Archiving Web Content

Jean-Christophe Peyssard

ARCHIVING WEB CONTENT

> DHIB. 2019

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

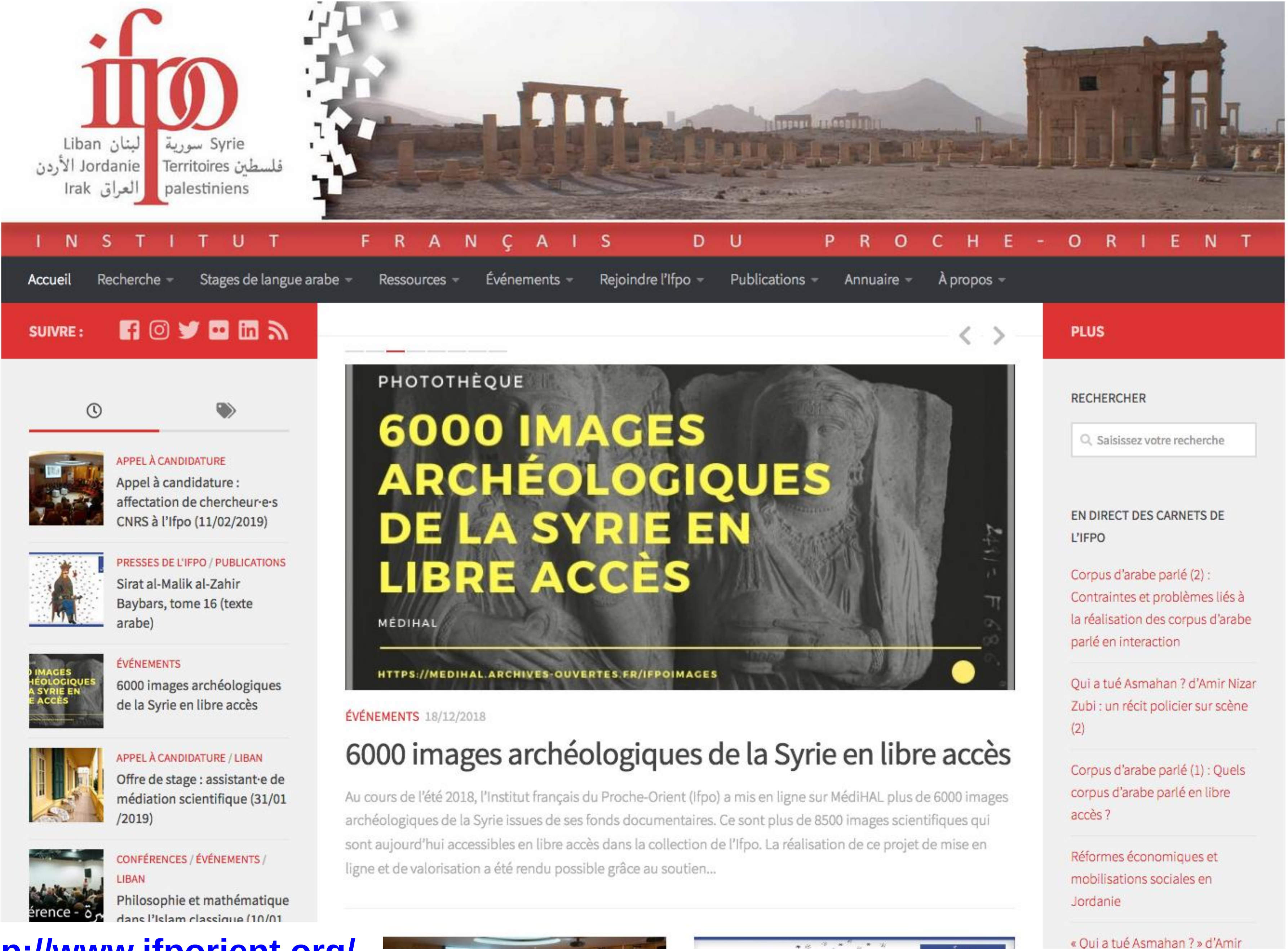
Who Am I ?

Research Ingineer 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
- 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

Ifpo : Institut français du Proche-Orient / French Institute of the Near East



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

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...

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)

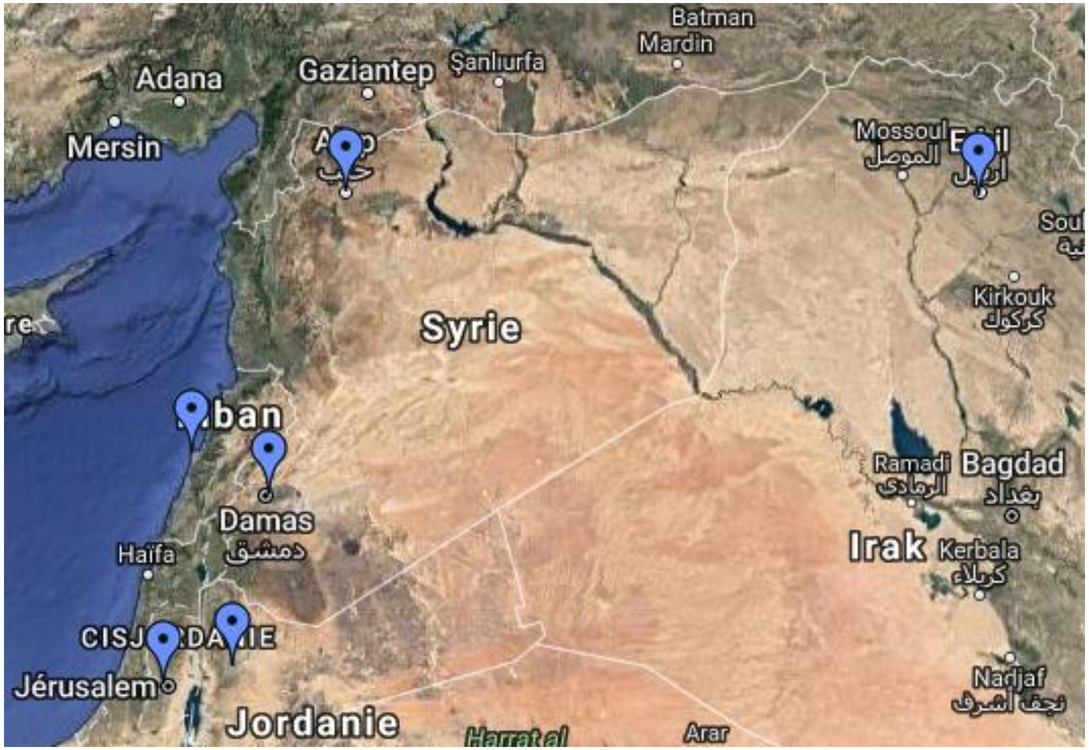
Corpus d'arabe parlé (1) : Quels corpus d'arabe parlé en libre accès ?

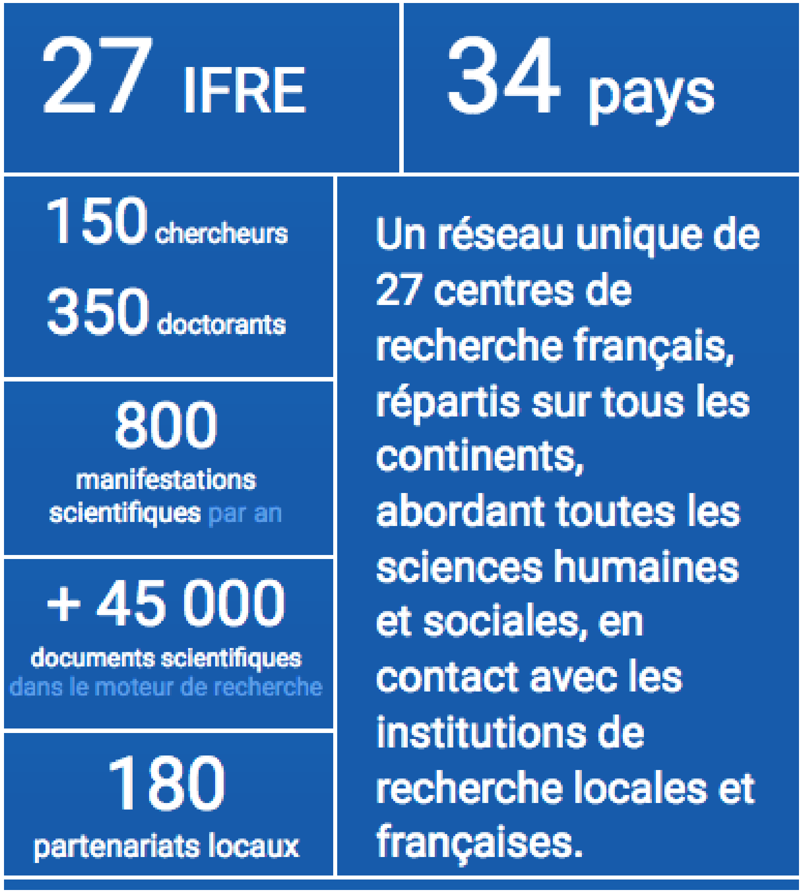
Réformes économiques et mobilisations sociales en Jordanie

« Qui a tué Asmahan ? » d'Amir

http://www.ifporient.org/

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





27 IFRE

34 pays

150 chercheurs

350 doctorants

800 manifestations scientifiques par an

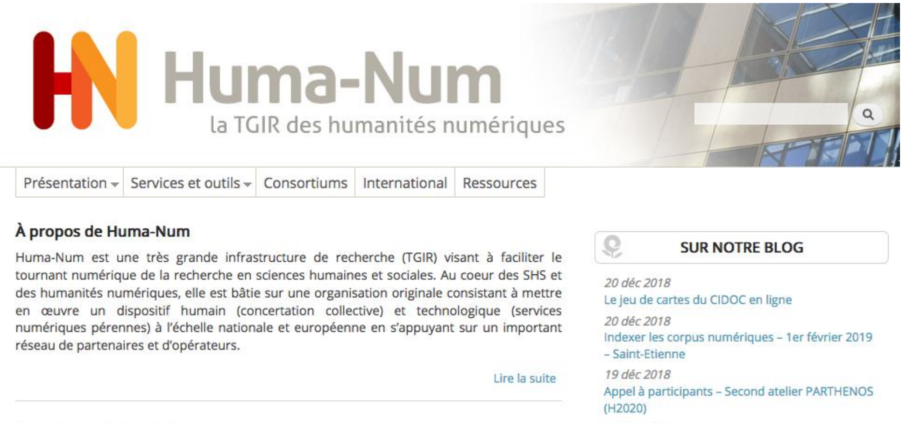
+ 45 000 documents scientifiques dans le moteur de recherche

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.

http://www.ifre.fr/

French research infrastructures for Scholarly Communication & Digital Humanities



Huma-Num
la TGIR des humanités numériques

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 coeur 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.

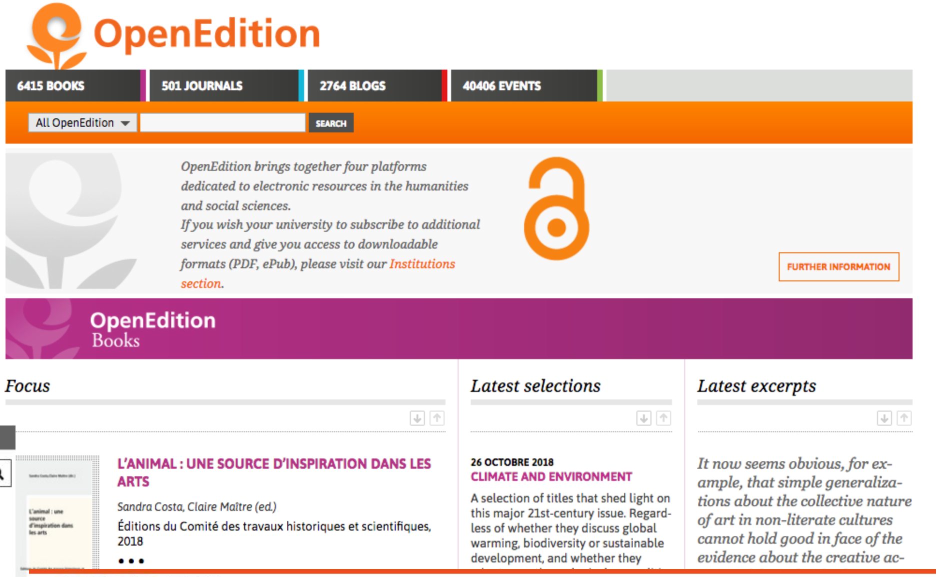
Lire la suite

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)



OpenEdition

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

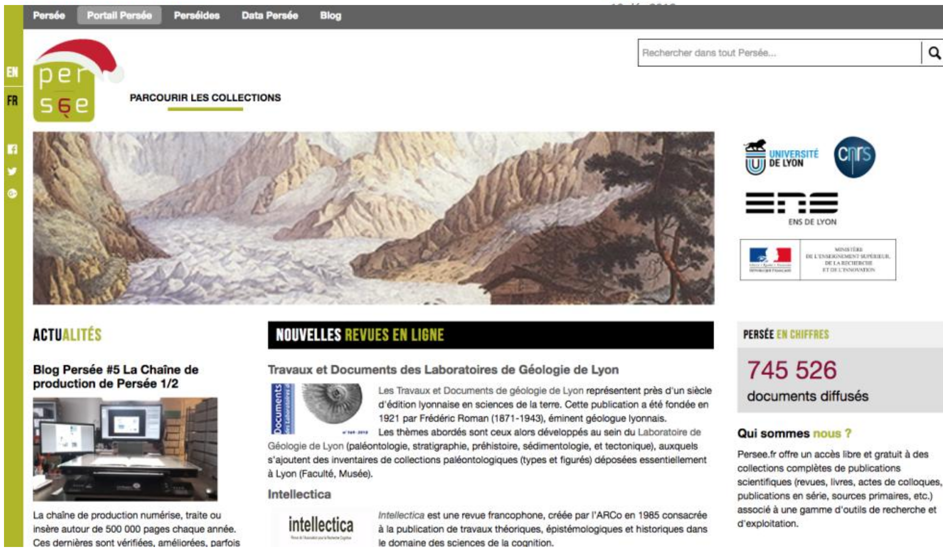
L'ANIMAL : UNE SOURCE D'INSPIRATION DANS LES ARTS
Sandra Costa, Claire Maître (ed.)
Éditions du Comité des travaux historiques et scientifiques, 2018
• • •

Latest selections

26 OCTOBRE 2018
CLIMATE AND ENVIRONMENT
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

Latest excerpts

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-



Persée | Portail Persée | Perséides | Data Persée | Blog

EN | FR

Rechercher dans tout Persée...

