



**HAL**  
open science

## Archiving Web Content

Jean-Christophe Peyssard

► **To cite this version:**

Jean-Christophe Peyssard. Archiving Web Content. École thématique. Archiving Web Content, American University of Beirut, Beirut, Lebanon, Lebanon. 2019. cel-02130558

**HAL Id: cel-02130558**

**<https://shs.hal.science/cel-02130558>**

Submitted on 15 May 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# ARCHIVING WEB CONTENT

> DHIJ. 2019

---

Jean-Christophe Peyssard  
Head of Digital Humanities  
Institut français du Proche-Orient (Ifpo)  
[jc.peyssard@ifporient.org](mailto:jc.peyssard@ifporient.org)

Research Engineer from CNRS at the French Institute of the Near East, Head of Digital Humanities

- Ifpo : <http://www.ifporient.org/jean-christophe-peyssard/>
- LinkedIn : <http://fr.linkedin.com/pub/jean-christophe-peyssard/22/705/782/>
- ORCID : [0000-0002-8503-2217](https://orcid.org/0000-0002-8503-2217)
- Google Scholar : <https://scholar.google.com/citations?user=32cZHPsAAAAJ>
- HAL : <https://cv.archives-ouvertes.fr/jcpeyssard>



Institut français du Proche-Orient (Ifpo) MEAE / CNRS – UMIFRE 6 / USR 3135





**APPEL À CANDIDATURE**  
Appel à candidature :  
affectation de chercheur-e-s  
CNRS à l'Ifpo (11/02/2019)



**PRESSES DE L'IFPO / PUBLICATIONS**  
Sirat al-Malik al-Zahir  
Baybars, tome 16 (texte  
arabe)



**ÉVÉNEMENTS**  
6000 images archéologiques  
de la Syrie en libre accès



**APPEL À CANDIDATURE / LIBAN**  
Offre de stage : assistant-e de  
médiation scientifique (31/01  
/2019)



**CONFÉRENCES / ÉVÉNEMENTS / LIBAN**  
Philosophie et mathématique  
dans l'Islam classique (10/01

PHOTOTHÈQUE

## 6000 IMAGES ARCHÉOLOGIQUES DE LA SYRIE EN LIBRE ACCÈS

MÉDIHAL

[HTTPS://MEDIHAL.ARCHIVES-OUVERTES.FR/IFPOIMAGES](https://medihal.archives-ouvertes.fr/ifpoimages)

ÉVÉNEMENTS 18/12/2018

## 6000 images archéologiques de la Syrie en libre accès

Au cours de l'été 2018, l'Institut français du Proche-Orient (Ifpo) a mis en ligne sur MédiHAL plus de 6000 images archéologiques de la Syrie issues de ses fonds documentaires. Ce sont plus de 8500 images scientifiques qui sont aujourd'hui accessibles en libre accès dans la collection de l'Ifpo. La réalisation de ce projet de mise en ligne et de valorisation a été rendu possible grâce au soutien...

PLUS

RECHERCHER

EN DIRECT DES CARNETS DE  
L'IFPO

Corpus d'arabe parlé (2) :  
Contraintes et problèmes liés à  
la réalisation des corpus d'arabe  
parlé en interaction

Qui a tué Asmahan ? d'Amir Nizar  
Zubi : un récit policier sur scène  
(2)

Corpus d'arabe parlé (1) : Quels  
corpus d'arabe parlé en libre  
accès ?

Réformes économiques et  
mobilisations sociales en  
Jordanie

« Qui a tué Asmahan ? » d'Amir

# Ifpo is present in 5 countries, part of the IFRE network



**27** IFRE

**34** pays

**150** chercheurs

**350** doctorants

**800**

manifestations  
scientifiques par an

**+ 45 000**

documents scientifiques  
dans le moteur de recherche

**180**

partenariats locaux

Un réseau unique de 27 centres de recherche français, répartis sur tous les continents, abordant toutes les sciences humaines et sociales, en contact avec les institutions de recherche locales et françaises.



- Présentation ▾
- Services et outils ▾
- Consortiums
- International
- Ressources

## À propos de Huma-Num

Huma-Num est une très grande infrastructure de recherche (TGIR) visant à faciliter le tournant numérique de la recherche en sciences humaines et sociales. Au cœur des SHS et des humanités numériques, elle est bâtie sur une organisation originale consistant à mettre en œuvre un dispositif humain (concertation collective) et technologique (services numériques pérennes) à l'échelle nationale et européenne en s'appuyant sur un important réseau de partenaires et d'opérateurs.

[Lire la suite](#)

## SUR NOTRE BLOG

- 20 déc 2018  
Le jeu de cartes du CIDOC en ligne
- 20 déc 2018  
Indexer les corpus numériques – 1er février 2019 – Saint-Etienne
- 19 déc 2018  
Appel à participants – Second atelier PARTHENOS (H2020)



6415 BOOKS | 501 JOURNALS | 2764 BLOGS | 40406 EVENTS

All OpenEdition ▾  SEARCH

OpenEdition brings together four platforms dedicated to electronic resources in the humanities and social sciences. If you wish your university to subscribe to additional services and give you access to downloadable formats (PDF, ePub), please visit our [Institutions section](#).

[FURTHER INFORMATION](#)

## OpenEdition Books

Focus	Latest selections	Latest excerpts
 Sandra Costa, Claire Maitre (ed.) Éditions du Comité des travaux historiques et scientifiques, 2018	26 OCTOBRE 2018 <b>CLIMATE AND ENVIRONMENT</b> A selection of titles that shed light on this major 21st-century issue. Regardless of whether they discuss global warming, biodiversity or sustainable development, and whether they	<i>It now seems obvious, for example, that simple generalizations about the collective nature of art in non-literate cultures cannot hold good in face of the evidence about the creative ac-</i>

**CCSD** CENTRE POUR LA COMMUNICATION SCIENTIFIQUE DIRECTE

La Science partagée

[HOME PAGE](#) | [ABOUT US](#) ▾ | [OUR SERVICES](#) ▾ | [BLOG](#) | [PRACTICAL INFORMATIONS](#)

## The Center for Direct Scientific Communication (CCSD)

works for the researchers/communities and their institutional environment (research organizations, universities). Its main mission is to provide tools, in the respect of open access principles, for archiving and the dissemination of scientific publications and data.

The CCSD has created, develops and manages the HAL open archive, the conference management platform SciencesConf.org and the hosting platform of epi-journals Episciences.org. It is part of a network of national and international partners, stakeholders and operators of scientific and technical information.

Created by the CNRS in 2000, the CCSD has become a service unit (UMS33668) joining INRIA and the University of Lyon. It is a major actor in the national policy of free access.

