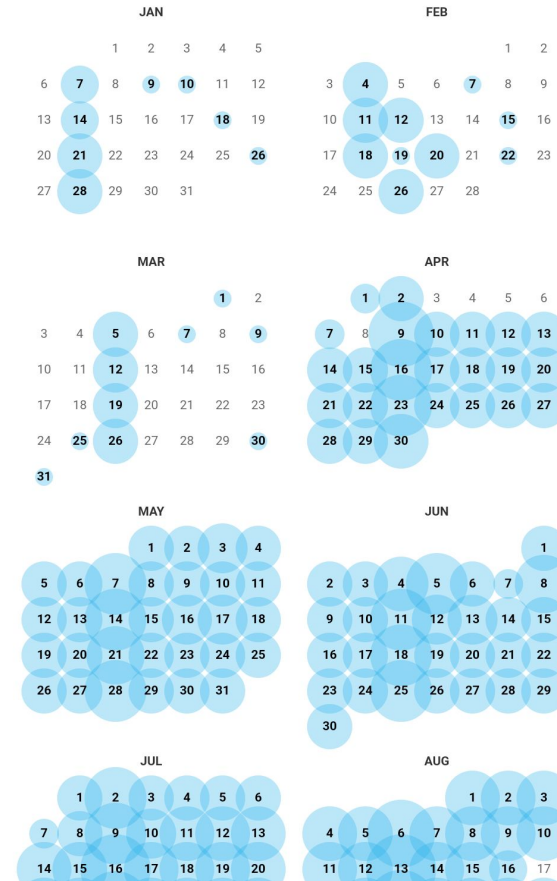


# 404: Archiving the Web

Ryder Kouba  
Digital Collections Archivist  
Rare Books and Special Collections Library  
The American University in Cairo

# What is web archiving?

- Collecting and preserving content from the web for future use
- Began in 1996 with the Internet Archive and various national libraries
- Seeds, crawls, captures, warc



# Why archive the web?

- Necessary to document society
- Temporary lifespan of content online
- Continuation of current collections (e.g. newspapers)
- More evidentiary value than printing



# Web content is vulnerable

You are viewing an archived web page, collected at the request of [American University in Cairo](#) using [Archive-It](#). This page was captured on 10:07:49 Jun 07, 2012, and is part of the [Egypt Revolution and Politics](#) collection. The information on this web page may be out of date. See [All versions](#) of this archived page. [Videos Metadata](#) [Enable QA](#) hide



# Accountability

Menu | 

The CHRISTIAN SCIENCE  
MONITOR®

Subscribe: \$11 / Month or [Login](#)

## Web evidence points to pro-Russia rebels in downing of MH17

There are strong, though not conclusive, indications that pro-Russian separatist rebels fired the missile that downed Malaysian Airlines Flight MH17 today.

### Latin American Government Documents Archive, LAGDA

**Collected by:** [The University of Texas at Austin, LILAS Benson Latin American Studies and Collections](#)

**Archived since:** Sep, 2005

**Description:** The Latin American Government Documents Archive (LAGDA) seeks to preserve and facilitate access to a wide range of ministerial and presidential documents from 18 Latin American and Caribbean countries. The Archive contains copies of the Web sites of approximately 300 government ministries and presidencies. Capture of sites began on multiple dates in 2005 and 2006, and will continue with regularly scheduled captures. Content in the Archive includes not only the full-text versions of official documents, but also original video and audio recordings of key regional leaders. Archive contents include thousands of annual and "state of the nation" reports; plans and programs; and speeches by presidents and government ministers. LAGDA is a joint project of the University of Texas Libraries, The Nettie Lee Benson Latin American Collection, and the Latin American Network Information Center at The University of Texas at Austin.