PARCOURIR LES COLLECTIONS

UNIVERSITÉ DE LYON | cnrs | ENS DE LYON | MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION

ACTUALITÉS

Blog Persée #5 La Chaîne de production de Persée 1/2

La chaîne de production numérise, traite ou insère autour de 500 000 pages chaque année. Ces dernières sont vérifiées, améliorées, parfois

NOUVELLES REVUES EN LIGNE

Travaux et Documents des Laboratoires de Géologie de Lyon

Les Travaux et Documents de géologie de Lyon représentent près d'un siècle d'édition lyonnaise en sciences de la terre. Cette publication a été fondée en 1921 par Frédéric Roman (1871-1943), éminent géologue lyonnais. Les thèmes abordés sont ceux alors développés au sein du Laboratoire de Géologie de Lyon (paléontologie, stratigraphie, préhistoire, sédimentologie, et tectonique), auxquels s'ajoutent des inventaires de collections paléontologiques (types et figurés) déposées essentiellement à Lyon (Faculté, Musée).

Intellectica

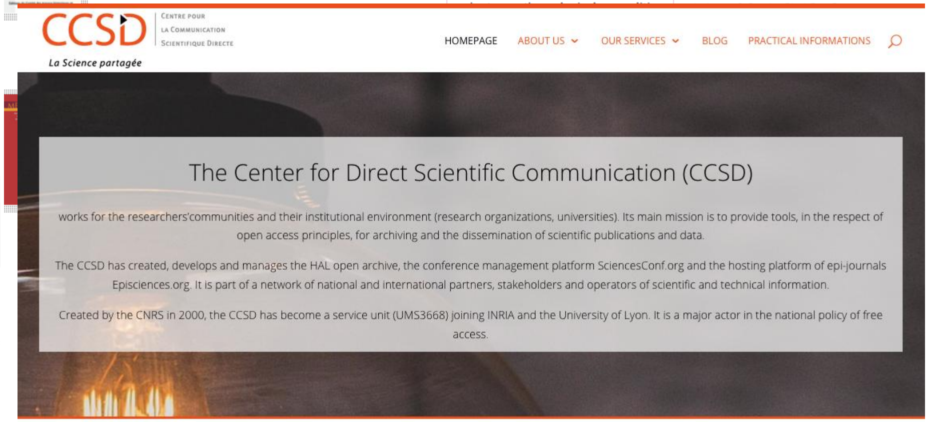
Intellectica est une revue francophone, créée par l'ARCo en 1985 consacrée à la publication de travaux théoriques, épistémologiques et historiques dans le domaine des sciences de la cognition.

PERSÉE EN CHIFFRES

745 526
documents diffusés

Qui sommes nous ?

Persee.fr offre un accès libre et gratuit à des collections complètes de publications scientifiques (revues, livres, actes de colloques, publications en série, sources primaires, etc.) associé à une gamme d'outils de recherche et d'exploitation.



CCSD | Centre pour la Communication Scientifique Directe
La Science partagée

HOMEPAGE | 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.

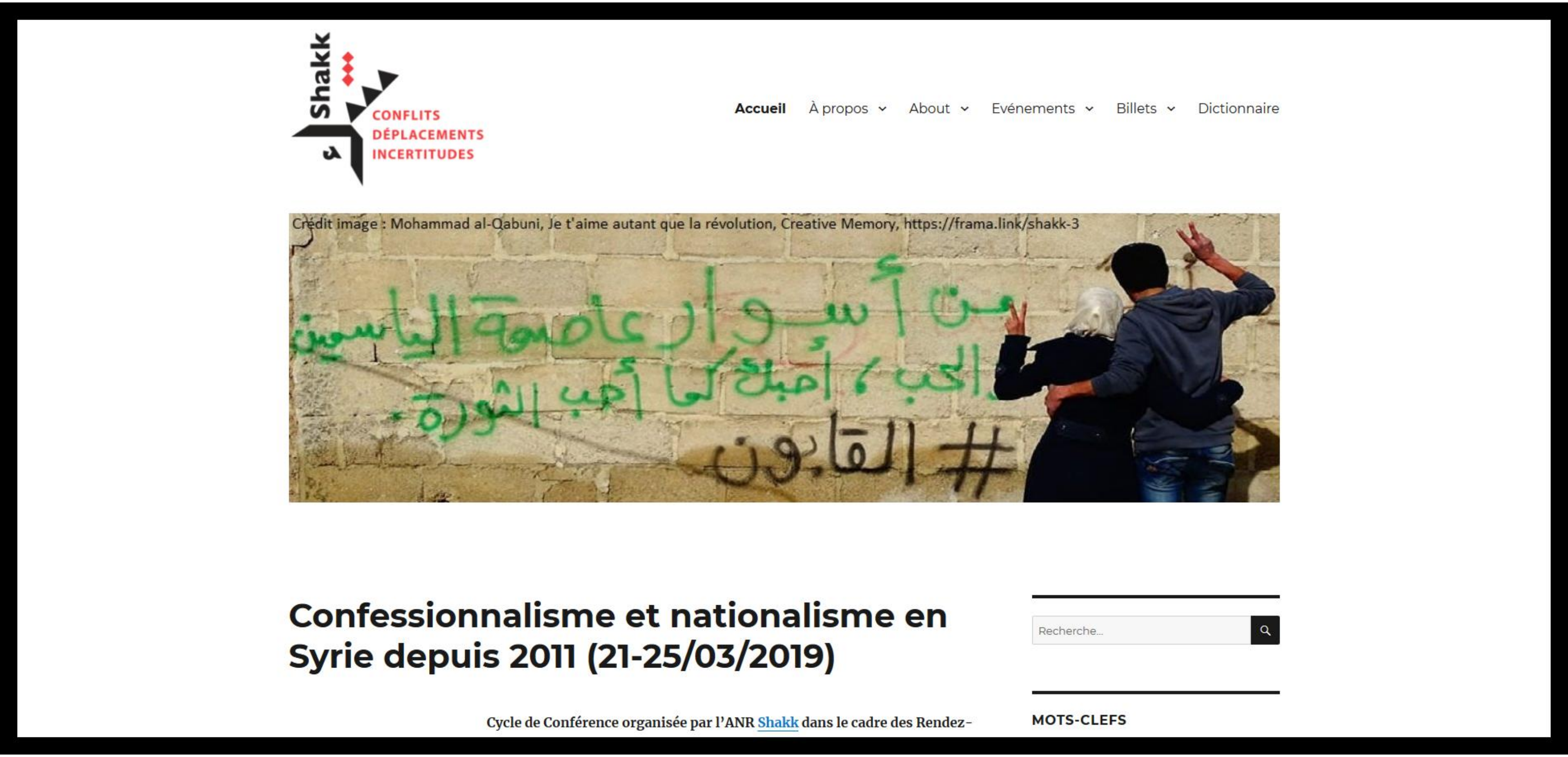
The CCSD has created, develops and manages the HAL open archive, the conference management platform SciencesConf.org and the hosting platform of epi-journals Episciences.org. It is part of a network of national and international partners, stakeholders and operators of scientific and technical information.

Created by the CNRS in 2000, the CCSD has become a service unit (UMS3668) joining INRIA and the University of Lyon. It is a major actor in the national policy of free access.

OPEN ARCHIVES	CONFERENCES MANAGEMENT	EPI JOURNALS
HAL archives-ouvertes.fr	Sciencesconf.org plateforme de gestion de conférences	episciences.org

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

Shakk, from revolt to War in Syria: Conflict, displacements, uncertainties



Crédit image : Mohammad al-Qabuni, Je t'aime autant que la révolution, Creative Memory, https://frama.link/shakk-3

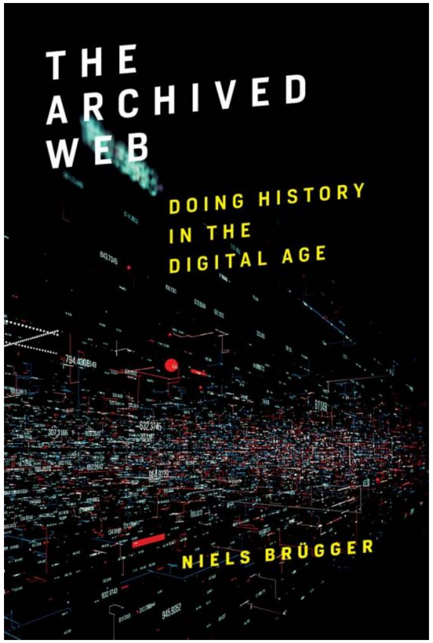
Confessionnalisme et nationalisme en Syrie depuis 2011 (21-25/03/2019)

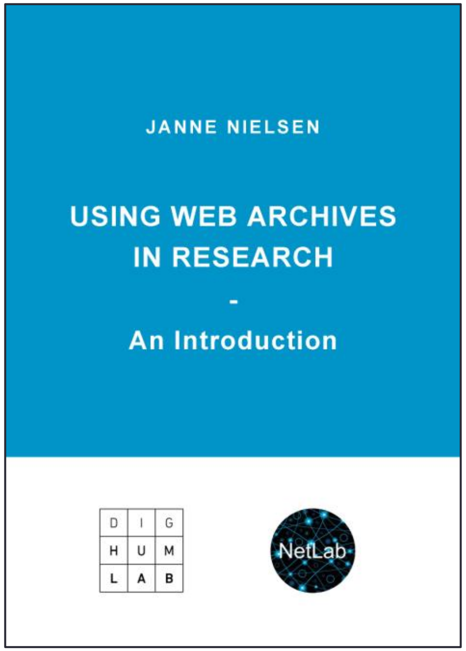
Cycle de Conférence organisée par l'ANR Shakk dans le cadre des Rendez-

https://shakk.hypotheses.org/

Three main references for this workshop

QU'EST-CE QU'UNE
ARCHIVE DU WEB ?
FRANCESCA MUSIANI, CAMILLE PALOQUE-BERGÈS, VALÉRIE SCHAFER, BENJAMIN G. THIERRY
OpenEdition press





- 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

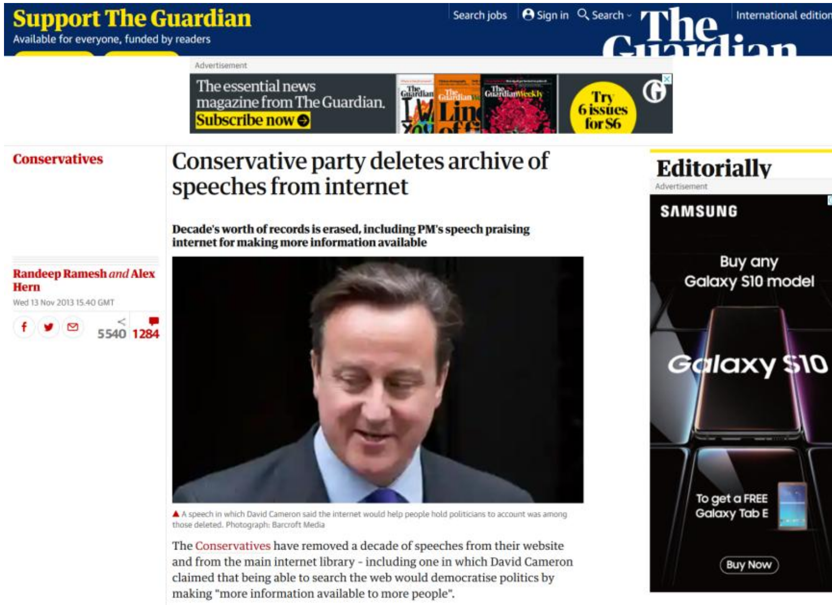
The World Wide Web by numbers

- “The first-ever website (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.worldwidewebsize.com/).

Life and death of Web pages

- “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



Conservatives

Conservative party deletes archive of speeches from internet

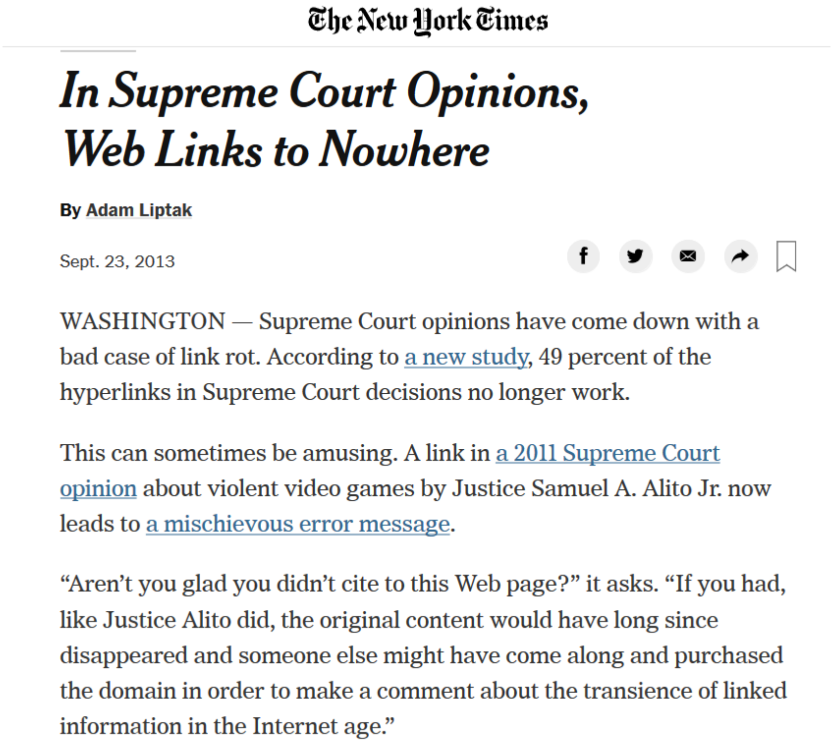
Decade's worth of records is erased, including PM's speech praising internet for making more information available

Randeep Ramesh *and* **Alex Hern**
Wed 13 Nov 2013 15.40 GMT

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".