- <https://www.huma-num.fr/>
- <https://www.openedition.org/>
- <http://persee.fr/>
- <https://www.ccsd.cnrs.fr/>

OPEN ARCHIVES	CONFERENCES MANAGEMENT	EPI JOURNALS



Accueil À propos ▾ About ▾ Événements ▾ Billets ▾ Dictionnaire



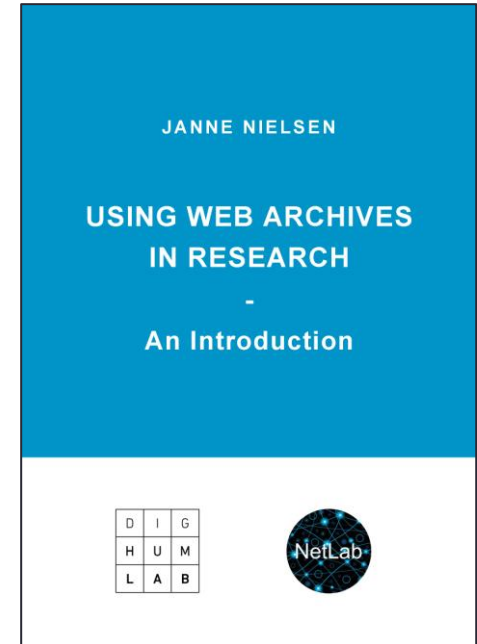
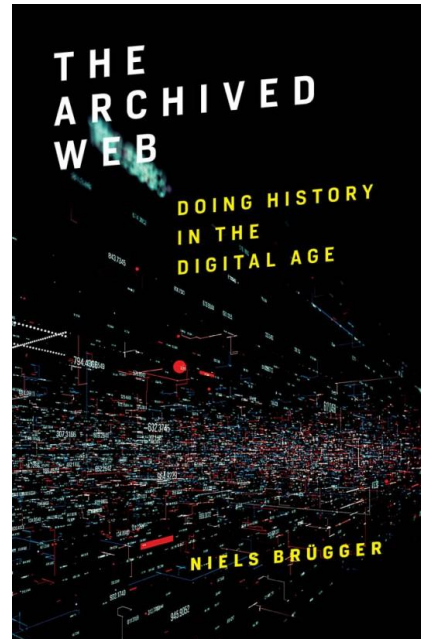
## Confessionnalisme et nationalisme en Syrie depuis 2011 (21-25/03/2019)

Recherche...

Cycle de Conférence organisée par l'ANR [Shakk](#) dans le cadre des Rendez-

MOTS-CLEFS

# Three main references for this workshop



- Musiani, F., Paloque-Bergès, C., Schafer, V., & G. Thierry, B. (2019). *Qu'est-ce qu'une archive du web ?* Retrieved from <http://books.openedition.org/oep/8713> ; DOI : 10.4000/books.oep.8713
- Brügger, N. (2018). *The Archived Web*. Retrieved from <http://www.worldcat.org/oclc/1057466800>
- Nielsen, J. (2016). *Using web archives in research: an introduction*. Retrieved from <http://www.worldcat.org/oclc/960018046>



- “The [first-ever website \(info.cern.ch\)](http://info.cern.ch) was published on August 6, 1991 by British physicist Tim Berners-Lee while at CERN, in Switzerland”
- “There are over 1.5 billion websites on the World Wide Web today. Of these, less than 200 million are active”  
(<http://www.internetlivestats.com>)
- “In the April 2019 survey we received responses from 1,445,266,139 sites across 233,886,577 unique domain names and 8,613,630 web-facing computers. **This reflects a loss of 16.8 million sites** [from previous March 2019 survey]”  
<https://news.netcraft.com/archives/2019/04/22/april-2019-web-server-survey.html>)
- “The Indexed Web contains at least 5.22 billion pages”  
(Tuesday, 30 April, 2019, <https://www.worldwidewebsite.com/>).

- “The average life span of a Web page is only 44 days, and 44 percent of the Web sites found in 1998 could not be found in 1999. [...] As ubiquitous as the Web seems to be, it is also ephemeral, and much of today's Web will have disappeared by tomorrow.” (Lyman, 2002 p. 38)
- “40% of the material on the Internet disappears within a year, while another 40% has been changed, which is why today we can only expect to find 20% of the material that was on the Internet one year ago.” (Brügger, 2005 p. 15)
- “We now know that Web pages only last about 100 days on average before they change or disappear.” (Kahle, 2015)
- In 2013, the average life span of a URL is 9.3 years (Musiani at al., 2019, <https://books.openedition.org/oep/8743>)
- In 2019, according to the Wayback Machine team the life span of a Web page is 92 days

# Many reasons for Web content disappearance

The New York Times

## *In Supreme Court Opinions, Web Links to Nowhere*

By Adam Liptak

Sept. 23, 2013



WASHINGTON — Supreme Court opinions have come down with a bad case of link rot. According to [a new study](#), 49 percent of the hyperlinks in Supreme Court decisions no longer work.

This can sometimes be amusing. A link in [a 2011 Supreme Court opinion](#) about violent video games by Justice Samuel A. Alito Jr. now leads to [a mischievous error message](#).

“Aren’t you glad you didn’t cite to this Web page?” it asks. “If you had, like Justice Alito did, the original content would have long since disappeared and someone else might have come along and purchased the domain in order to make a comment about the transience of linked information in the Internet age.”

<https://www.nytimes.com/2013/09/24/us/politics/in-supreme-court-opinions-clicks-that-lead-nowhere.html>

Support The Guardian Available for everyone, funded by readers

Search jobs Sign in Search The Guardian International edition

Advertisement

The essential news magazine from The Guardian. **Subscribe now**

Try 6 issues for \$6

Conservatives

### Conservative party deletes archive of speeches from internet

Decade's worth of records is erased, including PM's speech praising internet for making more information available

Randeep Ramesh and Alex Hern

Wed 13 Nov 2013 15:40 GMT

5540 1284



▲ A speech in which David Cameron said the internet would help people hold politicians to account was among those deleted. Photograph: Barcroft Media

The Conservatives have removed a decade of speeches from their website and from the main internet library - including one in which David Cameron claimed that being able to search the web would democratise politics by making "more information available to more people".

Editorially

SAMSUNG

Buy any Galaxy S10 model

Galaxy S10

To get a FREE Galaxy Tab E

Buy Now

<https://www.theguardian.com/politics/2013/nov/13/conservative-party-archive-speeches-internet>

## A problem for research and scholarly communication

- Error 404, Broken links, Link rot, Reference rot, Infosuicide, digital ruins, content drift, zombie media,..
- Shut down & take down, mergers and acquisitions:
  - On March 18, 2019, it was revealed that MySpace lost all of their user content from 2016 and earlier in "a server migration gone wrong". It was widely reported that over 50 million songs and 12 years worth of content was permanently lost, and there was no backup (<https://en.wikipedia.org/wiki/Myspace>)
- History :
  - Yugoslavia (.yu) breakup (now Serbia and Montenegro, .rs and .me)
  - Czechoslovakia (.cs) dissolution (now Czech Republic and Slovakia, .cz and .sk)
- **Reference rot**, a combination of:
  - **Content decay**: The content of the linked resource may change over time and, as a result, the degree to which that content remains representative of the content that was intended to be linked to may decrease over time.
- **Link rot**: The linked resource may disappear altogether. (Thoughts on Referencing, Linking, Reference Rot <http://mementoweb.org/missing-link/>)
- The integrity of research is at risk ! (James G. Neal, <http://library.ifla.org/id/eprint/907>)

# Questions

- What is your interest in Web archiving?
- Did you ever experience a Web content loss?
- What is your experience in Web archiving?

# E-corpus 2009-2016

INTERNET ARCHIVE WayBackMachine  Go NOV JAN FEB 03 2017 2018 2019 688 captures 9 Oct 2009 - 27 Feb 2018 Create an account Login About this capture



Search | News | Links | Dedicated websites | Contribute to E-corpus | Contact | Language : English

## E-corpus : une source vive de documentation pour tous les chercheurs et les amateurs

Le taux de consultation de E-corpus augmente, ses besoins aussi, n'hésitez pas à vous manifester pour nous montrer votre intérêt. Si vous êtes partenaires nous serions heureux d'avoir votre retour. Si vous êtes lecteurs et que vous avez trouvé de quoi attiser votre intérêt communiquez le nous via l'onglet contact ou sur les réseaux sociaux Facebook et Twitter.

