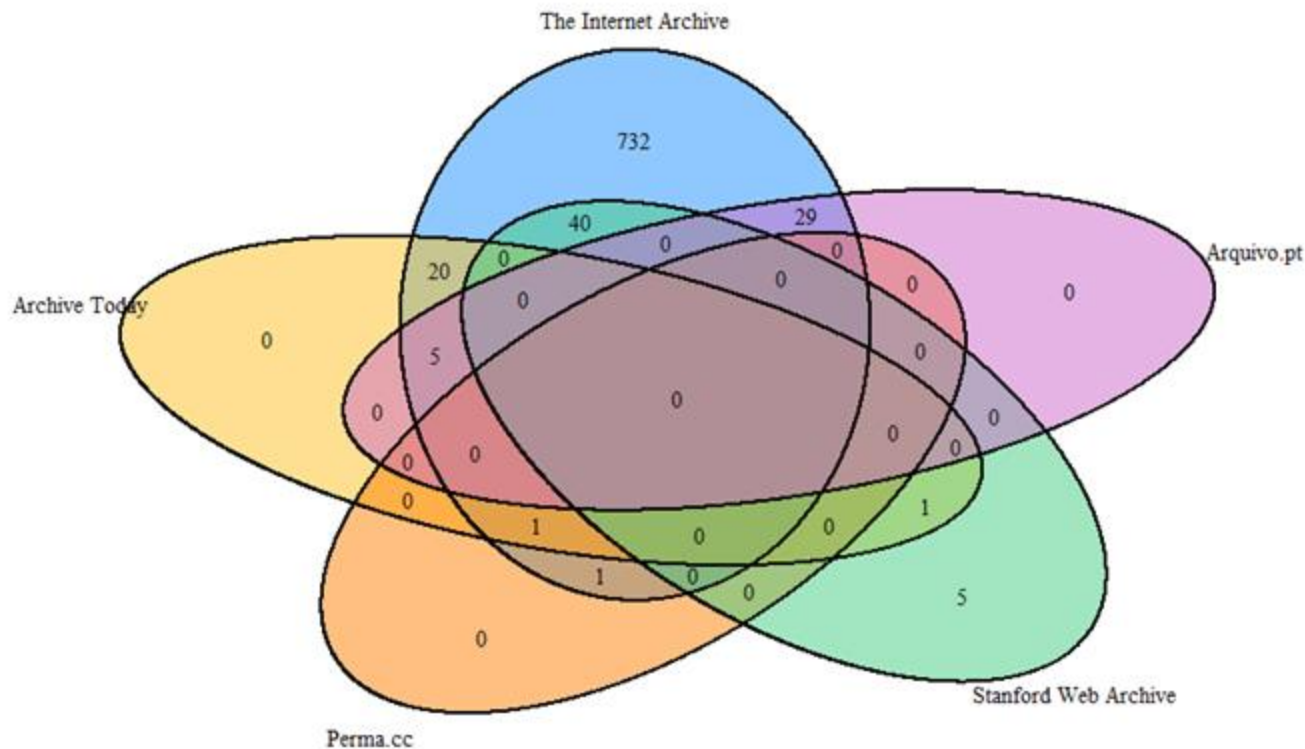
The Internet Archive Holds More Exclusive URIs than the Sum of All Other Web Archives



Almost No Redundancy in Web Archives Holdings of Arabic News Pages



Little Redundancy in Web Archives Holdings of English News Pages



Only the IA Increased its Holdings

- Arabic archiving growth mostly due to IA, not other archives.
- Extreme centralization: losing IA would erase nearly all archived digital history in Arabic.
- English also heavily dependent on the IA but less extreme.

Conclusions

- Contrary to expectations: Arabic news archival rate is growing faster than English.
- The gap between Arabic and English archival rates is shrinking.
- The IA is essential for both, but it is more critical for Arabic.
- Urgent need for decentralized Arabic-focused web archives.
- The dataset is available on [GitHub](#)