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



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

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:

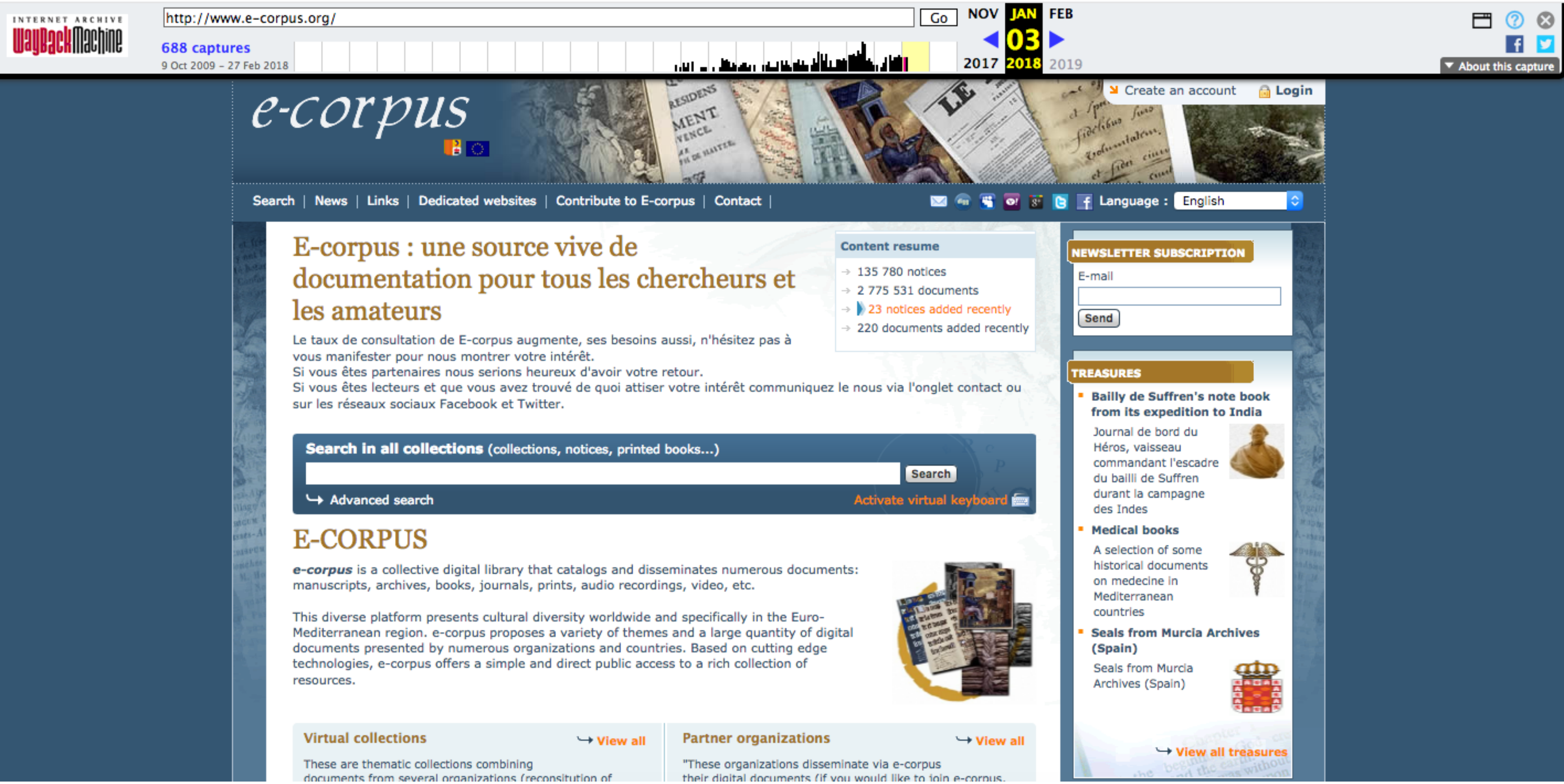
  On March 18, 2019, it was revealed that MySpace lost all of their user content from 2016 and earlier in "a server migration gone wrong". It was widely reported that over 50 million songs and 12 years worth of content was permanently lost, and there was no backup (https://en.wikipedia.org/wiki/Myspace)
- History :
  - Yougoslavia (.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?

My own experience in digital loss

E-corpus 2009-2016



INTERNET ARCHIVE WaybackMachine

http://www.e-corpus.org/ Go

688 captures
9 Oct 2009 – 27 Feb 2018

NOV JAN FEB
03
2017 2018 2019

About this capture

e-corpus

Create an account Login

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...)

Search

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

Send

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 medecine in Mediterranean countries
- **Seals from Murcia Archives (Spain)**
  Seals from Murcia Archives (Spain)

View all treasures

https://web.archive.org/web/20180103235239/http://www.e-corpus.org/

Why Web Archiving?

- 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).

Nielsen, J. (2016). *Using web archives in research: an introduction*. Retrieved from http://www.worldcat.org/oclc/960018046 p. 7

A definition of Web archiving

- 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/)

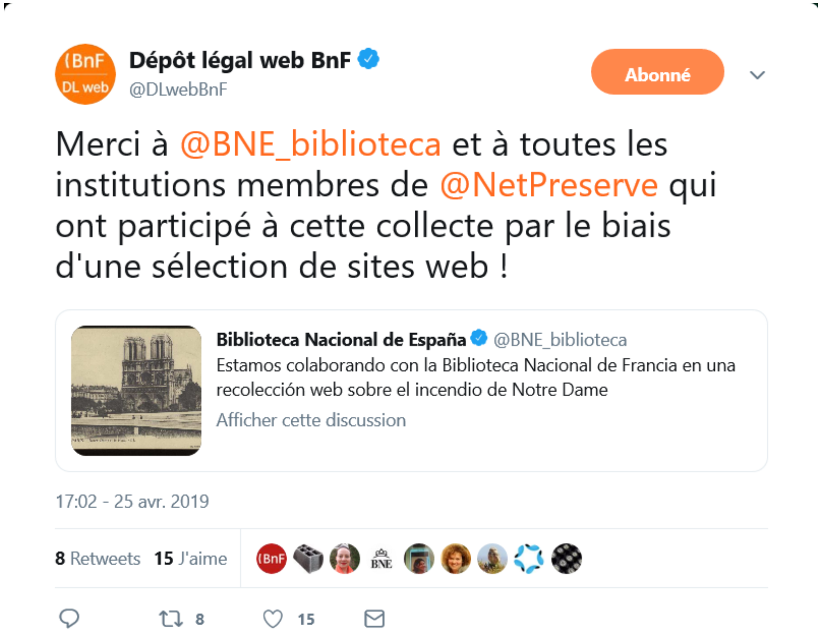
- “Web archiving is the process of gathering up data that has been recorded on the World Wide Web, storing it, ensuring the data is preserved in an archive, and making the collected data available for future research.” (Niu, J. (2012). An Overview of Web Archiving. *D-Lib Magazine*, *18* (3/4). https://doi.org/10.1045/march2012-niu1)

A short chronology of Web archives

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

Different strategies & methods for Web archiving

- 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)
- …



Dépôt légal web BnF @DLwebBnF

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

Different ways of accessing Web archives

- 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

- 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)

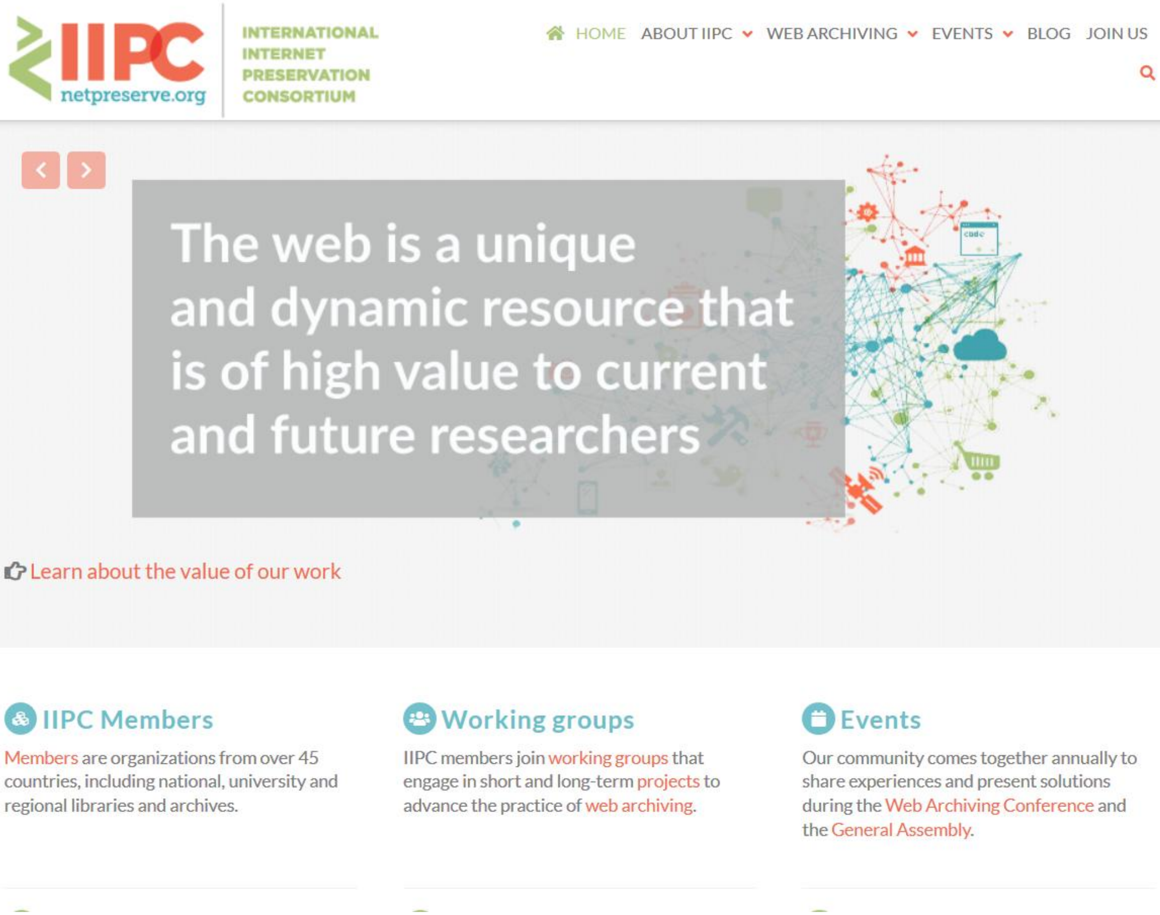
Web archiving projects often needs to gather diverse and multiples expertises and skills : archivists and librarians, researchers, legal officers, IT and computer specialists and… users and stakeholders

Legal and ethical issues

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/)

IIPC: International Internet Preservation Consortium



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

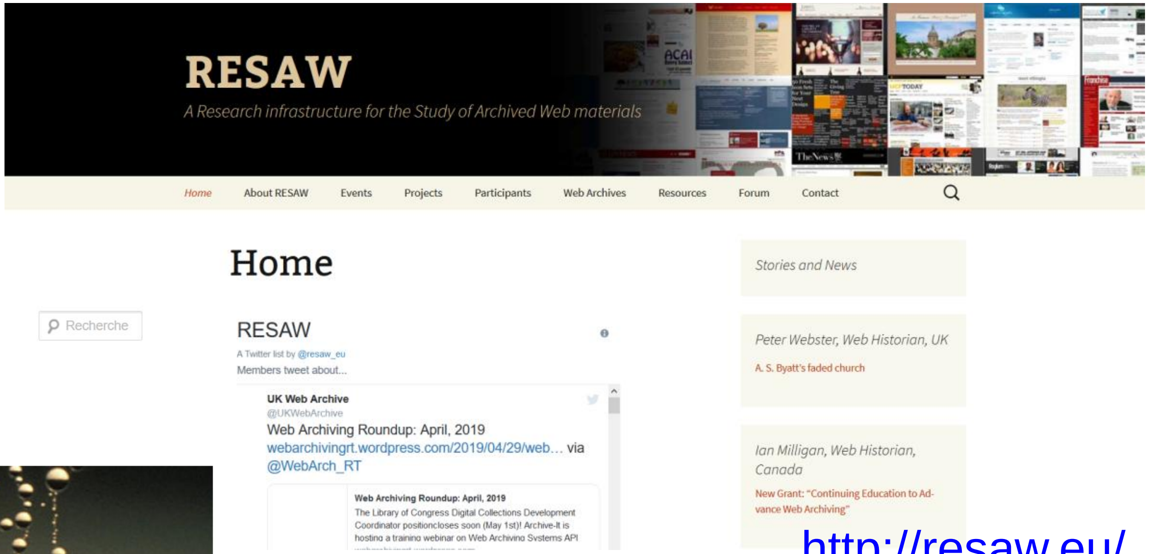
IIPC members join working groups that engage in short and long-term projects to advance the practice of web archiving.

Events

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

http://netpreserve.org/

New research fields and projects



RESAW

A Research infrastructure for the Study of Archived Web materials

Home | About RESAW | Events | Projects | Participants | Web Archives | Resources | Forum | Contact

Home

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… via @WebArch_RT

Web Archiving Roundup: April, 2019

The Library of Congress Digital Collections Development Coordinator positioncloses soon (May 1st)! Archive-It is hosting a training webinar on Web Archiving Systems API

Stories and News

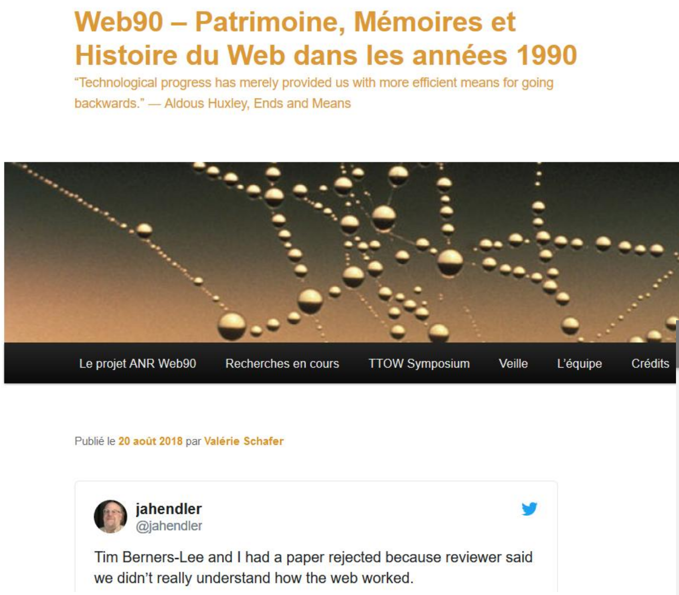
Peter Webster, Web Historian, UK

A. S. Byatt's faded church

Ian Milligan, Web Historian, Canada

New Grant: "Continuing Education to Advance Web Archiving"

http://resaw.eu/



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

Recherche

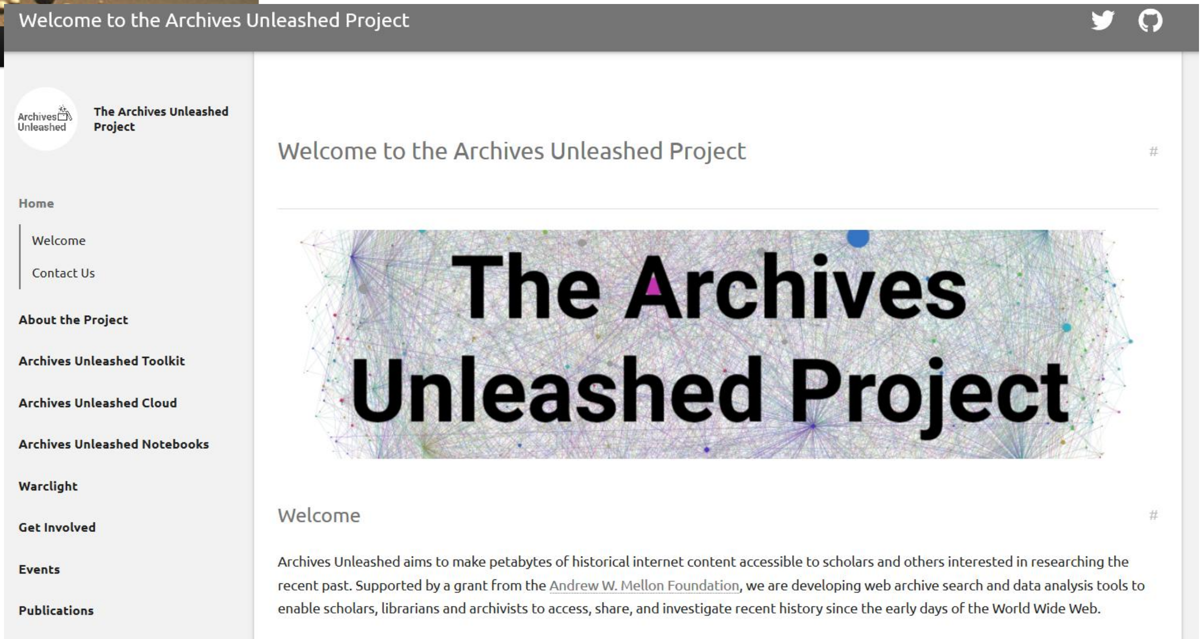
Le projet ANR Web90 | Recherches en cours | TTOW Symposium | Veille | L'équipe | Crédits