**Content resume**  
→ 135 780 notices  
→ 2 775 531 documents  
→ 23 notices added recently  
→ 220 documents added recently

**Search in all collections** (collections, notices, printed books...)  
  
Advanced search      Activate virtual keyboard

## E-CORPUS

**e-corpus** is a collective digital library that catalogs and disseminates numerous documents: manuscripts, archives, books, journals, prints, audio recordings, video, etc.

This diverse platform presents cultural diversity worldwide and specifically in the Euro-Mediterranean region. e-corpus proposes a variety of themes and a large quantity of digital documents presented by numerous organizations and countries. Based on cutting edge technologies, e-corpus offers a simple and direct public access to a rich collection of resources.



**Virtual collections** [View all](#)  
These are thematic collections combining documents from several organizations (reconstitution of

**Partner organizations** [View all](#)  
"These organizations disseminate via e-corpus their digital documents (if you would like to join e-corpus.

**NEWSLETTER SUBSCRIPTION**  
E-mail

**TREASURES**

- Bailly de Suffren's note book from its expedition to India**  
Journal de bord du Héros, vaisseau commandant l'escadre du bailli de Suffren durant la campagne des Indes
- Medical books**  
A selection of some historical documents on medicine in Mediterranean countries
- Seals from Murcia Archives (Spain)**  
Seals from Murcia Archives (Spain)

[View all treasures](#)

- To maintain our digital cultural heritage
- To stabilize and preserve web materials as a research object
- To be able to document and illustrate claims based on analyses of web materials (whether the web itself is the research object or a source of knowledge about other research objects).

- Web archiving is the process of collecting portions of the World Wide Web, preserving the collections in an archival format, and then serving the archives for access and use. (IIPC Web site, <http://netpreserve.org/web-archiving/>)
- “Web archiving is the process of gathering up data that has been recorded on the World Wide Web, storing it, ensuring the data is preserved in an archive, and making the collected data available for future research.” (Niu, J. (2012). An Overview of Web Archiving. *D-Lib Magazine*, 18 (3/4). <https://doi.org/10.1045/march2012-niu1>)



**1537** Legal deposit in France (1619 Spain, 1710 United Kingdom)

**1989** The World Wide Web was invented by Tim Berners-Lee

**1996** Internet Archive is founded by Brewster Kahle

**1996** Kulturarw3 in Sweden for archiving the .se top level domain name

**1998** Google is launched

**2001** The Wayback Machine (Internet Archive)

**2003** UNESCO Charter on the Preservation of Digital Heritage

**2003** the International Internet Preservation Consortium is formally chartered at the National Library of France with 12 participating institution

**2005** Youtube is launched

**2006** [The] Facebook and Twitter are launched

**2006** National Library of France is in charge of collecting and preserving the “French Internet” (new French Copyright Law) alongside with the National Audiovisual Institute (Ina)

**2013** EU Web Archive

- Macro archiving
- Micro archiving
- Thematic or selective archiving
- Bulk or snapshot harvesting, or broad crawls
- “Exhaustivity” of a National domain name archiving (.se in Sweden)
- Event and ‘real-time’ institutional archiving (after 2015 terrorist attacks and Notre Dame fire in 2019 France)
- Shared archiving among institutions (in France btwn BnF and Ina)
- ...



**BnF** Dépôt légal web BnF ✓  
@DLwebBnF

Abonné

Merci à [@BNE\\_biblioteca](#) et à toutes les institutions membres de [@NetPreserve](#) qui ont participé à cette collecte par le biais d'une sélection de sites web !

**Biblioteca Nacional de España** ✓ @BNE\_biblioteca  
Estamos colaborando con la Biblioteca Nacional de Francia en una recolección web sobre el incendio de Notre Dame  
Afficher cette discussion

17:02 - 25 avr. 2019

8 Retweets 15 J'aime

🗨️ 8 🍷 15 ✉️

- Finnish Web Archive (since 2006)

<https://www.kansalliskirjasto.fi/en/collections-and-content-online#finnish-web-archive>

The contents of the archive can be only accessed from special legal deposit workstations that are available in selected libraries within Finland (including The National Library of Finland).

- Portugal (since 2008) accessible in Open Access at [Arquivo.pt](https://arquivo.pt)
- The Wayback Machine (since 2006) accessible in Open Access at <https://archive.org/web/>

## There is a lot of different types of Web archives

As much as for other kind of archives, one must know the history of a Web archive and how it was constructed to better understand it and use it in research work. What you see is a reconstruction, not a copy of the site

- “What is harvested is both a point in time (the time of harvesting) and a period of time (the period up to the time of harvesting).” (Brügger, 2008 p. 158)
- “On the one hand the archive does not look like the internet as it actually was in the past (we have lost something), but on the other hand the archive might look like the internet as it never was in the past (we get something different).” (Brügger, 2001 p. 6)

Web archiving projects often needs to gather diverse and multiples expertises and skills : archivists and librarians, researchers, legal officers, IT and computer specialists and... users and stakeholders

As for any other kind of archives one must act lawfully and ethically when archiving and using Web archives:

- The materials in the web archives are protected by copyright law as they were on the live web
- There is “tensions around the archival principles of preserving the public record vs the individual’s expectation of the right to be forgotten”  
(<http://netpreserve.org/ga2018/programme/abstracts/#paper21>)
- The processing of personal data is submitted to laws and even more to the research project ethics
- New laws to take into account ex. General Data Protection Regulation (GDPR, <https://gdpr-info.eu/>)



The web is a unique  
and dynamic resource that  
is of high value to current  
and future researchers



[Learn about the value of our work](#)

## IIPC Members

**Members** are organizations from over 45 countries, including national, university and regional libraries and archives.

## Working groups

IIPC members join **working groups** that engage in short and long-term **projects** to advance the practice of **web archiving**.

## Events

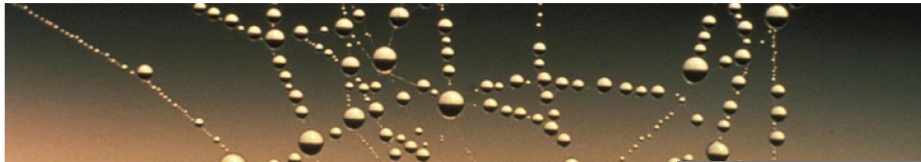
Our community comes together annually to share experiences and present solutions during the **Web Archiving Conference** and the **General Assembly**.

# New research fields and projects



## Web90 – Patrimoine, Mémoires et Histoire du Web dans les années 1990

"Technological progress has merely provided us with more efficient means for going backwards." — Aldous Huxley, Ends and Means



Le projet ANR Web90   Recherches en cours   TTOW Symposium   Veille   L'équipe   Crédits

Publié le 20 août 2018 par Valérie Schafer

 jahendler  
@jahendler

Tim Berners-Lee and I had a paper rejected because reviewer said we didn't really understand how the web worked.

<https://web90.hypotheses.org/>

## Home

Recherche

### RESAW

A Twitter list by @resaw\_eu  
Members tweet about...