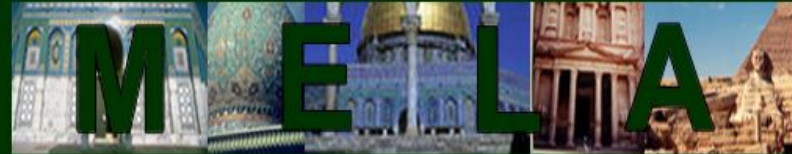
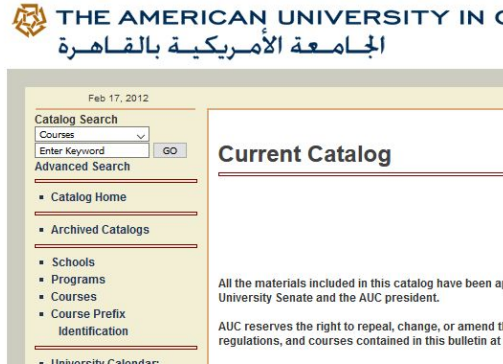
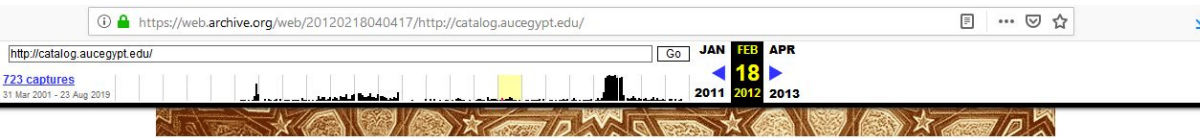
**Subject:** [Government - National](#), [government ministries](#)



THE AMERICAN UNIVERSITY IN CAIRO

الجامعة الأمريكية بالقاهرة

# Websites as records



## Middle East Librarians Association

It is the purpose of the Middle East Librarians' Association to facilitate communication among members through meetings and publications; to improve the quality of area librarianship through the development of standards for the profession and education of Middle East library specialists; to compile and disseminate information concerning Middle East libraries and collections and to represent the judgment of the members in matters affecting them; to encourage cooperation among members and Middle East libraries, especially in the acquisition of materials and the development of bibliographic control; to cooperate with other library and area organizations in projects of mutual concern and benefit; to promote research in and development of indexing and automated techniques as applied to Middle East materials.





President Trump Retweeted



**Donald J. Trump** @realDonaldTrump · Jan 26

The U.S. has a 60 billion dollar trade deficit with Mexico. It has been a one-sided deal from the beginning of NAFTA with massive numbers...



12K



25K



96K



President Trump Retweeted



**Donald J. Trump** @realDonaldTrump · Jan 26

of jobs and companies lost. If Mexico is unwilling to pay for the badly needed wall, then it would be better to cancel the upcoming meeting.



40K



27K



105K



President Trump Retweeted



**Donald J. Trump** @realDonaldTrump · Jan 26

Ungrateful TRAITOR Chelsea Manning, who should never have been released from prison, is now calling President Obama a weak leader. Terrible!



25K



27K



117K



THE AMERICAN UNIVERSITY IN CAIRO

الجامعة الأمريكية بالقاهرة

# Culture

COLLECTION

## Web Cultures Web Archive

About this Collection

Collection Items

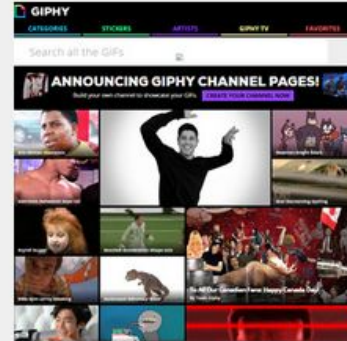
### Featured Content



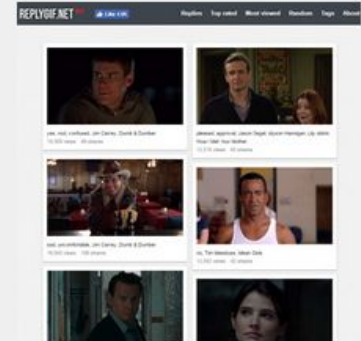
Save The Pacific Northwest Tree Octopus



YTMND: You're the man now dog!