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/



Welcome to the Archives Unleashed Project

Archives Unleashed | The Archives Unleashed Project

- Home
  - Welcome
  - Contact Us
- About the Project
- Archives Unleashed Toolkit
- Archives Unleashed Cloud
- Archives Unleashed Notebooks
- Warclight
- Get Involved
- Events
- Publications

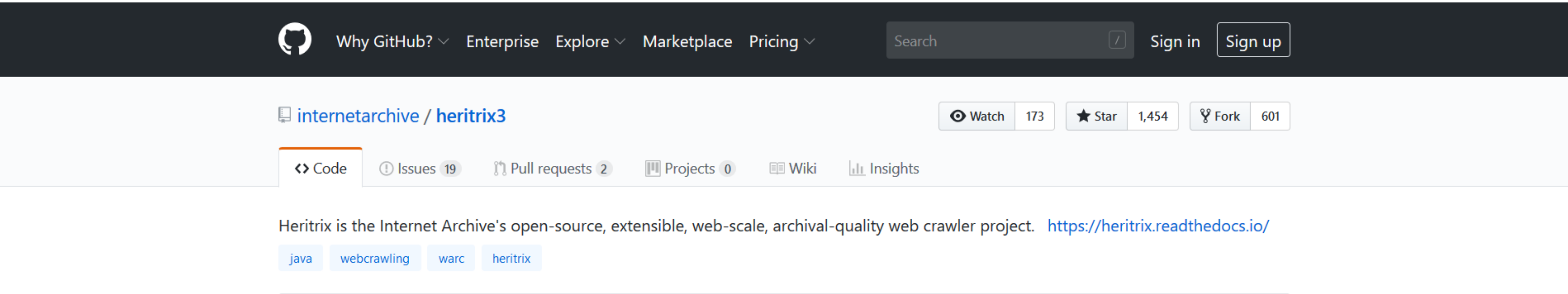
Welcome to the Archives Unleashed Project

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/

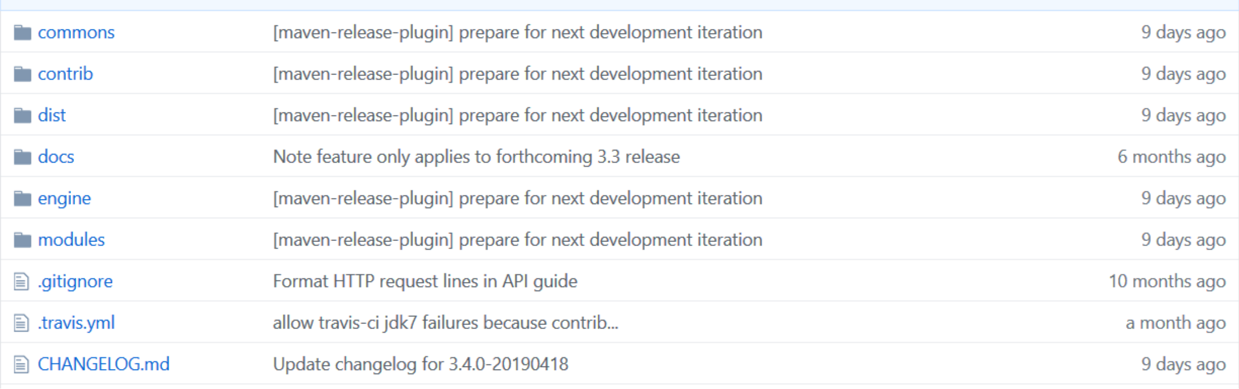
Technologies for Web archiving: Heritrix and many other tools



internetarchive / heritrix3

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

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

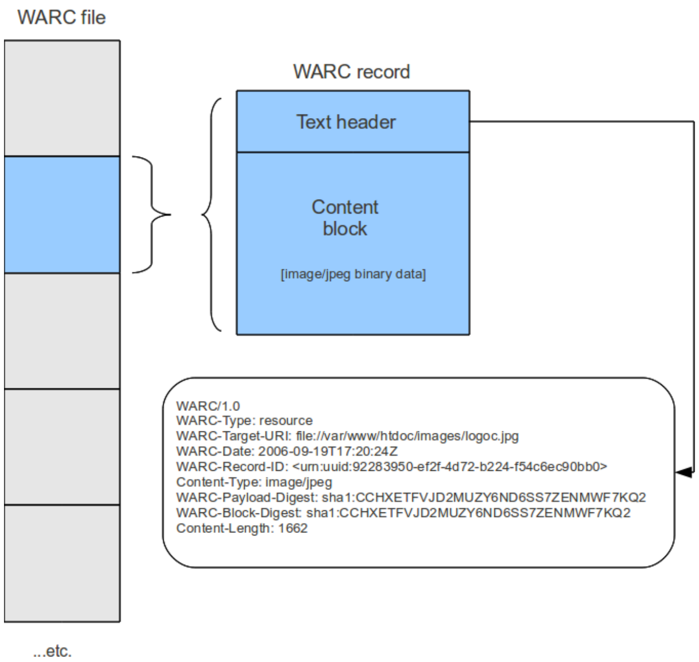


commons	[maven-release-plugin] prepare for next development iteration	9 days ago
contrib	[maven-release-plugin] prepare for next development iteration	9 days ago
dist	[maven-release-plugin] prepare for next development iteration	9 days ago
docs	Note feature only applies to forthcoming 3.3 release	6 months 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

A standard format for Web archives

- 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



```
WARC/1.0
WARC-Type: resource
WARC-Target-URI: file://var/www/htdoc/images/logoc.jpg
WARC-Date: 2006-09-19T17:20:24Z
WARC-Record-ID: <urn:uuid:92283950-ef2f-4d72-b224-f54c6ec90bb0>
Content-Type: image/jpeg
WARC-Payload-Digest: sha1:CCHXETFVJD2MUZY6ND6SS7ZENMWF7KQ2
WARC-Block-Digest: sha1:CCHXETFVJD2MUZY6ND6SS7ZENMWF7KQ2
Content-Length: 1662
```

https://en.wikipedia.org/wiki/Web_ARChive
https://wiki.archivematica.org/Significant_characteristics_of_websites
https://wiki.archivematica.org/File:WARCdiagram.png

Challenges and setbacks in Web archiving

- 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
- …

Method and tools for Web archiving

- 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

URL : Uniform Resource Locator

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

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

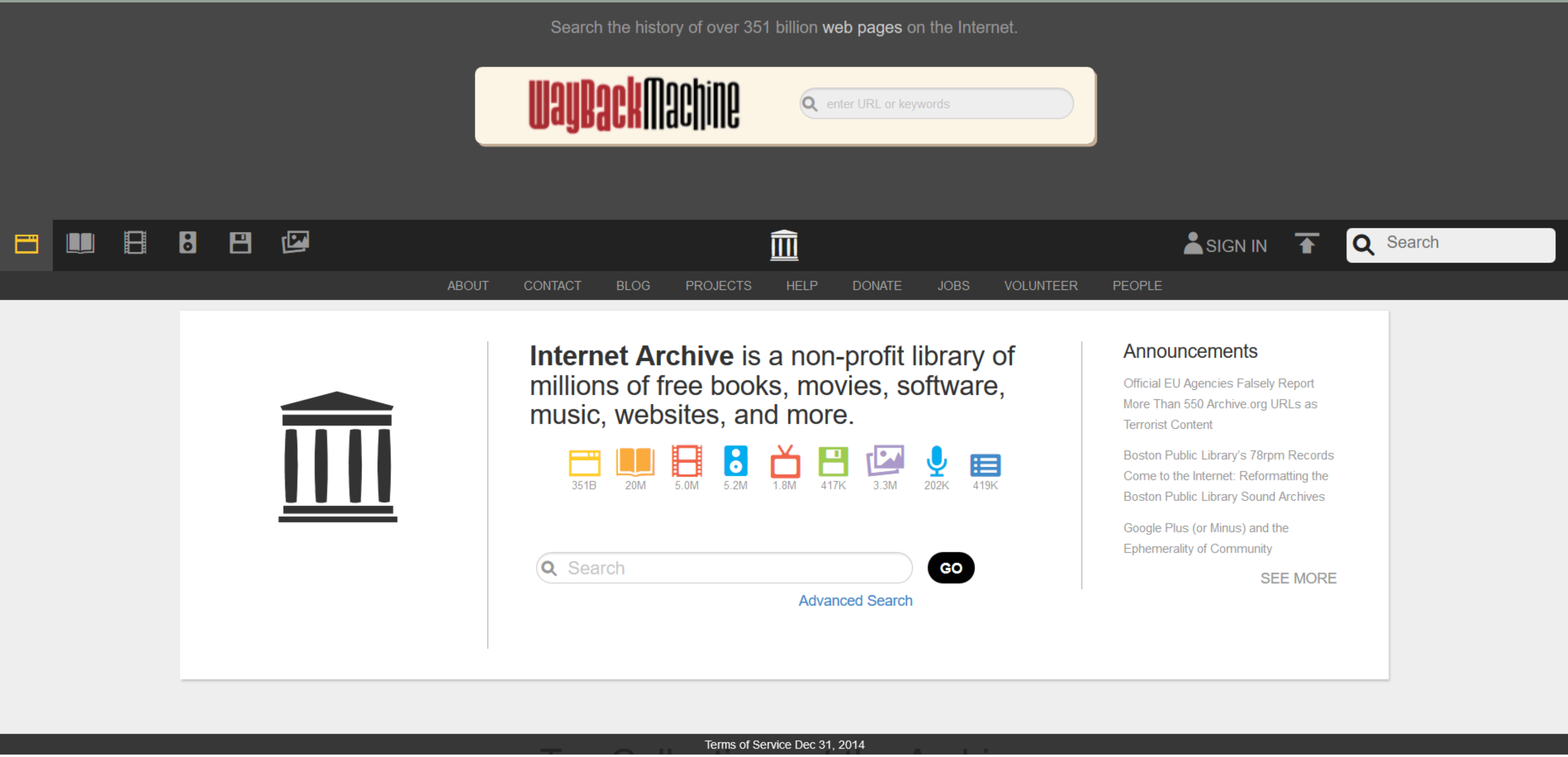
Ex:

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

https://en.wikipedia.org/wiki/URL

http://dac.au.dk/forskning/forskningsprogrammer/ p. 51

Internet Archive’s Wayback Machine



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

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

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

- Launched in 2001
- 357 billion archived Web pages so far
- Archived content going back to 1996

https://archive.org/

URLs in the Wayback Machine

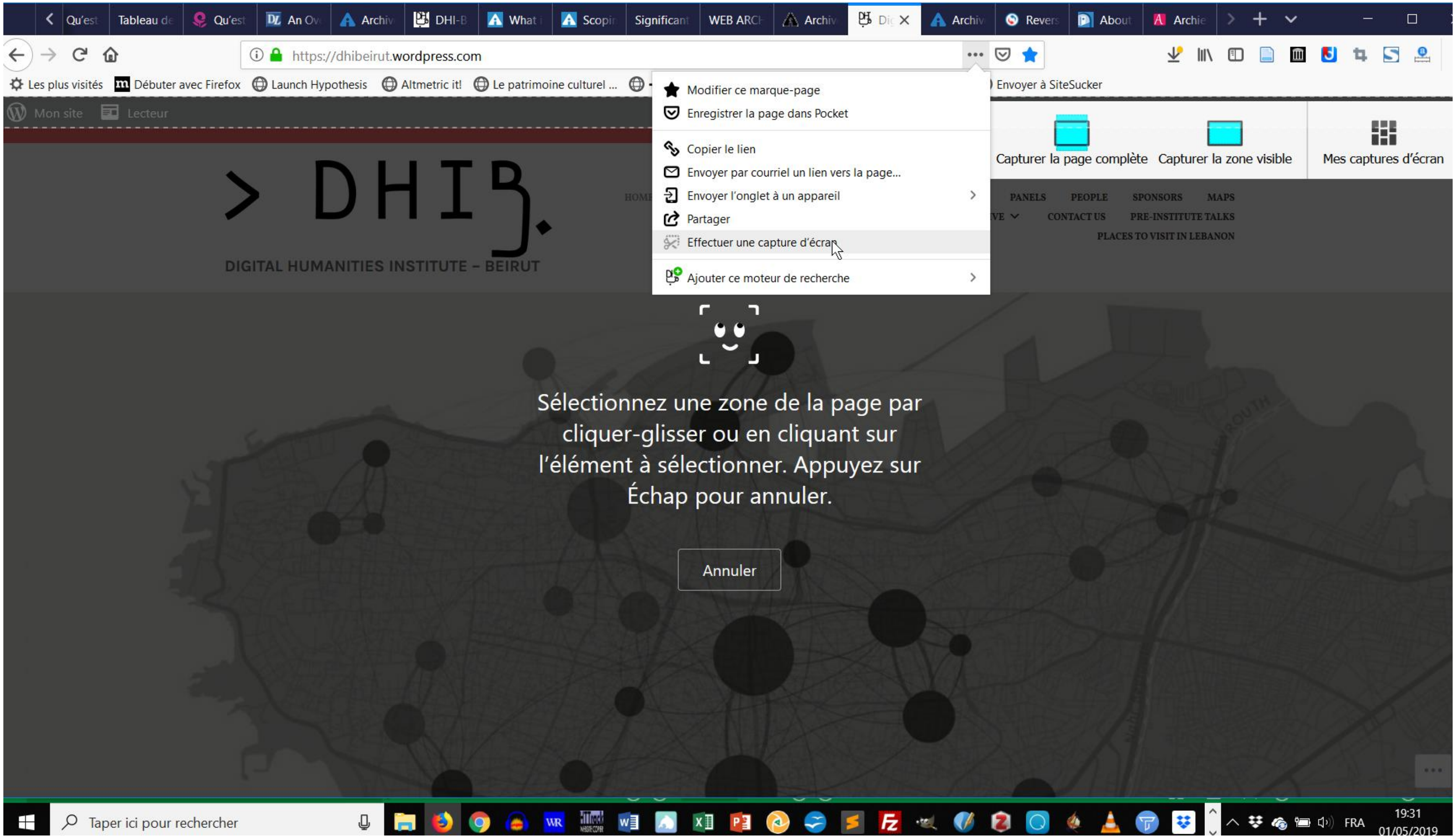
Structure of URLs in the Internet Archive’s Wayback Machine:

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

Ex:

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

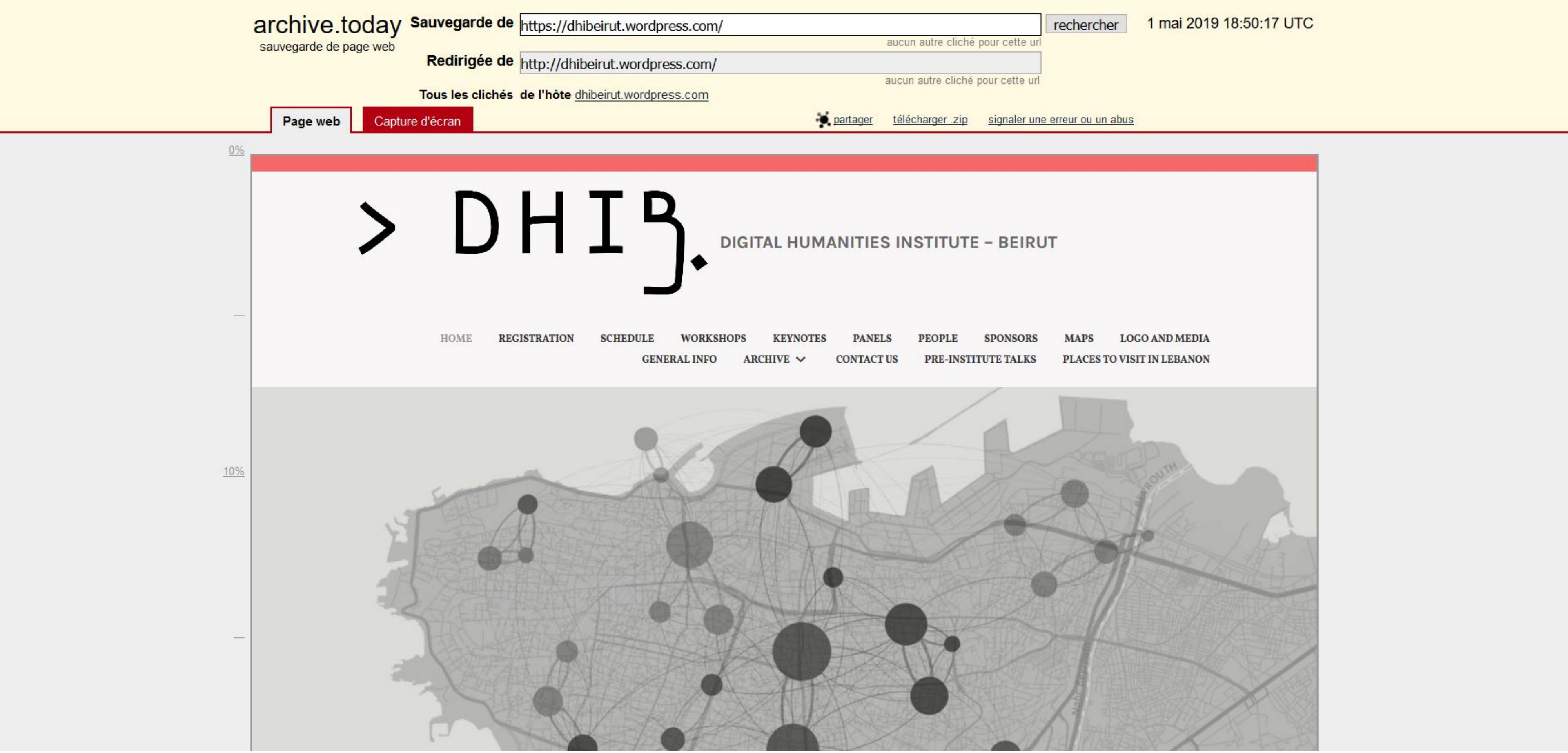
Screen capture with Firefox



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

https://dhibeirut.wordpress.com
Les plus visités
Débuter avec Firefox
Launch Hypothesis
Altmetric it!
Le patrimoine culturel ...
Envoyer à SiteSucker
Mon site
Lecteur
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
Capturer la page complète
Capturer la zone visible
Mes captures d'écran
> DHIB.
DIGITAL HUMANITIES INSTITUTE – BEIRUT
PANELS
PEOPLE
SPONSORS
MAPS
CONTACT US
PRE-INSTITUTE TALKS
PLACES TO VISIT IN LEBANON
Annuler
Taper ici pour rechercher
FRA
19:31
01/05/2019

Snapshot with Archive.fo



archive.today
sauvegarde de page web

Sauvegarde de https://dhibeirut.wordpress.com/
aucun autre cliché pour cette url

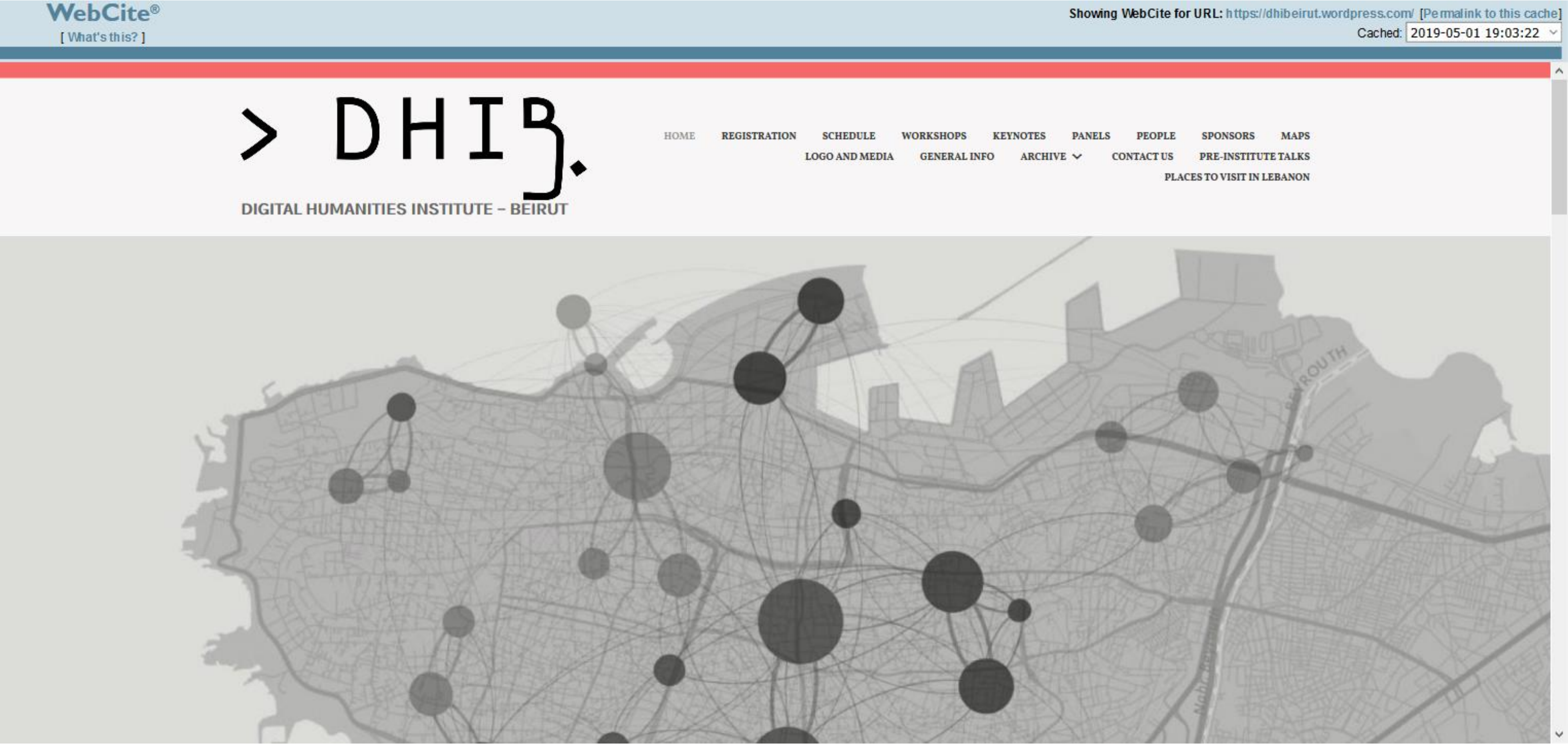
Redirigée de http://dhibeirut.wordpress.com/
aucun autre cliché pour cette url

Tous les clichés de l'hôte dhibeirut.wordpress.com

1 mai 2019 18:50:17 UTC

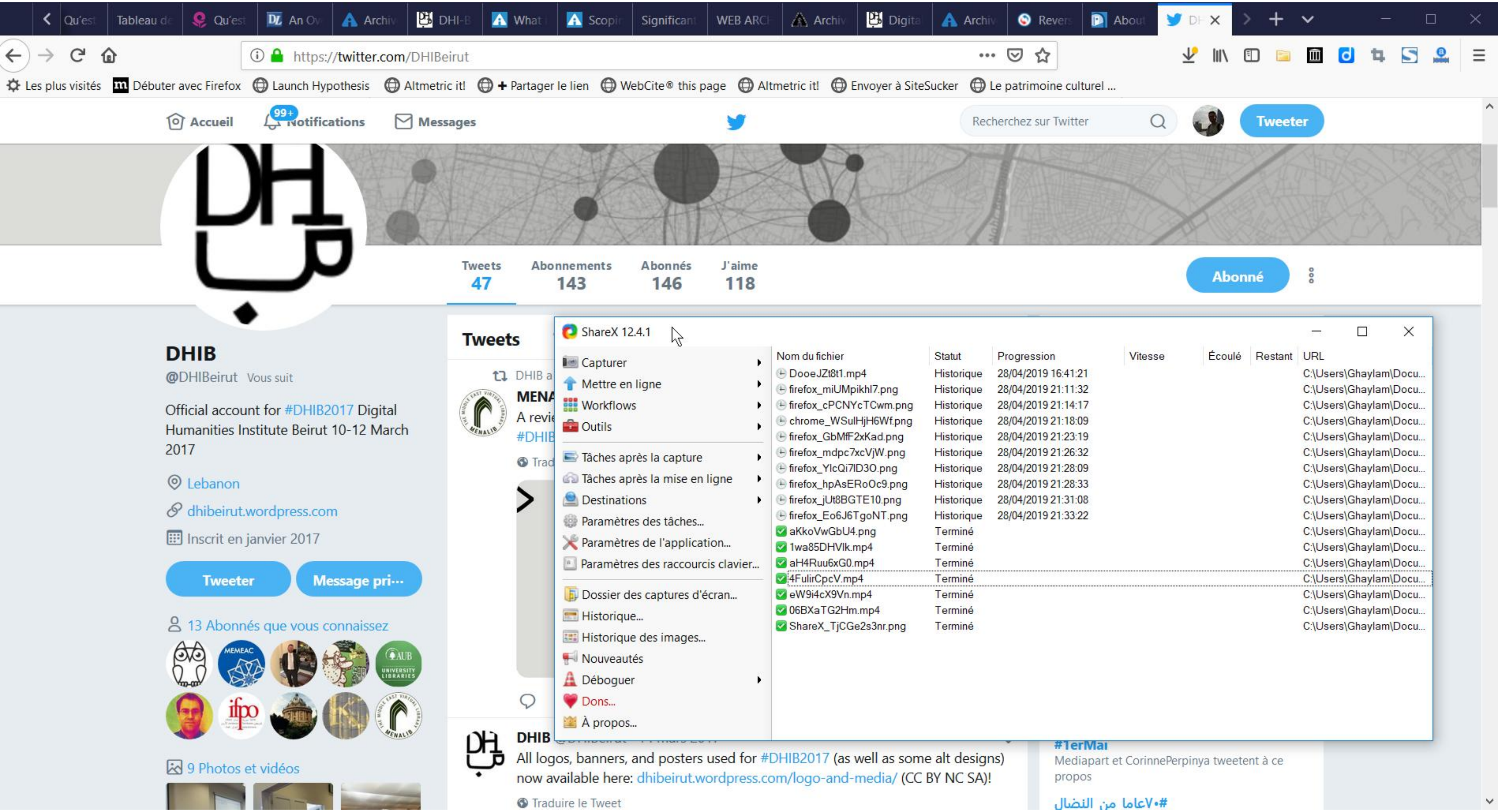
http://archive.fo/ZAFsE

Citation and cached version with Webcite


WebCite®
[What's this?]
Showing WebCite for URL: https://dhibeirut.wordpress.com/ [Permalink to this cache]
Cached: 2019-05-01 19:03:22
> 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
BEYROUTH

http://www.webcitation.org/783EYmwuV

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



Nom du fichier	Statut	Progression	Vitesse	Écoulé	Restant	URL
DooeJZt8t1.mp4	Historique	28/04/2019 16:41:21				C:\Users\Ghaylam\Docu...
firefox_miUMpikhl7.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_GbMfF2xKad.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_YlcQi7lD3O.png	Historique	28/04/2019 21:28:09				C:\Users\Ghaylam\Docu...
firefox_hpAsERoOc9.png	Historique	28/04/2019 21:28:33				C:\Users\Ghaylam\Docu...
firefox_jUt8BGTE10.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...

DHIB

@DHIBeirut Vous suit

Official account for #DHIB2017 Digital Humanities Institute Beirut 10-12 March 2017

Lebanon

dhibeirut.wordpress.com

Inscrit en janvier 2017

DHIB

All logos, banners, and posters used for #DHIB2017 (as well as some alt designs) now available here: dhibeirut.wordpress.com/logo-and-media/ (CC BY NC SA)!

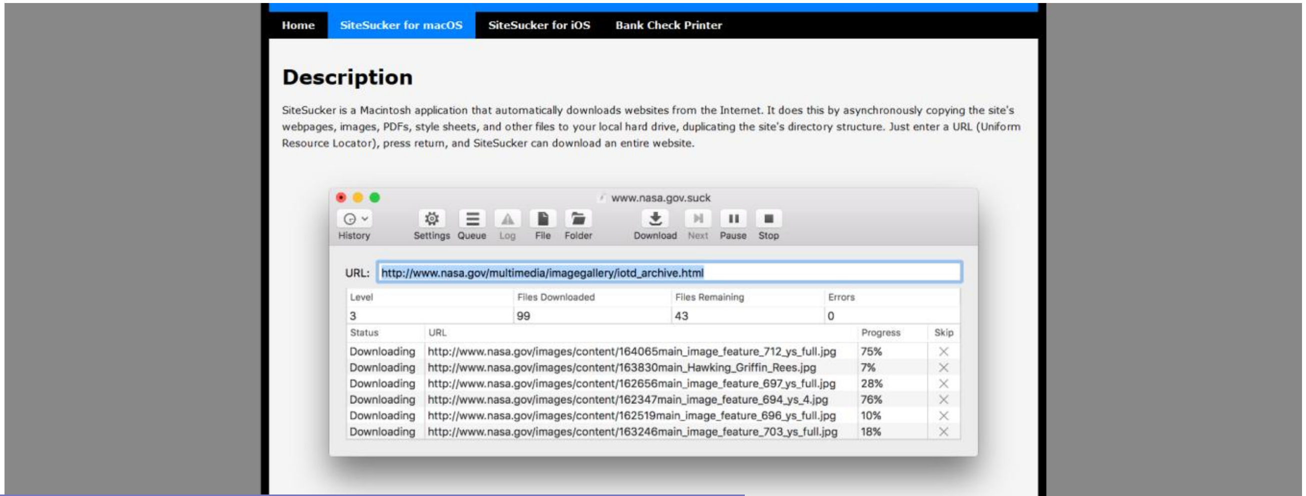
#1erMai

Mediapart et CorinnePerpinya tweetent à ce propos

#٧٠عاما من النضال

https://github.com/ShareX/ShareX

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



Description

SiteSucker is a Macintosh application that automatically downloads websites from the Internet. It does this by asynchronously copying the site's webpages, images, PDFs, style sheets, and other files to your local hard drive, duplicating the site's directory structure. Just enter a URL (Uniform Resource Locator), press return, and SiteSucker can download an entire website.