**UK Web Archive**  
@UKWebArchive  
Web Archiving Roundup: April, 2019  
[webarchivingrt.wordpress.com/2019/04/29/web...](http://webarchivingrt.wordpress.com/2019/04/29/web...) via @WebArch\_RT

**Web Archiving Roundup: April, 2019**  
The Library of Congress Digital Collections Development Coordinator position closes soon (May 1st)! Archive-It is hosting a training webinar on Web Archiving Systems API

Stories and News

Peter Webster, Web Historian, UK

A. S. Byatt's faded church

Ian Milligan, Web Historian, Canada

New Grant: "Continuing Education to Advance Web Archiving"

<http://resaw.eu/>

## Welcome to the Archives Unleashed Project

 The Archives Unleashed Project

Home

Welcome

Contact Us

About the Project

Archives Unleashed Toolkit

Archives Unleashed Cloud

Archives Unleashed Notebooks

Warlight

Get Involved

Events

Publications

## Welcome to the Archives Unleashed Project



## Welcome

Archives Unleashed aims to make petabytes of historical internet content accessible to scholars and others interested in researching the recent past. Supported by a grant from the [Andrew W. Mellon Foundation](#), we are developing web archive search and data analysis tools to enable scholars, librarians and archivists to access, share, and investigate recent history since the early days of the World Wide Web.

<https://archivesunleashed.org/>

# Technologies for Web archiving: Heritrix and many other tools

internetarchive / heritrix3

Watch 173 Star 1,454 Fork 601

Code Issues 19 Pull requests 2 Projects 0 Wiki Insights

Heritrix is the Internet Archive's open-source, extensible, web-scale, archival-quality web crawler project. <https://heritrix.readthedocs.io/>

java webcrawling warc heritrix

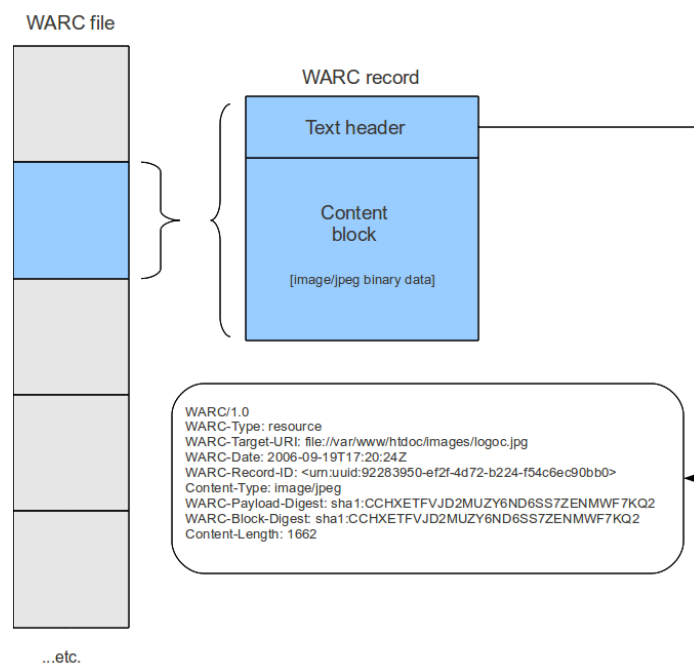
## Heritrix is the Internet Archive's open-source, extensible, web-scale, archival-quality web crawler project

commons	[maven-release-plugin] prepare for next development iteration	9 days ago
contrib	[maven-release-plugin] prepare for next development iteration	9 days ago
dist	[maven-release-plugin] prepare for next development iteration	9 days ago
docs	Note feature only applies to forthcoming 3.3 release	6 months ago
engine	[maven-release-plugin] prepare for next development iteration	9 days ago
modules	[maven-release-plugin] prepare for next development iteration	9 days ago
.gitignore	Format HTTP request lines in API guide	10 months ago
.travis.yml	allow travis-ci jdk7 failures because contrib...	a month ago
CHANGELOG.md	Update changelog for 3.4.0-20190418	9 days ago

<https://github.com/internetarchive/heritrix3>



- WARC file format = Web ARChive archive format
- ARC was accepted as an international standard in 2009 (ISO 28500:2009)
- WARC is now recognised by most national library systems as the standard to follow for web archival



[https://en.wikipedia.org/wiki/Web\\_ARChive](https://en.wikipedia.org/wiki/Web_ARChive)

[https://wiki.archivematica.org/Significant\\_characteristics\\_of\\_websites](https://wiki.archivematica.org/Significant_characteristics_of_websites)

<https://wiki.archivematica.org/File:WARCdiagram.png>

- Robots.txt
- Captcha (ie Completely Automated Public Turing-test to tell Computers and Humans Apart)
- User interaction needed
- Password protected content
- Technologies and dynamic content : Flash, java scripts,...
- Distant content
- Temporal inconsistencies
- Bot traps
- ...

- Define a strategy
- Use a log and write throughout the life of the project
- You may need to use additional methods and tools
  
- Screen capture and screen recording
- Link crawling
- On demand archiving
  
- We will practice:
  - The Wayback machine
  - Web recorder
  - Archive IT

The structure of URLs on the World Wide Web (www):

**protocol**://**subdomain**.**domain**.**top-level domain**/**path**/**page**/

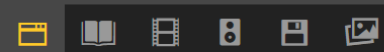
Ex:

**https**://**dhibeirut**.**wordpress**.**com**/**archive**/**dhi-b-2017**/

Search the history of over 351 billion web pages on the Internet.

WayBackMachine

enter URL or keywords



SIGN IN



Search

ABOUT

CONTACT

BLOG

PROJECTS

HELP

DONATE

JOBS

VOLUNTEER

PEOPLE



Internet Archive is a non-profit library of millions of free books, movies, software, music, websites, and more.



351B



20M



5.0M



5.2M



1.8M



417K



3.3M



202K



419K

Search

GO

[Advanced Search](#)

## Announcements

Official EU Agencies Falsely Report More Than 550 Archive.org URLs as Terrorist Content

Boston Public Library's 78rpm Records Come to the Internet: Reformatting the Boston Public Library Sound Archives

Google Plus (or Minus) and the Ephemerality of Community

[SEE MORE](#)

Terms of Service Dec 31, 2014

- Launched in 2001
- 357 billion archived [Web pages](#) so far
- Archived content going back to 1996

## Structure of URLs in the Internet Archive's Wayback Machine:

Wayback Machine URL/collection/ time shown as  
yyymmddhhmmss/URL

Ex:

<https://web.archive.org/web/19980118071158/http://www.aub.edu.lb/>

# Screen capture with Firefox

The screenshot shows a Firefox browser window with the address bar displaying `https://dhibeirut.wordpress.com`. The browser's menu bar is open, showing various options. The main content area of the browser displays a tutorial for screen capturing. The tutorial text reads: "Sélectionnez une zone de la page par cliquer-glisser ou en cliquant sur l'élément à sélectionner. Appuyez sur Échap pour annuler." Below the text is a button labeled "Annuler". The background of the tutorial is a dark map of Beirut with a network overlay of black dots and lines. The browser's address bar shows several bookmarks, including "Les plus visités", "Débuter avec Firefox", "Launch Hypothesis", "Altmetric it", and "Le patrimoine culturel...". The browser's toolbar includes icons for back, forward, refresh, home, search, and other functions. The Windows taskbar is visible at the bottom, showing the search bar with the text "Taper ici pour rechercher" and several application icons. The system tray at the bottom right shows the time as 19:31 and the date as 01/05/2019.