Search Animated GIFs on the Web - Giphy



Replygif.net - Reaction GIFs for every situation



# Common tools

- Wayback Machine
- Archive-It
- Webrecorder.io
- Mirrorweb
- Heritrix
- WARCcreate
- Archive.is



Subscription

Service

Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. [Visit Archive-It to build and browse the collections.](#)



Save Page Now

https://

SAVE PAGE

Capture a web page as it appears now for use as a trusted citation in the future.

A banner image showing a wooden desk with a laptop, a smartphone, and a document titled 'Statistcs'. Overlaid on the image is the text 'Archiving for digital marketing and compliance professionals' in large white font.

## Archiving for digital marketing and compliance professionals

We help regulated firms manage and evidence changes in their digital content.

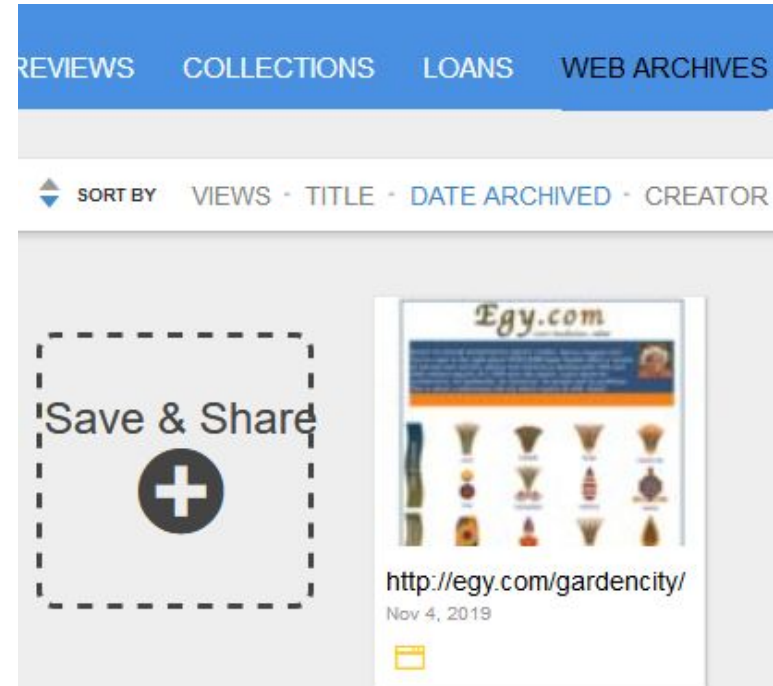


THE AMERICAN UNIVERSITY IN CAIRO

الجامعة الأمريكية بالقاهرة

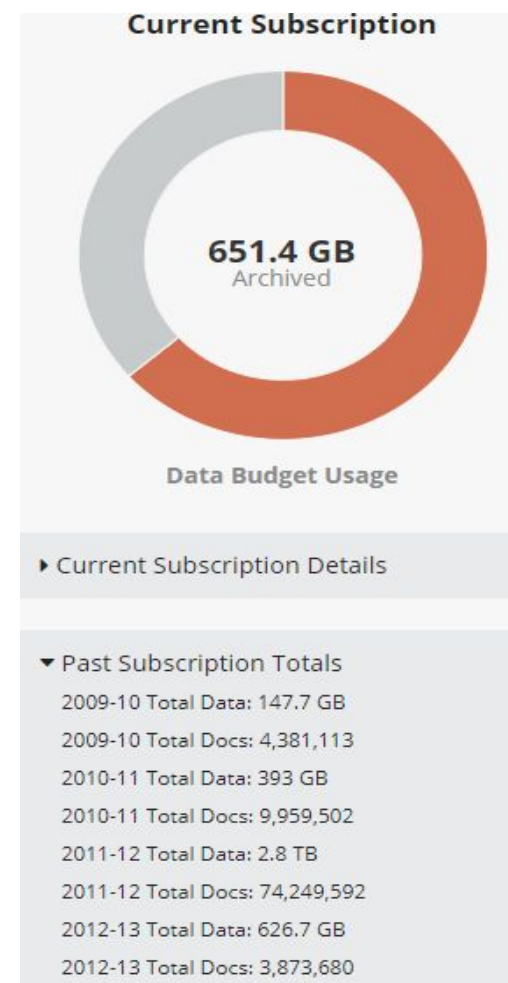
# Wayback Machine

- Easy to preserve and provide access to websites
- Difficult to organize seeds and metadata
- Currently free
- No control over archived data



# Archive-It

- Subscription service AUC currently uses
- Ability to organize collections, create metadata, etc.
- More technical options
- Limits on amount of data that can be saved
- Requires more time and resources than other options



# Archive-It

- Automated crawls
- Flexible scheduling

## Scheduled Crawls

Frequency	Active Seeds	Next Crawl	Last Crawl	Time Limit	Data Limit	Doc. Limit	Edit Schedule
Monthly	<a href="#">36</a>	Nov 24, 2019	<a href="#">Oct 24, 2019</a>	3 Days	3 GB	5,000	<a href="#">Edit &gt;</a>
Bimonthly	<a href="#">3</a>	Nov 24, 2019	<a href="#">Sep 24, 2019</a>	3 Days	3 GB	5,000	<a href="#">Edit &gt;</a>
Quarterly	<a href="#">7</a>	Feb 7, 2020	<a href="#">Nov 7, 2019</a>	3 Days	2 GB	5,000	<a href="#">Edit &gt;</a>