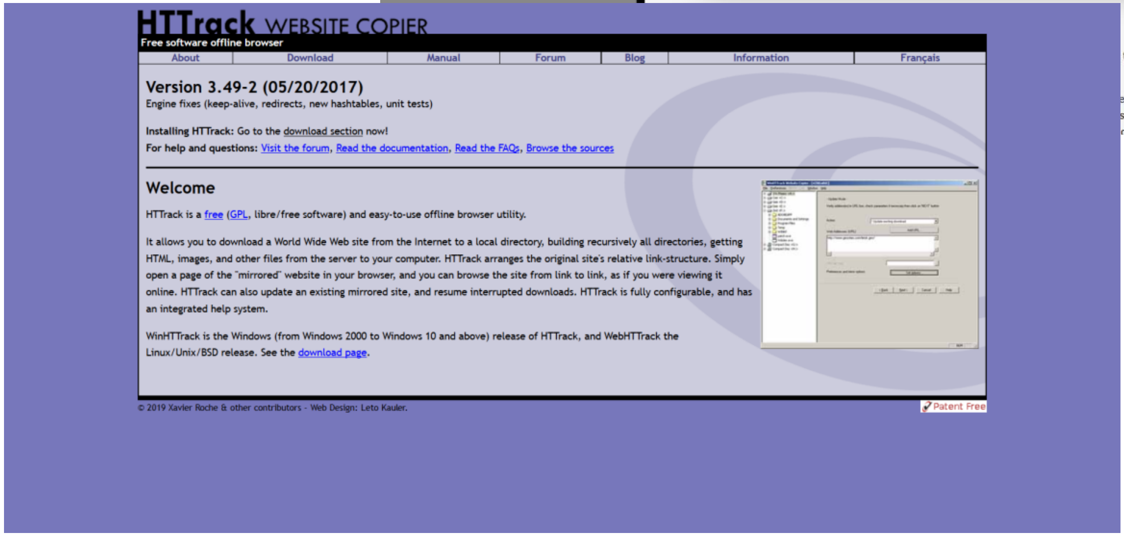
URL: http://www.nasa.gov/multimedia/imagegallery/iotd_archive.html

Level	Files Downloaded	Files Remaining	Errors
3	99	43	0

Status	URL	Progress	Skip
Downloading	http://www.nasa.gov/images/content/164065main_image_feature_712_ys_full.jpg	75%	×
Downloading	http://www.nasa.gov/images/content/163830main_Hawking_Griffin_Rees.jpg	7%	×
Downloading	http://www.nasa.gov/images/content/162656main_image_feature_697_ys_full.jpg	28%	×
Downloading	http://www.nasa.gov/images/content/162347main_image_feature_694_ys_4.jpg	76%	×
Downloading	http://www.nasa.gov/images/content/162519main_image_feature_696_ys_full.jpg	10%	×
Downloading	http://www.nasa.gov/images/content/163246main_image_feature_703_ys_full.jpg	18%	×

the files it downloads, allowing you to browse a site

e 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



Version 3.49-2 (05/20/2017)

Engine fixes (keep-alive, redirects, new hashtables, unit tests)

Installing HTTrack: Go to the download section now!

For help and questions: Visit the forum, Read the documentation, Read the FAQs, Browse the sources

Welcome

HTTrack is a free (GPL, libre/free software) and easy-to-use offline browser utility.

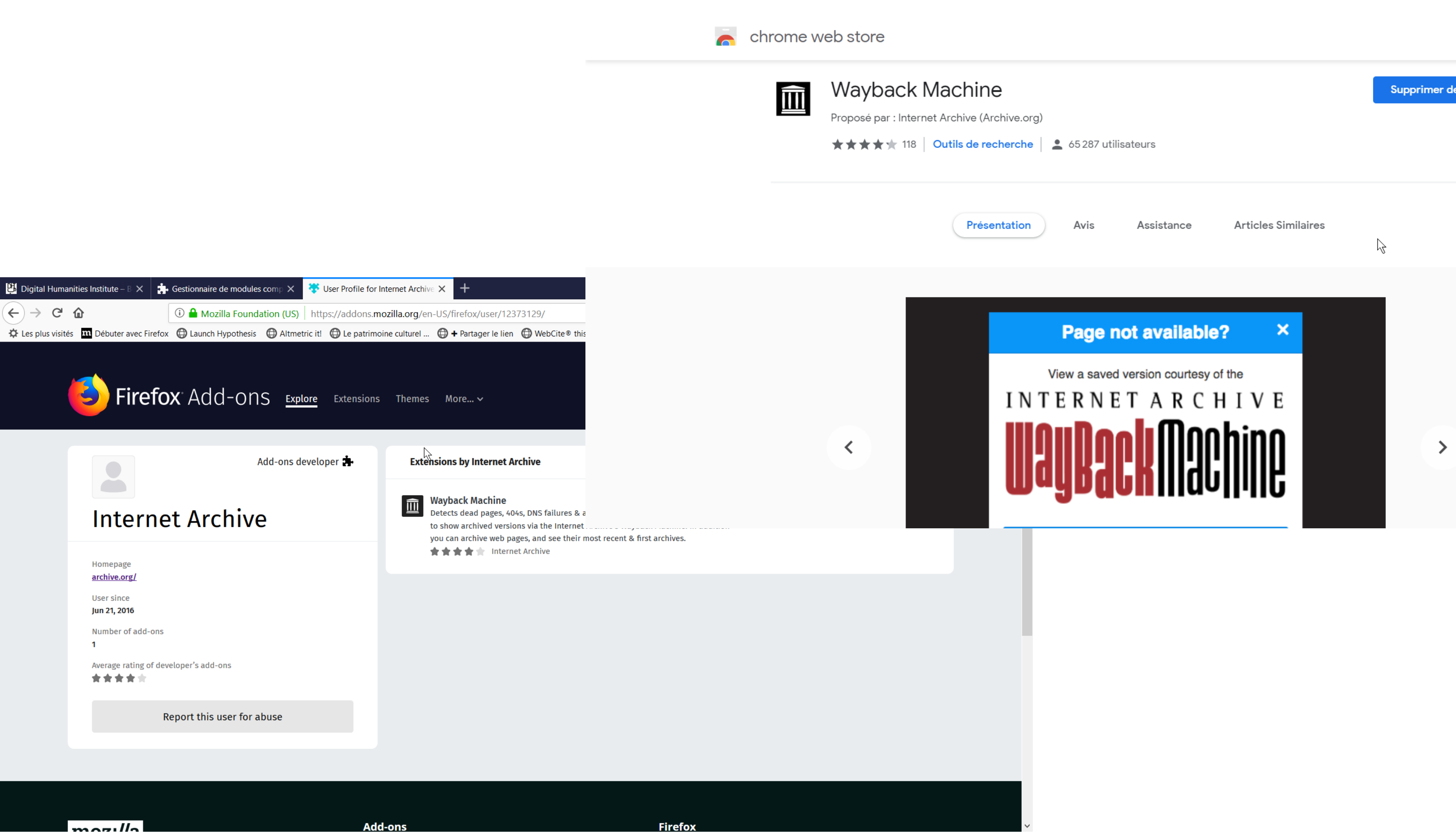
It allows you to download a World Wide Web site from the Internet to a local directory, building recursively all directories, getting HTML, images, and other files from the server to your computer. HTTrack arranges the original site's relative link-structure. Simply open a page of the "mirrored" website in your browser, and you can browse the site from link to link, as if you were viewing it online. HTTrack can also update an existing mirrored site, and resume interrupted downloads. HTTrack is fully configurable, and has an integrated help system.

WinHTTrack is the Windows (from Windows 2000 to Windows 10 and above) release of HTTrack, and WebHTTrack the Linux/Unix/BSD release. See the download page.

© 2019 Xavier Roche & other contributors - Web Design: Leto Kauler.

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

Internet Archive / Wayback Machine extension for browsers



Wayback Machine

Proposé par : Internet Archive (Archive.org)

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

Page not available?

View a saved version courtesy of the

INTERNET ARCHIVE

WayBackMachine

Mozilla Foundation (US) https://addons.mozilla.org/en-US/firefox/user/12373129/

Add-ons developer

Internet Archive

Homepage
archive.org/

User since
Jun 21, 2016

Number of add-ons
1

Average rating of developer's add-ons
★★★★☆

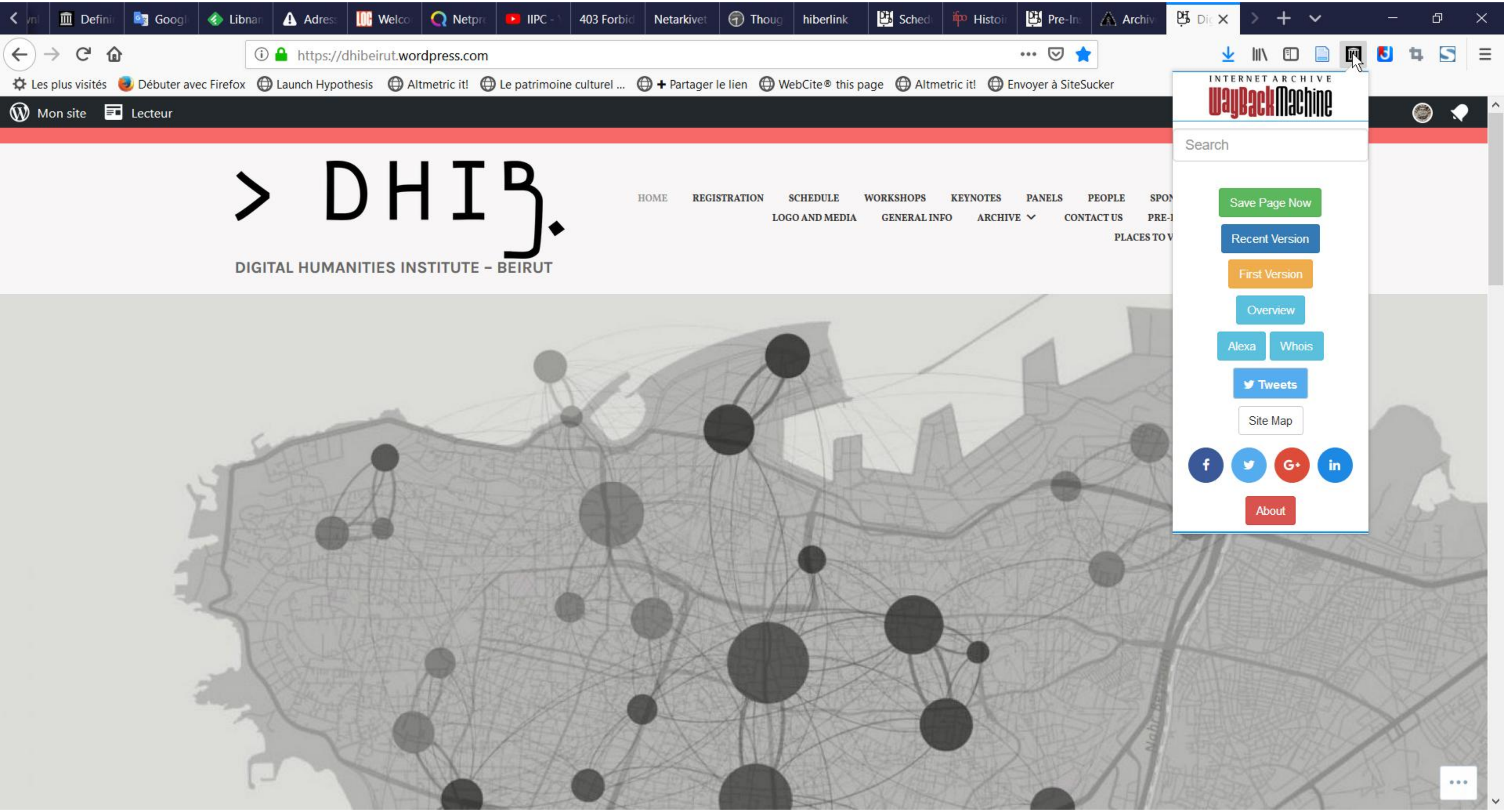
Extensions by Internet Archive

Wayback Machine
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.
★★★★☆ Internet Archive

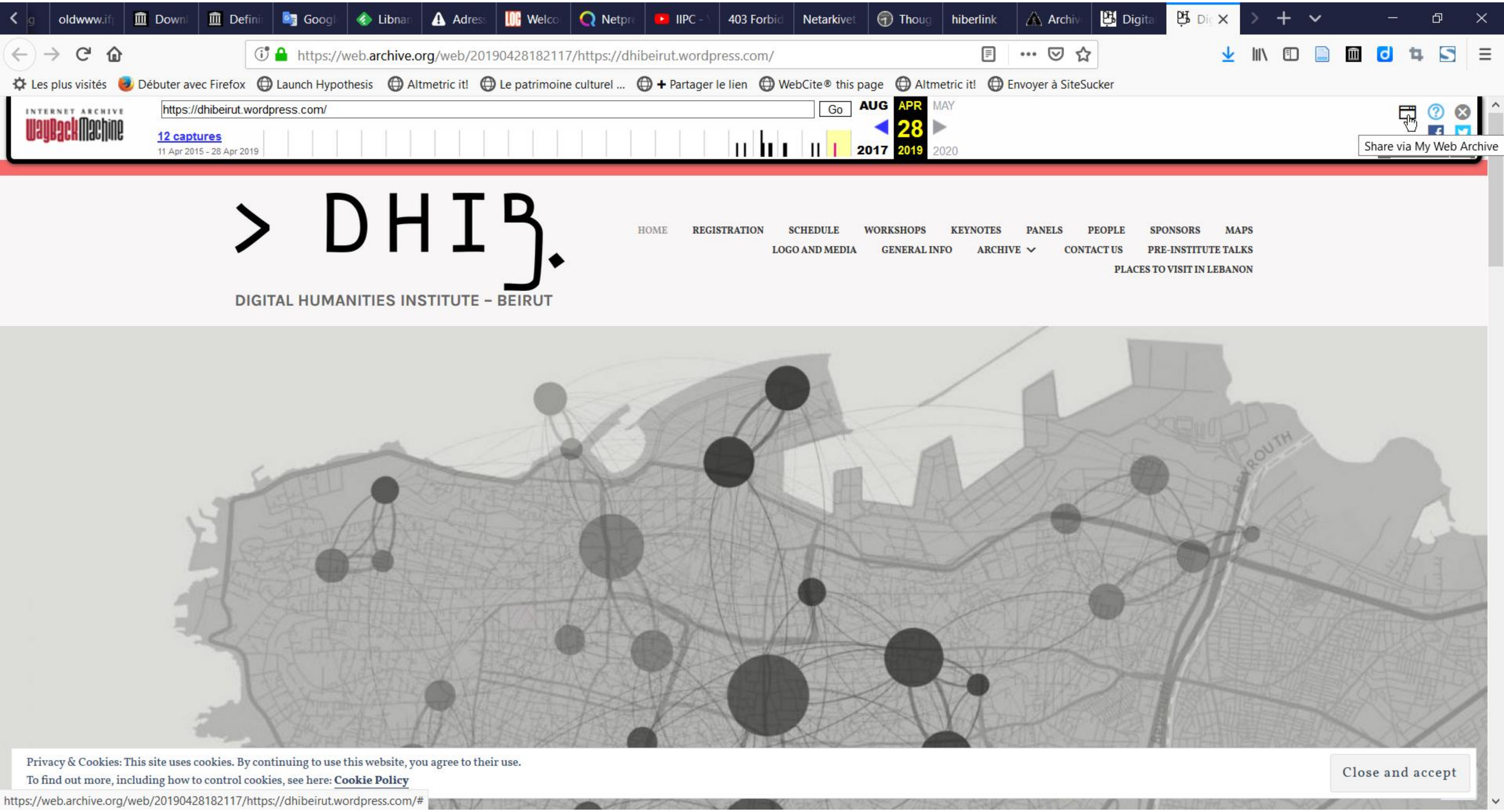
https://chrome.google.com/webstore/detail/wayback-machine/fpnmgdkabkmnadcjpehmlllkndpkmiak?hl=fr