Firefox menu options:

- Modifier ce marque-page
- Enregistrer la page dans Pocket
- Copier le lien
- Envoyer par courriel un lien vers la page...
- Envoyer l'onglet à un appareil
- Partager
- Effectuer une capture d'écran
- Ajouter ce moteur de recherche

Tutorial text:

Sélectionnez une zone de la page par cliquer-glisser ou en cliquant sur l'élément à sélectionner. Appuyez sur Échap pour annuler.

Annuler

archive.today  
sauvegarde de page web

Sauvegarde de

1 mai 2019 18:50:17 UTC

Redirigée de

Tous les clichs de l'hôte [dhibeirut.wordpress.com](http://dhibeirut.wordpress.com)

Page web

Capture d'écran

 partager

 télécharger .zip

 signaler une erreur ou un abus

# > DHI B. DIGITAL HUMANITIES INSTITUTE - BEIRUT

HOME REGISTRATION SCHEDULE WORKSHOPS KEYNOTES PANELS PEOPLE SPONSORS MAPS LOGO AND MEDIA  
GENERAL INFO ARCHIVE  CONTACT US PRE-INSTITUTE TALKS PLACES TO VISIT IN LEBANON





# > DHI B.

DIGITAL HUMANITIES INSTITUTE - BEIRUT

HOME   REGISTRATION   SCHEDULE   WORKSHOPS   KEYNOTES   PANELS   PEOPLE   SPONSORS   MAPS  
LOGO AND MEDIA   GENERAL INFO   ARCHIVE ▾   CONTACT US   PRE-INSTITUTE TALKS  
PLACES TO VISIT IN LEBANON



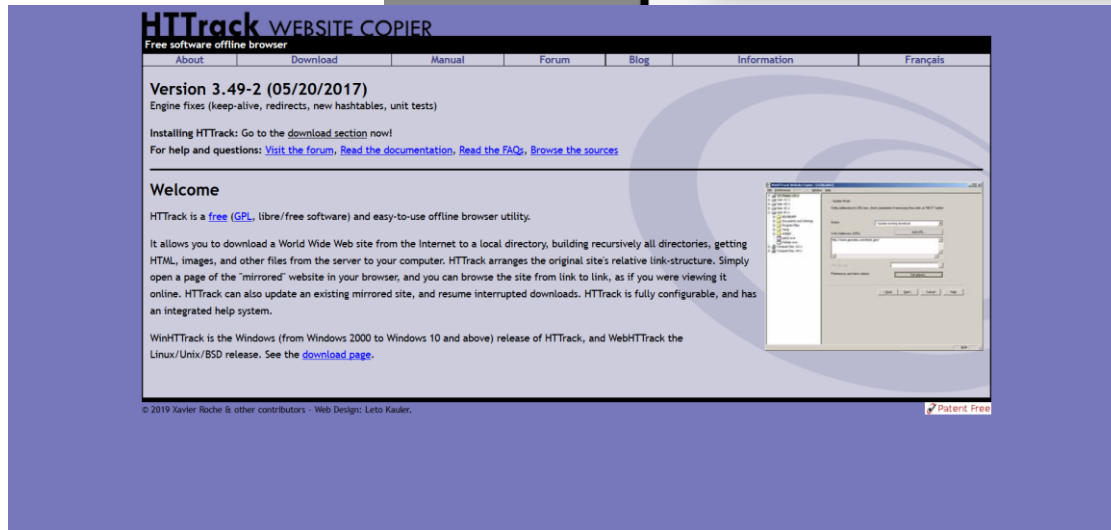
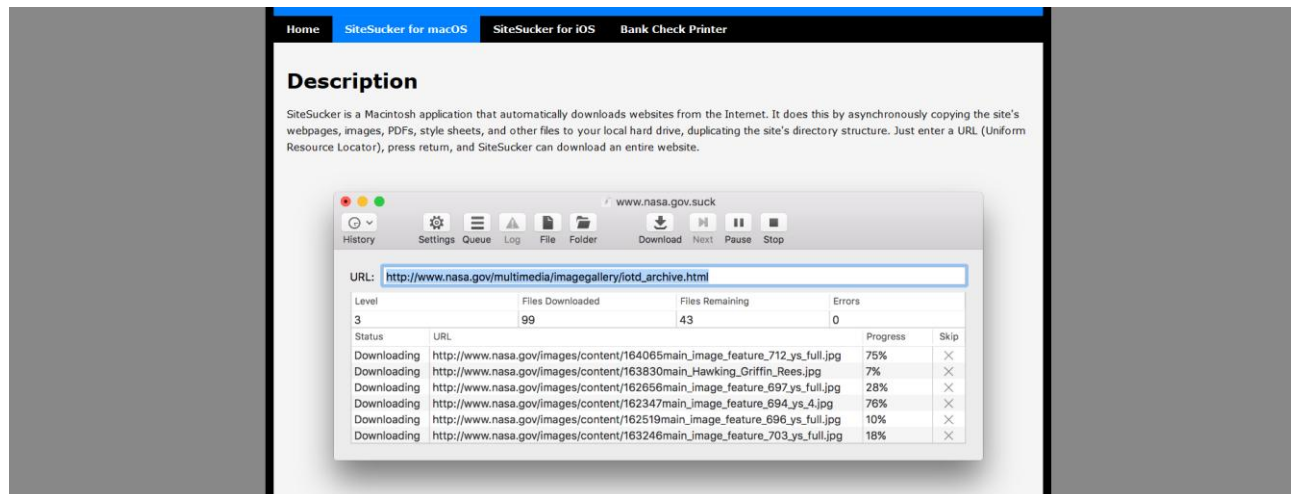
# Screen recording with ShareX (Win) / QuickTime (Mac OS)

The screenshot shows a web browser window displaying the Twitter profile of DHIB Beirut. The profile includes the name 'DHIB', the handle '@DHIBeirut', and the bio 'Official account for #DHIB2017 Digital Humanities Institute Beirut 10-12 March 2017'. The profile has 47 tweets, 143 followers, and 146 subscribers. A ShareX 12.4.1 window is open over the tweets, showing a list of captured files with columns for file name, status, progress, speed, and URL. The files include various screenshots and recordings from the browser. A tweet from 'TerMal' is partially visible at the bottom right.

Nom du fichier	Statut	Progression	Vitesse	Écoulé	Restant	URL
DooeJZ8t1.mp4	Historique	28/04/2019 16:41:21				C:\Users\Ghaylam\Docu...
firefox_miUMpikh7.png	Historique	28/04/2019 21:11:32				C:\Users\Ghaylam\Docu...
firefox_cPCNYcTCwm.png	Historique	28/04/2019 21:14:17				C:\Users\Ghaylam\Docu...
chrome_WSulHjH6Wf.png	Historique	28/04/2019 21:18:09				C:\Users\Ghaylam\Docu...
firefox_GbMF2xKad.png	Historique	28/04/2019 21:23:19				C:\Users\Ghaylam\Docu...
firefox_mdpc7xcVjW.png	Historique	28/04/2019 21:26:32				C:\Users\Ghaylam\Docu...
firefox_YlcQi7ID3O.png	Historique	28/04/2019 21:28:09				C:\Users\Ghaylam\Docu...
firefox_hpAsERoOe9.png	Historique	28/04/2019 21:28:33				C:\Users\Ghaylam\Docu...
firefox_lU8BGTE10.png	Historique	28/04/2019 21:31:08				C:\Users\Ghaylam\Docu...
firefox_Eo6J6TgoNT.png	Historique	28/04/2019 21:33:22				C:\Users\Ghaylam\Docu...
aKkoVwGbU4.png	Terminé					C:\Users\Ghaylam\Docu...
1wa85DHVlk.mp4	Terminé					C:\Users\Ghaylam\Docu...
aH4Ruu6xG0.mp4	Terminé					C:\Users\Ghaylam\Docu...
4FulirCpcV.mp4	Terminé					C:\Users\Ghaylam\Docu...
eW9i4cX9Vn.mp4	Terminé					C:\Users\Ghaylam\Docu...
06BXaTG2Hm.mp4	Terminé					C:\Users\Ghaylam\Docu...
ShareX_TjCGe2s3nr.png	Terminé					C:\Users\Ghaylam\Docu...