Annual	<a href="#">134</a>	Mar 24, 2020	<a href="#">Mar 24, 2019</a>	5 Days	7 GB	10,000	<a href="#">Edit &gt;</a>
Semiannual	<a href="#">26</a>	Dec 25, 2019	<a href="#">Jun 26, 2019</a>	5 Days	7 GB	5,000	<a href="#">Edit &gt;</a>
Daily	<a href="#">2</a>	--	<a href="#">Apr 4, 2016</a>	1 Day	2 GB	5,000	<a href="#">Edit &gt;</a>
One-Time	<a href="#">557</a>	--		23 Hours	No Limit	5,000	<a href="#">Edit &gt;</a>
Weekly	<a href="#">11</a>	Nov 14, 2019	<a href="#">Nov 7, 2019</a>	3 Days	3 GB	5,000	<a href="#">Edit &gt;</a>



# Other options

- Webrecorder.io - the least mysterious web crawler in history
- Heritrix: Open-source Internet Archive crawler

## New Capture

<https://www.mela.us/>

Add to collection

Default Collection

Select browser

+ Create new collection

Filter collections...

▼ Session settings

Session Notes

Default Collection

Add notes about this session. Visible only to you.

```
# URLs HERE
http://example.example/example

</prop>
</props>
</property>
</bean>

<!-- CRAWL METADATA: including identification of crawler/operator -->
<bean id="metadata" class="org.archive.modules.CrawlMetadata" autowire="byName">
  <property name="operatorContactUrl" value="[see override above]"/>
  <property name="jobName" value="[see override above]"/>
  <property name="description" value="[see override above]"/>
  <!-- <property name="operator" value="" --> -->
  <!-- <property name="operatorFrom" value="" --> -->
  <!-- <property name="organization" value="" --> -->
  <!-- <property name="audience" value="" --> -->
  <!-- <property name="userAgentTemplate"
        value="Mozilla/5.0 (compatible; heritrix/@VERSION@ +@OPERATOR_CONTACT_URL@)"/> -->
</bean>
```



# The harsh reality



## Not in Archive

The page you requested has not been archived in Archive-It.

This could be for a number of reasons.