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

Internet Archive / Wayback Machine extension for browsers



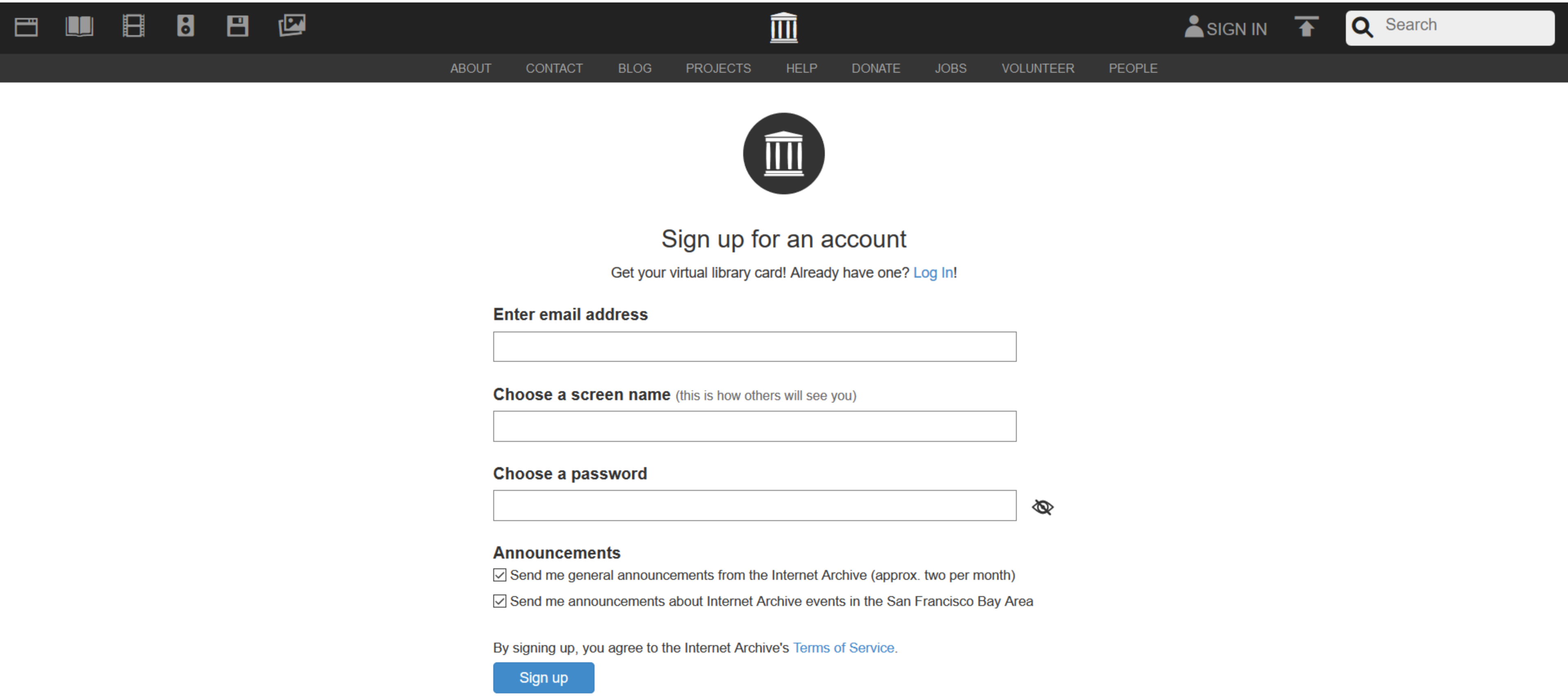
Internet Archive / Wayback Machine extension for browsers



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

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

Your own collection of archived Web pages in Internet Archive



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

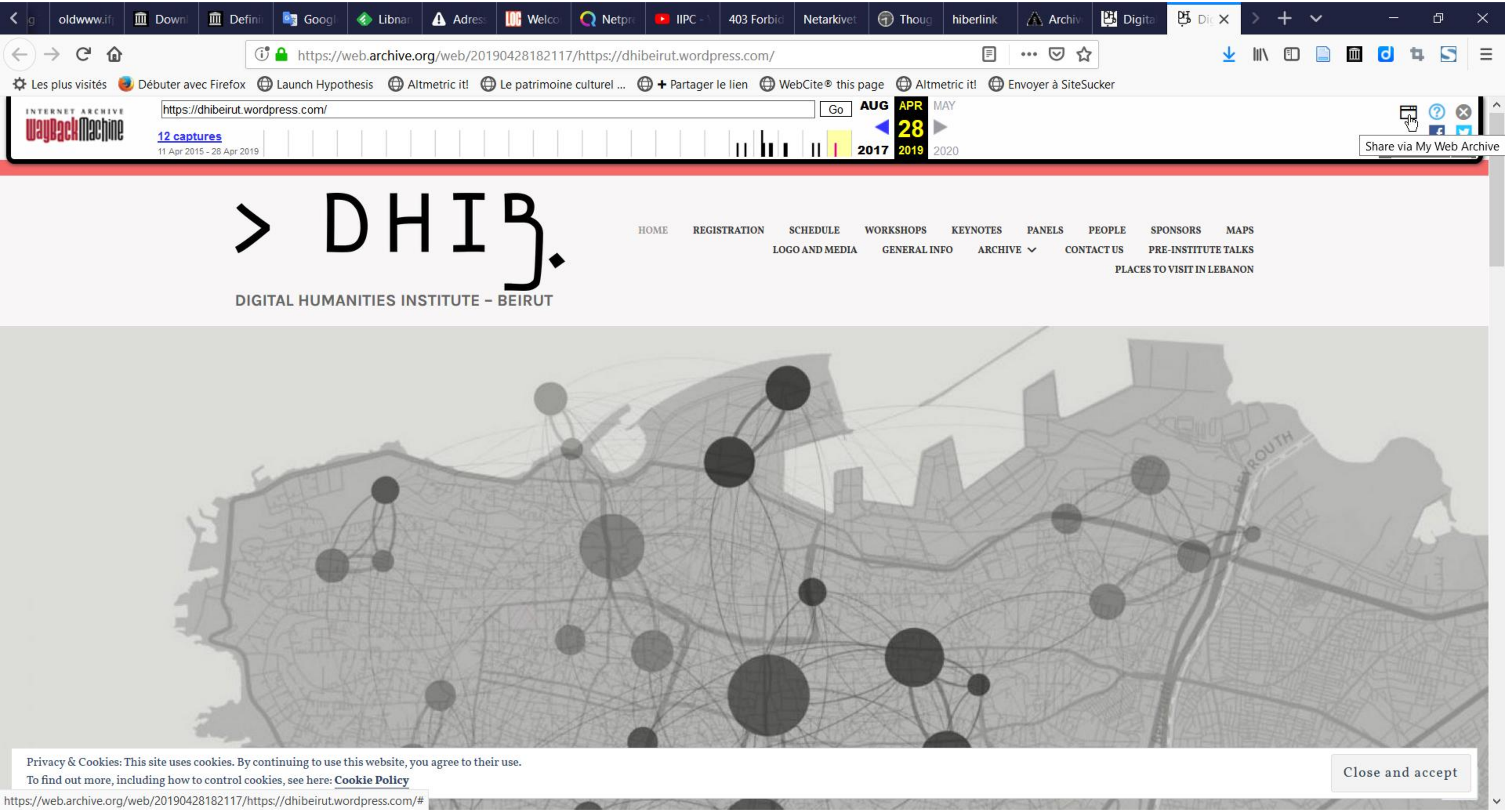
☑ Send me general announcements from the Internet Archive (approx. two per month)

☑ 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.

https://archive.org/account/signup

Your own collection of archived Web pages in Internet Archive



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

https://dhibeirut.wordpress.com/

12 captures

11 Apr 2015 - 28 Apr 2019

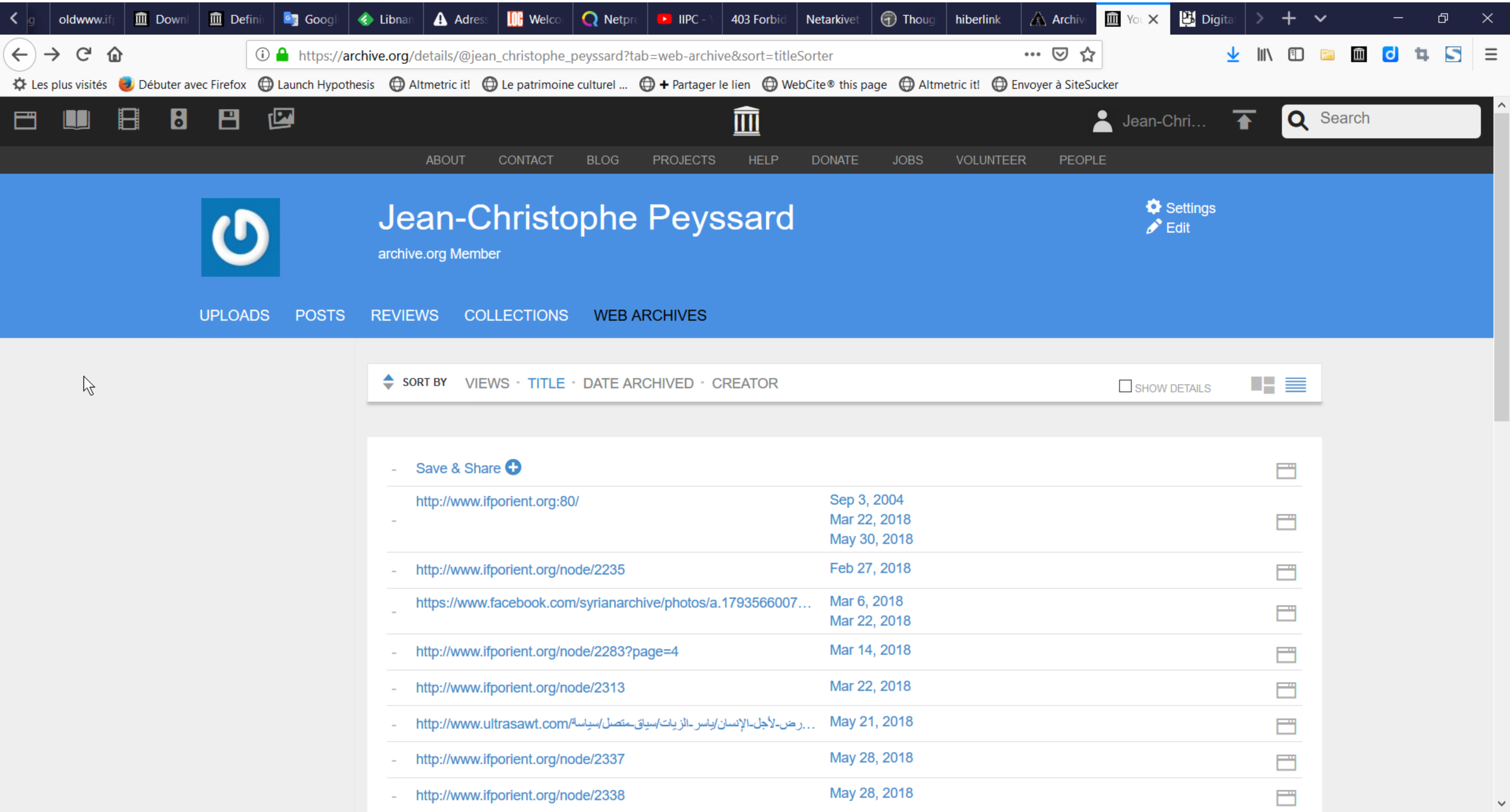
Share via My Web Archive

DIGITAL HUMANITIES INSTITUTE – BEIRUT

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

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

Your own collection of archived Web pages in Internet Archive



Jean-Christophe Peyssard

archive.org Member

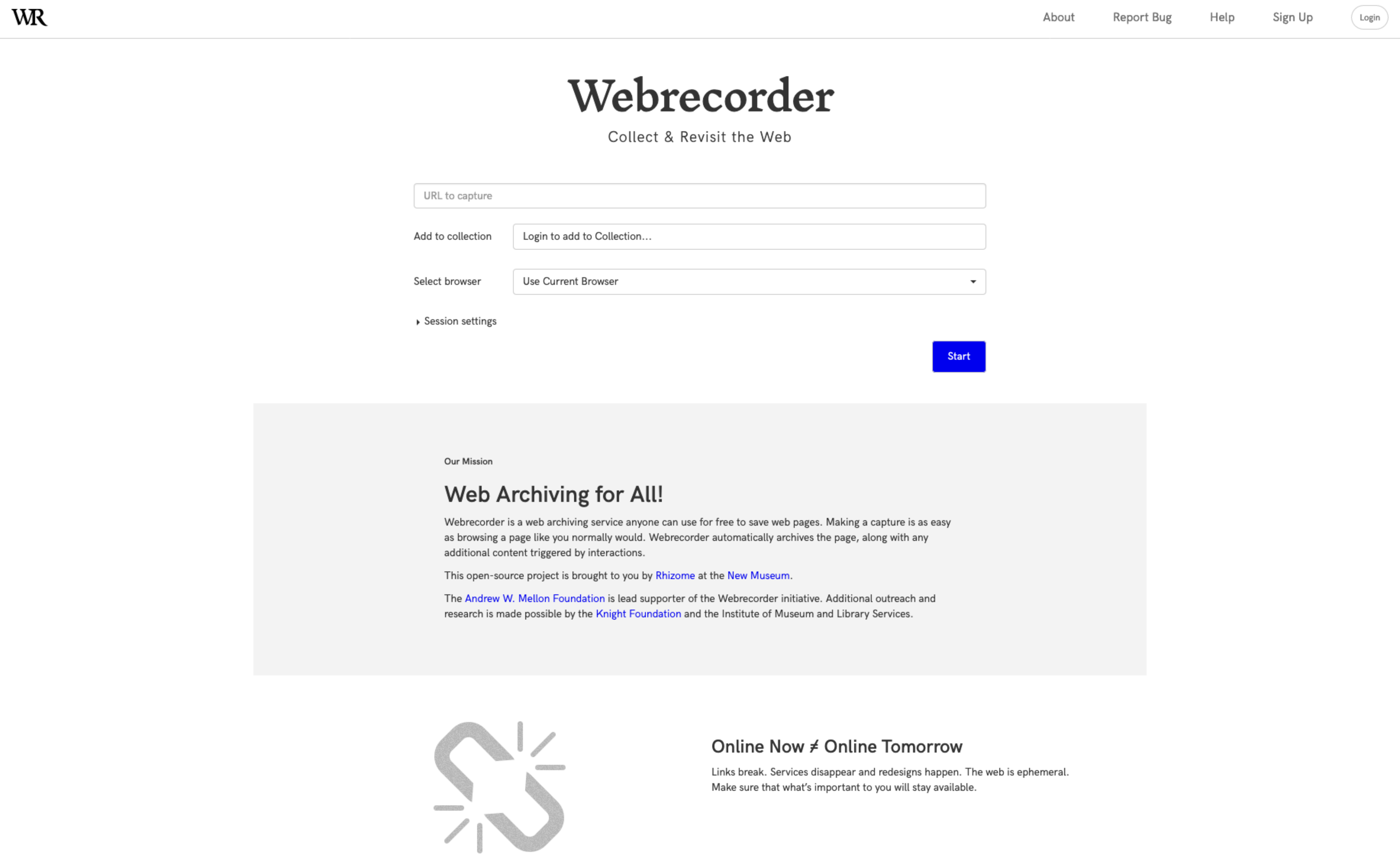
UPLOADS POSTS REVIEWS COLLECTIONS WEB ARCHIVES