<https://github.com/ShareX/ShareX>

# Link crawling with HTTrack (Win) / SiteSucker (Mac OS)



the files it downloads, allowing you to browse a site

se a document that you can use to perform the same  
se the Save command, SiteSucker will pause the  
can restart the download from where it left off by

<http://ricks-apps.com/>  
<https://www.httrack.com/>

# Internet Archive / Wayback Machine extension for browsers

chrome web store



## Wayback Machine

Proposé par : Internet Archive (Archive.org)

★★★★★ 118 | Outils de recherche | 65 287 utilisateurs

Supprimer d

Présentation

Avis

Assistance

Articles Similaires

The screenshot shows the Firefox Add-ons page for the Internet Archive extension. The browser's address bar displays the URL <https://addons.mozilla.org/en-US/firefox/user/12373129/>. The page header includes the Firefox logo and the text "Firefox Add-ons" with navigation links for "Explore", "Extensions", "Themes", and "More...". The main content area features a profile card for "Internet Archive" with the following details: "Add-ons developer", a profile icon, the name "Internet Archive", a homepage link to [archive.org/](http://archive.org/), "User since Jun 21, 2016", "Number of add-ons 1", and an "Average rating of developer's add-ons" of 5 stars. A "Report this user for abuse" button is located at the bottom of the profile card. To the right, a section titled "Extensions by Internet Archive" lists the "Wayback Machine" extension with a description: "Detects dead pages, 404s, DNS failures & a to show archived versions via the Internet. you can archive web pages, and see their most recent & first archives." and a 5-star rating.

The screenshot shows a "Page not available?" error message in a blue box. Below the error message, the text reads: "View a saved version courtesy of the INTERNET ARCHIVE WayBack Machine". The "WayBack Machine" logo is prominently displayed in red and black. Navigation arrows are visible on either side of the error message box.

<https://chrome.google.com/webstore/detail/wayback-machine/fpnmgdkabkmnadcjpehmlllkndpkmiak?hl=fr>

<https://addons.mozilla.org/en-US/firefox/user/12373129/>

The screenshot shows a browser window with the address bar displaying <https://dhibeirut.wordpress.com>. The browser's top bar includes various icons and a search field. The main content area of the browser shows the DHI Beirut website, which features a navigation menu with items like HOME, REGISTRATION, SCHEDULE, WORKSHOPS, KEYNOTES, PANELS, PEOPLE, SPONSORS, LOGO AND MEDIA, GENERAL INFO, ARCHIVE, CONTACT US, and PLACES TO VISIT. The website's logo is prominently displayed as '> DHI B.' with the text 'DIGITAL HUMANITIES INSTITUTE - BEIRUT' below it. The background of the website is a stylized map of Beirut with a network of nodes and connections overlaid on it.

Overlaid on the right side of the browser window is the Internet Archive Wayback Machine extension interface. It features a search bar at the top, followed by several buttons: 'Save Page Now' (green), 'Recent Version' (blue), 'First Version' (orange), 'Overview' (light blue), 'Alexa' (light blue), 'Whois' (light blue), 'Tweets' (blue with a Twitter icon), 'Site Map' (white with a magnifying glass icon), and social media icons for Facebook, Twitter, Google+, and LinkedIn. At the bottom of the extension interface is an 'About' button (red).

Browser toolbar showing the address bar with the URL <https://web.archive.org/web/20190428182117/https://dhibeirut.wordpress.com/>. The Wayback Machine extension interface is visible, showing the date **28 APR 2019** and a calendar view. The extension also displays the text "12 captures" and "11 Apr 2015 - 28 Apr 2019".

# > DHIB.

DIGITAL HUMANITIES INSTITUTE - BEIRUT

- HOME
- REGISTRATION
- SCHEDULE
- WORKSHOPS
- KEYNOTES
- PANELS
- PEOPLE
- SPONSORS
- MAPS
- LOGO AND MEDIA
- GENERAL INFO
- ARCHIVE
- CONTACT US
- PRE-INSTITUTE TALKS
- PLACES TO VISIT IN LEBANON

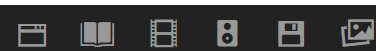


Privacy & Cookies: This site uses cookies. By continuing to use this website, you agree to their use.  
To find out more, including how to control cookies, see here: [Cookie Policy](#)

Close and accept

<https://web.archive.org/web/20190428182117/https://dhibeirut.wordpress.com/#>

<https://web.archive.org/web/20190428182117/https://dhibeirut.wordpress.com/>



 SIGN IN



 Search

[ABOUT](#) [CONTACT](#) [BLOG](#) [PROJECTS](#) [HELP](#) [DONATE](#) [JOBS](#) [VOLUNTEER](#) [PEOPLE](#)



## Sign up for an account

Get your virtual library card! Already have one? [Log In!](#)

### Enter email address

### Choose a screen name (this is how others will see you)

### Choose a password



### Announcements

- Send me general announcements from the Internet Archive (approx. two per month)
- Send me announcements about Internet Archive events in the San Francisco Bay Area

By signing up, you agree to the Internet Archive's [Terms of Service](#).

[Sign up](#)

# Your own collection of archived Web pages in Internet Archive

Internet Archive Wayback Machine interface showing the URL <https://web.archive.org/web/20190428182117/https://dhibeirut.wordpress.com/>. The interface includes a search bar with the URL, a "Go" button, and a calendar navigation showing the date **28** of **APR** 2019. The calendar shows a grid of days from 11 Apr 2015 to 28 Apr 2019, with 12 captures indicated. A "Share via My Web Archive" button is visible on the right.