# Building collections

- How can web archiving compliment existing collections?
- What is the collection scope?
- Who will be responsible for what tasks?

Collection Name	Data (this period) ▼	Docs (this period)	Active Seeds	Last Crawl
<a href="#">American University in Cairo</a>	224.2 GB	3,246,037	345	Oct 29, 2019
<a href="#">Egypt Non-Governmental Organi</a>	79.4 GB	683,883	331	Oct 31, 2019
<a href="#">Egypt Travel and Tourism</a>	76.3 GB	1,165,751	426	Nov 1, 2019
<a href="#">Egypt Gender &amp; Sexuality Issue</a>	59.3 GB	711,193	183	Oct 31, 2019
<a href="#">Egypt Newspapers, Magazines,</a>	55.9 GB	417,482	176	Oct 29, 2019
<a href="#">Egypt Ethnic, Foreign and Expat</a>	44.1 GB	540,723	236	Oct 25, 2019
<a href="#">Egypt and Middle East Architect</a>	40.6 GB	614,240	392	Oct 22, 2019
<a href="#">Egypt Revolution and Politics</a>	38.3 GB	486,363	793	Oct 31, 2019
<a href="#">Egypt Arts, Leisure, Entertainme</a>	30.7 GB	539,744	463	Oct 31, 2019
<a href="#">Egypt Religious Communities</a>	11.8 GB	104,930	547	Jun 25, 2019



# Other workflow challenges

- Constant changes and updates
- Managing/organizing seeds
- Quality Assurance

<http://www.nytimes.com/interactive/2013/08/21>.

Match: "tp://www.nytimes.com/inter"

<http://www.nytimes.com/packages/flash/newsgr>

Match: "tp://www.nytimes.com/packa"

<http://www.nytimes.com/interactive/2011/02/12>.

Match: "tp://www.nytimes.com/inter"

<http://topics.nytimes.com/top/news/internation>

Match: "://topics.nytimes.com/top/n"

<http://thelede.blogs.nytimes.com/2013/11/04/eg>

Match: "ede.blogs.nytimes.com/2013/"

<http://thelede.blogs.nytimes.com/2013/07/06/vic>

Match: "ede.blogs.nytimes.com/2013/"



<http://thelede.blogs.nytimes.com/2013/07/05/so>

Match: "ede.blogs.nytimes.com/2013/"



# Description and access

- Necessary?
- Level of description?
- OCLC recently released “Descriptive Metadata for Web Archiving”

Collection Guide	
Collection Title:	Web Archive (University of California, Irvine)
Collection Number:	AS.187
Get Items:	 Online items available  Contact UC Irvine::University Archives
Collection Overview	
Description	The University of California, Irvine Web Archive contains hundreds of websites documenting UCI. Most websites reside within the "uci.edu" domain, but it does contain external sites, videos, and social media pages. This continuously growing collection documents UCI's administrative and academic departments, student organizations, faculty and staff, events, milestones, and publications of UCI.
Extent	280 Gigabytes (200+ websites)



# Searching a phonebook

- Wayback search is not full-text
- Currently very helpful to know the URL of a (former) site

Explore more than 387 billion [web pages](#) saved over time

"middle east librarians association"

<http://mela.us/>

*middle east librarians association*

 375

 35

 0

 0

2,049 capture(s) from 2005 to 2017 | [Site stats](#)

<http://melcominternational.org/>

*melcom international*

 47

 22

 0

 0

437 capture(s) from 2011 to 2017 | [Site stats](#)



# Access through Archive-It

Group

Sort By: Count | (A-Z)

AUC News and Media (5)

Alumni Organizations (1)

Faculty Organizations (3)

RBSCl Journalists' Visit, 2016 (15)

Student Organizations (2)

More ▾

Subject

Sort By: Count | (A-Z)

College student newspapers and periodicals (2)

American University in Cairo (1)

American University in Cairo--Records and correspondence (1)

Creator

Sort By: Count | (A-Z)