## > DHI B.

DIGITAL HUMANITIES INSTITUTE - BEIRUT

- HOME
- REGISTRATION
- SCHEDULE
- WORKSHOPS
- KEYNOTES
- PANELS
- PEOPLE
- SPONSORS
- MAPS
- LOGO AND MEDIA
- GENERAL INFO
- ARCHIVE
- CONTACT US
- PRE-INSTITUTE TALKS
- PLACES TO VISIT IN LEBANON



Privacy & Cookies: This site uses cookies. By continuing to use this website, you agree to their use.  
To find out more, including how to control cookies, see here: [Cookie Policy](#)

Close and accept

<https://web.archive.org/web/20190428182117/https://dhibeirut.wordpress.com/#>



# Your own collection of archived Web pages in Internet Archive

Browser navigation bar showing the URL: [https://archive.org/details/@jean\\_christophe\\_peyssard?tab=web-archive&sort=titleSorter](https://archive.org/details/@jean_christophe_peyssard?tab=web-archive&sort=titleSorter)

ABOUT CONTACT BLOG PROJECTS HELP DONATE JOBS VOLUNTEER PEOPLE



## Jean-Christophe Peyssard

archive.org Member

Settings  
Edit

UPLOADS POSTS REVIEWS COLLECTIONS WEB ARCHIVES

▲ SORT BY VIEWS · TITLE · DATE ARCHIVED · CREATOR

SHOW DETAILS

- <a href="#">Save &amp; Share</a> +			
- <a href="http://www.ifporient.org:80/">http://www.ifporient.org:80/</a>	Sep 3, 2004		
-	Mar 22, 2018		
-	May 30, 2018		
- <a href="http://www.ifporient.org/node/2235">http://www.ifporient.org/node/2235</a>	Feb 27, 2018		
- <a href="https://www.facebook.com/syrianarchive/photos/a.1793566007...">https://www.facebook.com/syrianarchive/photos/a.1793566007...</a>	Mar 6, 2018		
-	Mar 22, 2018		
- <a href="http://www.ifporient.org/node/2283?page=4">http://www.ifporient.org/node/2283?page=4</a>	Mar 14, 2018		
- <a href="http://www.ifporient.org/node/2313">http://www.ifporient.org/node/2313</a>	Mar 22, 2018		
- <a href="http://www.ultrasawt.com/مرض-لأجل-الإنسان-ياسر-الزيت-سياق-متصل-سياسة">http://www.ultrasawt.com/مرض-لأجل-الإنسان-ياسر-الزيت-سياق-متصل-سياسة...</a>	May 21, 2018		
- <a href="http://www.ifporient.org/node/2337">http://www.ifporient.org/node/2337</a>	May 28, 2018		
- <a href="http://www.ifporient.org/node/2338">http://www.ifporient.org/node/2338</a>	May 28, 2018		

# Webrecorder

Collect & Revisit the Web

Add to collection

Select browser

▸ Session settings

Start

## Our Mission

### Web Archiving for All!

Webrecorder is a web archiving service anyone can use for free to save web pages. Making a capture is as easy as browsing a page like you normally would. Webrecorder automatically archives the page, along with any additional content triggered by interactions.

This open-source project is brought to you by [Rhizome](#) at the [New Museum](#).

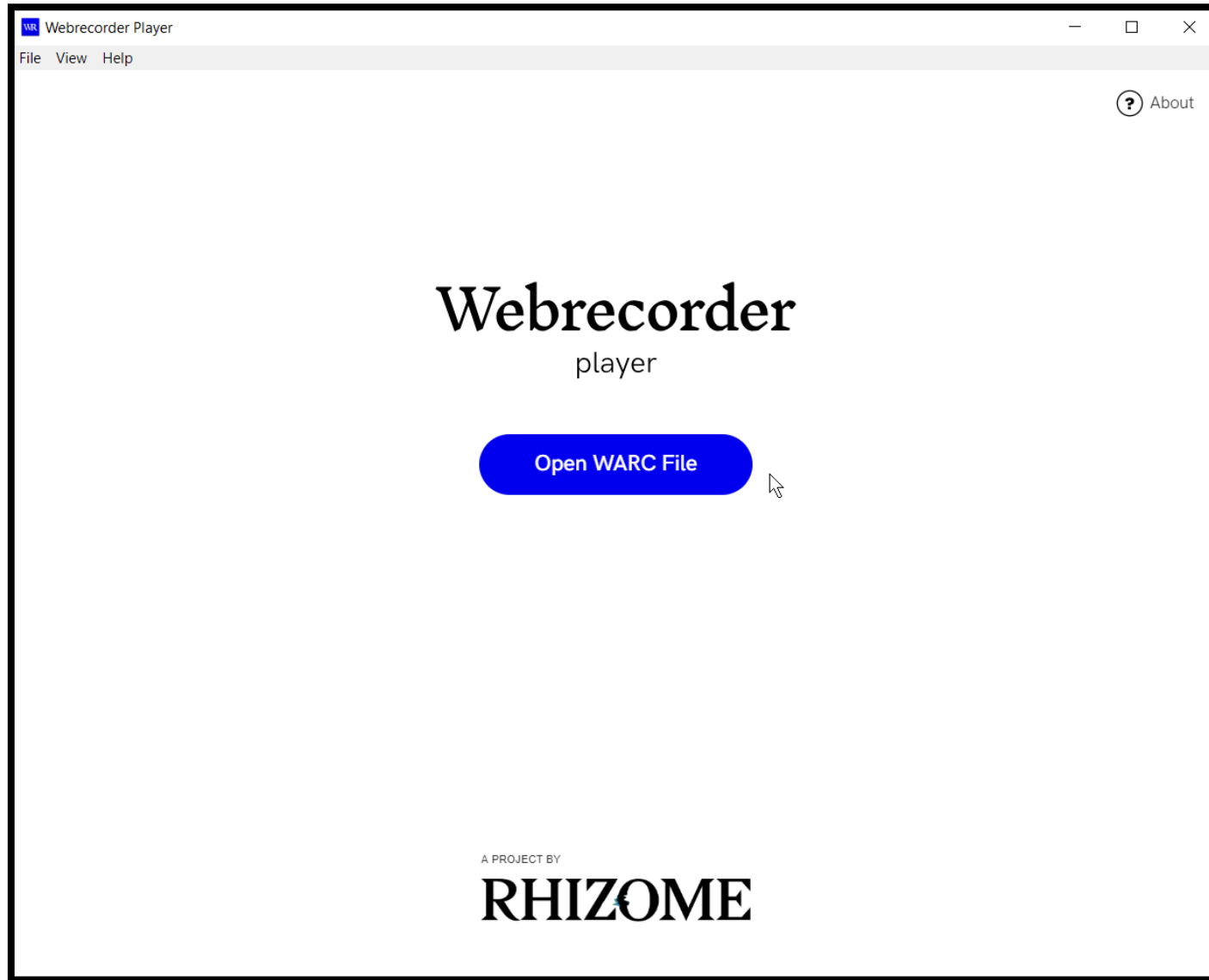
The [Andrew W. Mellon Foundation](#) is lead supporter of the Webrecorder initiative. Additional outreach and research is made possible by the [Knight Foundation](#) and the Institute of Museum and Library Services.



### Online Now ≠ Online Tomorrow

Links break. Services disappear and redesigns happen. The web is ephemeral. Make sure that what's important to you will stay available.


# Playing back WARC files with Webrecorder playe



<https://github.com/webrecorder/webrecorder-player>

# Rendering Web archives the old-fashioned way with oldweb.today

oldweb.today

browser  Netscape Navigator on Macintosh ▼  
[about this browser](#)

URL

date  ▼

▼ [About / How it works](#)

**oldweb.today** is built with open source tools:

- [Docker containers](#) to manage user sessions and scale across machines.
- [Wine](#), [Basilisk](#) and [Sheepshaver](#) and [Previous](#) emulators are used to run old browsers on legacy platforms.
- [pywb](#) is used as the web archive replay proxy system.
- [noVNC](#) provides VNC support in HTML over Websockets.
- The [Memento Web API](#) and [MemGator](#) are used to communicate with web archives.
- Random page list compiled with help from [@muffinista](#)

[Full source code available on Github.](#)

The archived web pages are available courtesy of these public web archives:

<http://oldweb.today/>

# Archive-It: a service for collecting and accessing Web archives

The screenshot shows the Archive-It website interface. At the top right, there are social media icons for Facebook, Twitter, and LinkedIn, along with a 'Login' button. The main navigation bar includes the Archive-It logo, 'HOME', 'EXPLORE', 'LEARN MORE', and 'CONTACT US'. A tagline reads: 'The leading web archiving service for collecting and accessing cultural heritage on the web. Built at the Internet Archive', accompanied by the Internet Archive logo.

A prominent orange banner contains the following text:  
Welcome to Archive-It!  
Attend a live informational webinar and demo to learn more about the service  
Contact Us to sign up for an upcoming session:  
mai 02 2019, 11:00 AM PDT  
mai 16 2019, 11:00 AM PDT


Below the banner is a search section for collections: 'Explore Collections' with a search box 'Find a Collection by Name', a 'Search' button, and a link 'Show All Collections'.

The main content area features three featured collections, each with a thumbnail image and descriptive text:  
1. **Clinical Translational Science Award**  
By National Institutes of Health  
The Clinical Translational Science Award Collection document and preserve the progress of the Clinical Translational Science Award

2. **Culture and Sensitivity**  
By NYC Public School 56 Q - The Harry Eichler School  
This collection, from fifth graders at New York City PS 56 in Queens is their student's guide to the wonderful arts and varieties of cultures.


3. **Jasmine Revolution - Tunisia 2011**  
By Internet Archive Global Events  
A collection of websites, news coverage, and commentary surrounding the 2011 Jasmine Revolution in Tunisia. Our partners at

At the bottom, there is a search section for organizations: 'Explore Collecting Organizations' with a search box 'Find an Organization by Name', a 'Search' button, and a link 'Show All Organizations'.




HOME | EXPLORE | LEARN MORE | CONTACT US

The leading web archiving service  
for collecting and accessing  
cultural heritage on the web  
*Built at the Internet Archive*



Explore >> Digital Humanities Institute Beirut 2019



## Digital Humanities Institute Beirut 2019

Archive-It Partner Since: avr., 2019  
Organization URL: <http://dhibeirut.wordpress.com/>  
Description: Digital Humanities Institute Beirut

**Narrow Your Results**

There are no further ways to narrow your results.

Sites and collections from this organization are listed below. Narrow your results at left, or enter a search query below to find a collection, site, specific URL or to search the text of archived webpages.

Collections | Sites | Search Page Text

Page 1 of 1 (2 Total Results)

Sort By: Collection Name (A-Z) | Collection Name (Z-A)

**DHIB 2017**  
Archived since: mai, 2019  
No description.

---

**DHIB 2019**



ARCHIVE-IT

Home

Collections

Crawls

Archives

ARS

Help Center

Welcome, Jean-Christophe Peysard

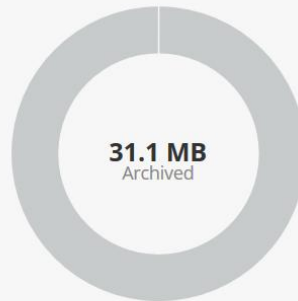
Home

## Digital Humanities Institute Beirut 2019

31.1 MB archived since Apr 26, 2019

InstaCrawl

### Current Subscription



Data Budget Usage

▶ Current Subscription Details

▶ Past Subscription Totals

### Active Collection List (2 Active Collections)

Type to Filter Active Collections

Create a Collection

Download Collection List

Collection Name	Data (this period) ▾	Docs (this period)	Active Seeds	Last Crawl
DHIB 2019	31.1 MB	469	13	May 1, 2019
DHIB 2017	0 bytes	0	5	May 1, 2019

Aide

# Archive-It: example of an archived Web page

You are viewing an archived web page, collected at the request of [Digital Humanities Institute Beirut 2019](#) using [Archive-It](#). This page was captured on 12:20:14 mai 01, 2019, and is part of the [DHIB 2017](#) collection. The information on this web page may be out of date. See [All versions](#) of this archived page. [Metadata](#) [Enable QA](#) hide

# > DHIB.

## DIGITAL HUMANITIES INSTITUTE – BEIRUT

[HOME](#)   [REGISTRATION](#)   [SCHEDULE](#)   [WORKSHOPS](#)   [KEYNOTES](#)  
[PANELS](#)   [PEOPLE](#)   [SPONSORS](#)   [MAPS](#)   [LOGO AND MEDIA](#)  
[GENERAL INFO](#)   [ARCHIVE](#) ▾   [CONTACT US](#)   [PRE-INSTITUTE TALKS](#)  
[PLACES TO VISIT IN LEBANON](#)



Privacy & Cookies: This site uses cookies. By continuing to use this website, you agree to their use.  
To find out more, including how to control cookies, see here: [Cookie Policy](#)









Close and accept

<https://wayback.archive-it.org/12102/20190501122014/https://dhibeirut.wordpress.com/archive/dhi-b-2017/>



[Home](#) > [People](#) > [Jean-Christophe Peyssard](#) > [Library](#) > [Web archiving](#)**Library**

- 404
- A lire
- About Zotero
- Aldébaran
- AO
- archives de la recherche
- assesment
- ated-tunis-09-2016
- bad science
- bibliométrie
- book-reviews
- Bookmarks
- Clés du référencement
- Cours IUT
- cours-introduction-HN
- Dakar

<input type="checkbox"/>	Title	Creator	Date Modified
<input type="checkbox"/>	 Archiving websites: general considerations and strategies	Brügger	28/04/2019 15:05
<input type="checkbox"/>	 ASAP – Archives sauvegarde attentats Paris   De #jesuischarl...		07/02/2019 11:58
<input type="checkbox"/>	 Conservative party deletes archive of speeches from internet...		26/04/2019 17:49
<input type="checkbox"/>	 Digital Humanities Institute – Beirut		28/04/2019 21:36
<input type="checkbox"/>	 In Jordan, the “Invisible Hand” Blocks Internet Archive		26/04/2019 17:49
<input type="checkbox"/>	 La sélection de sites web dans une bibliothèque nationale en...	Bonnel and Oury	15/12/2015 09:37
<input type="checkbox"/>	 Legal deposit of the French Web: harvesting strategies for a...	Lasfargues et al.	14/04/2019 18:23
<input type="checkbox"/>	 LibGuides: Preserving Web Sites using WinHTTrack: What is Wi...	Benalayt	14/04/2019 18:42
<input type="checkbox"/>	 List of Web archiving initiatives		21/04/2019 21:53
<input type="checkbox"/>	 Methods of collecting facebook material and their effects on...	Brügger and Sandvik	28/04/2019 14:48
<input type="checkbox"/>	 More than 9 million broken links on Wikipedia are now rescue...		03/02/2019 18:36
<input type="checkbox"/>	 Official EU Agencies Falsely Report More Than 550 Archive.or...		12/04/2019 10:00
<input type="checkbox"/>	 Peut-être n'est-il pas nécessaire d'archiver tous les tweets...		28/01/2018 18:15
<input type="checkbox"/>	 Qu'est-ce qu'une archive du web ?	Musiani et al.	28/02/2019 14:07
<input type="checkbox"/>	 RESAW   A Research infrastructure for the Study of Archived ...		07/02/2019 11:59
<input type="checkbox"/>	 Searching and Using Web Archives	Fuhrig	20/04/2019 16:59
<input type="checkbox"/>	 The integrity of research is at risk: Capturing and preservi...	Neal	15/12/2015 09:38

Thank you / Merci / شكراً

> DHIJ. 2019

[jc.peyssard@ifporient.org](mailto:jc.peyssard@ifporient.org)