American University in Cairo (2)

AUC Insider (1)

AUC Office of Student Development (1)

AUC Student Senate (1)

AUC Times Magazine (1)

More ▾

Enter search terms here

Sites

Search Page Text

Page 1 of 4 (355 Total)

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

URL: <http://24.com.eg/Mnwat/951517...>

Captured once on Mar 7, 2016

Group: RBSCl Journalists' Visit, 2016

URL: <http://abletoaspire.weebly.com/>

Description: Part of Kim Fox's JRMCMultimedia Journal

Captured 3 times between Nov 30, 2016 and Dec 2, 2016

Videos: 1 Videos Captured

Title: Faculty Bulletin

URL: <http://academic.aucegypt.edu/bulletin>

Captured 98 times between Mar 18, 2010 and Oct 1, 2016

Videos: 1 Videos Captured

THE AMERICAN UNIVERSITY IN CAIRO

الجامعة الأمريكية بالقاهرة

# Ethical issues

- Require permission from creators?
- Privacy by obscurity?
- Expectation of privacy?

INTERNET ARCHIVE  
**WayBackMachine**

Explore more than 387 billion web pages saved over time

"ryder kouba"

---

<http://getqualitybacklink.com/>  
*mozell klaers*  
11,223 32 0 0  
11,686 capture(s) from 2010 to 2016 | [Site stats](#)

---

<http://xn--mi-3sa.opole.pl/>  
*dwight weiden*  
1 0 0 0  
4 capture(s) from 2012 to 2012 | [Site stats](#)

---

<http://mir2h.com/>  
*pierre knatt*  
1,167 228 0 0  
2,441 capture(s) from 2006 to 2016 | [Site stats](#)



## We archive websites:

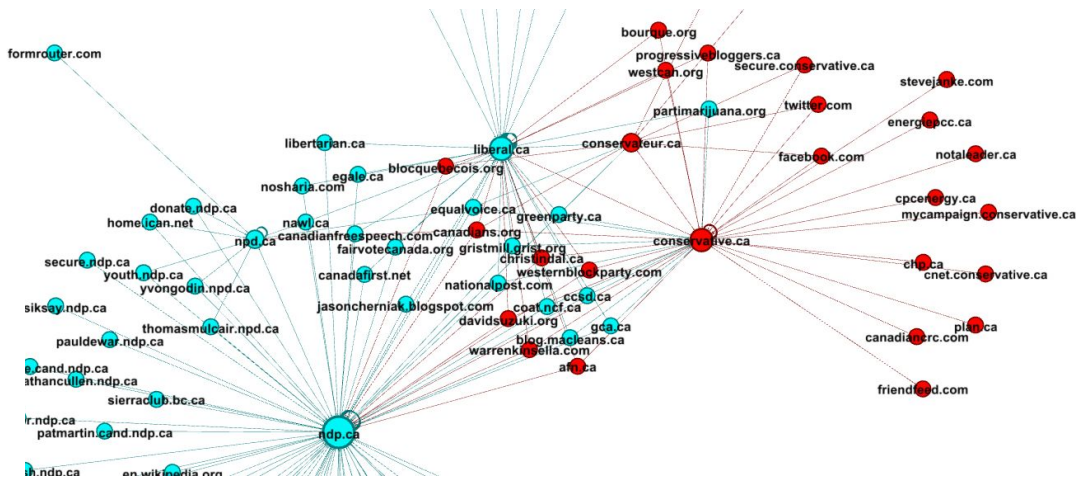
1. That are on a .uk or other UK geographic top-level domain such as .scot or .cymru
2. That are published in the UK.

## We do *not* archive:

1. Online Sound or Video platforms, in which audio-visual material is the predominant content
2. Private Intranets and Emails
3. Personal data in social networking sites or websites only available to restricted groups.

# Digital Humanities

# Archives Unleashed



Comparing WhiteHouse.gov (using Trump classifier)