SORT BY VIEWS · TITLE · DATE ARCHIVED · CREATOR

□ SHOW DETAILS

-	Save & Share	
-	http://www.ifporient.org:80/	Sep 3, 2004 Mar 22, 2018 May 30, 2018
-	http://www.ifporient.org/node/2235	Feb 27, 2018
-	https://www.facebook.com/syrianarchive/photos/a.1793566007…	Mar 6, 2018 Mar 22, 2018
-	http://www.ifporient.org/node/2283?page=4	Mar 14, 2018
-	http://www.ifporient.org/node/2313	Mar 22, 2018
-	http://www.ultrasawt.com/سياسة/سياق-متصل/ياسر-الزيات/الإنسان-لأجل-رض…	May 21, 2018
-	http://www.ifporient.org/node/2337	May 28, 2018
-	http://www.ifporient.org/node/2338	May 28, 2018

Archiving complete Web sites with Webrecorder



Webrecorder

Collect & Revisit the Web

URL to capture

Add to collection: Login to add to Collection…

Select browser: Use Current Browser

▸ Session settings

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.

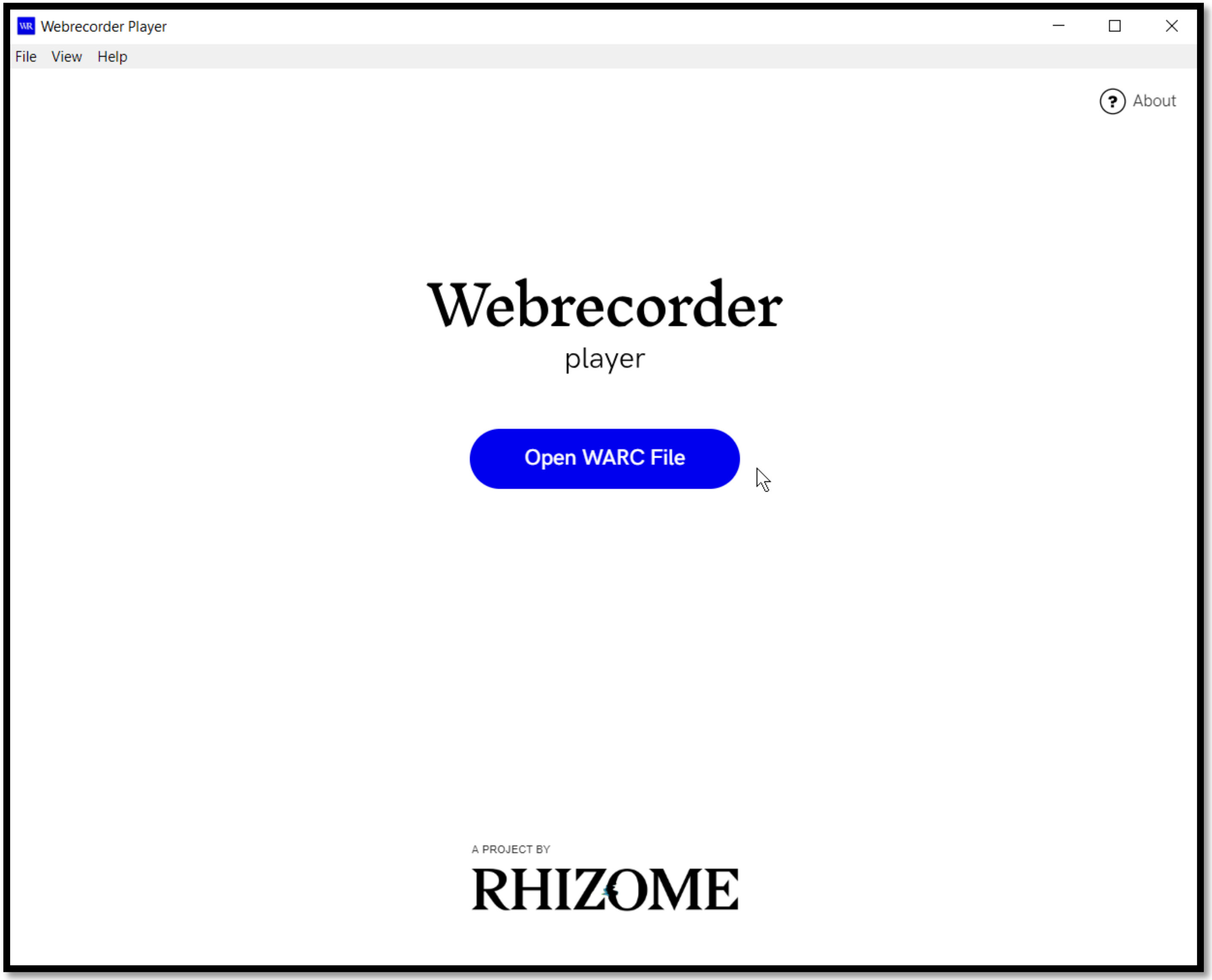
The Andrew W. Mellon Foundation is lead supporter of the Webrecorder initiative. Additional outreach and research is made possible by the Knight Foundation and the Institute of Museum and Library Services.

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.

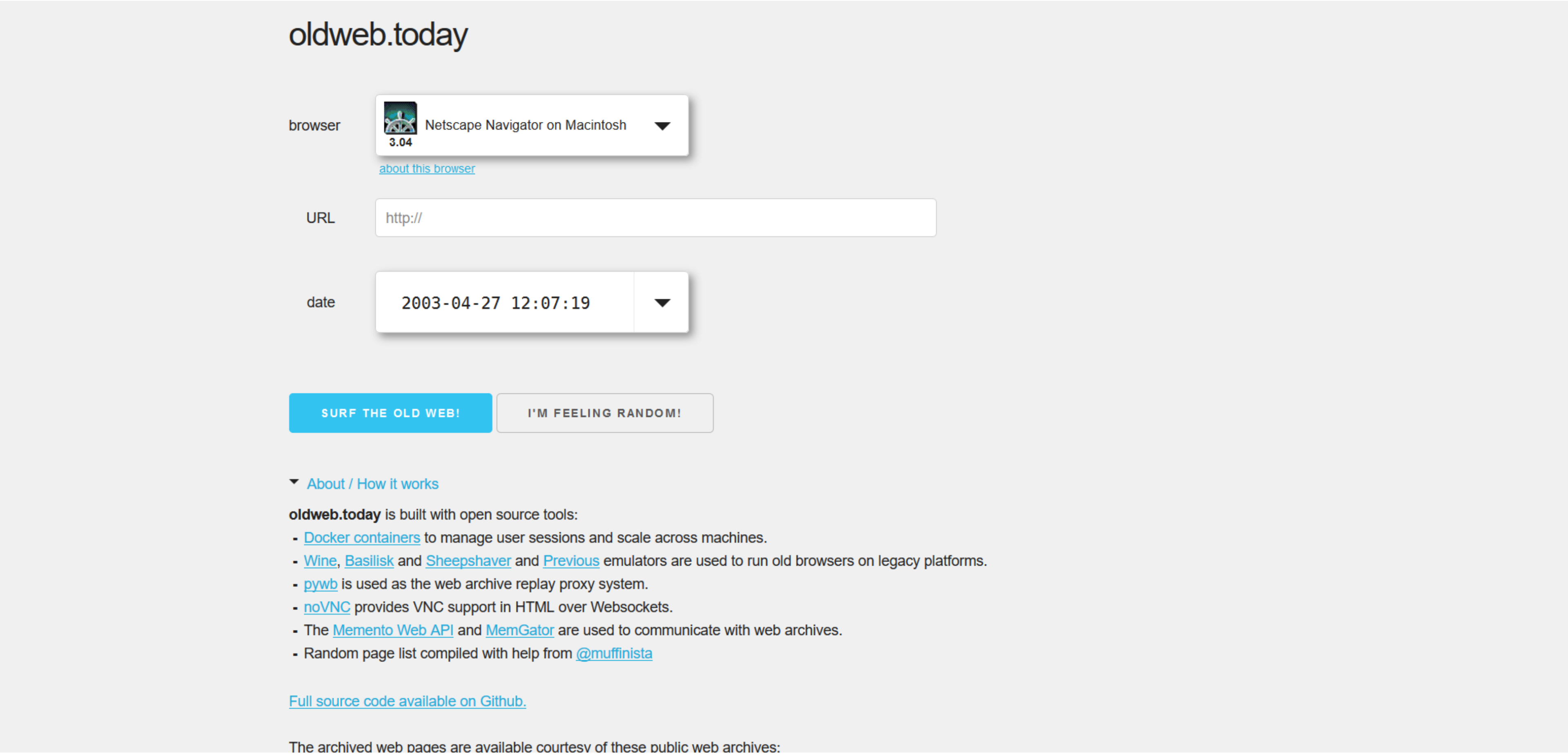
https://webrecorder.io/

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

date: 2003-04-27 12:07:19

oldweb.today is built with open source tools:

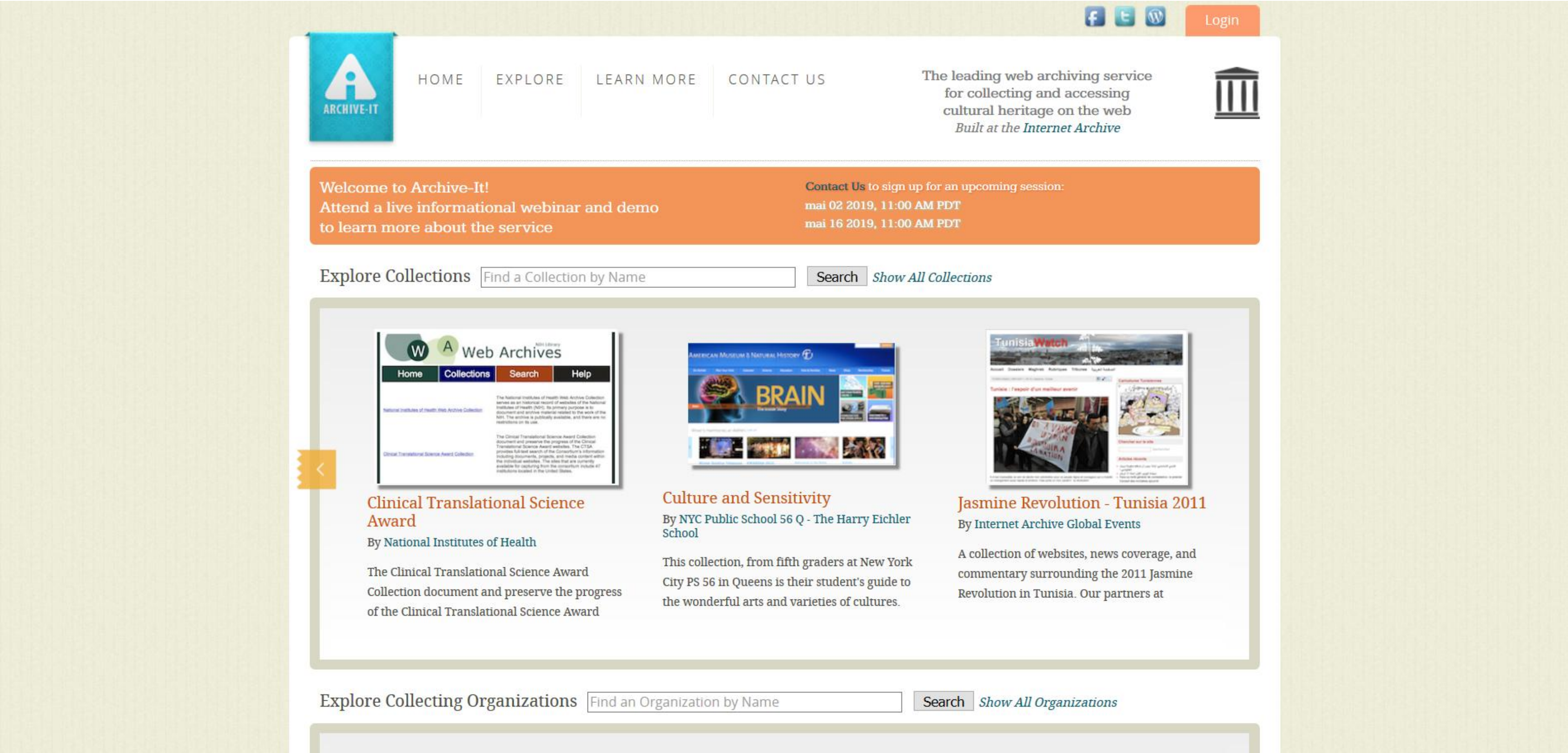
- Docker containers to manage user sessions and scale across machines.
- Wine, Basilisk and Sheepshaver and Previous emulators are used to run old browsers on legacy platforms.
- pywb is used as the web archive replay proxy system.
- noVNC provides VNC support in HTML over Websockets.
- The Memento Web 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



Login

ARCHIVE-IT

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

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

Explore Collections [Find a Collection by Name] Search *Show All Collections*

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

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.

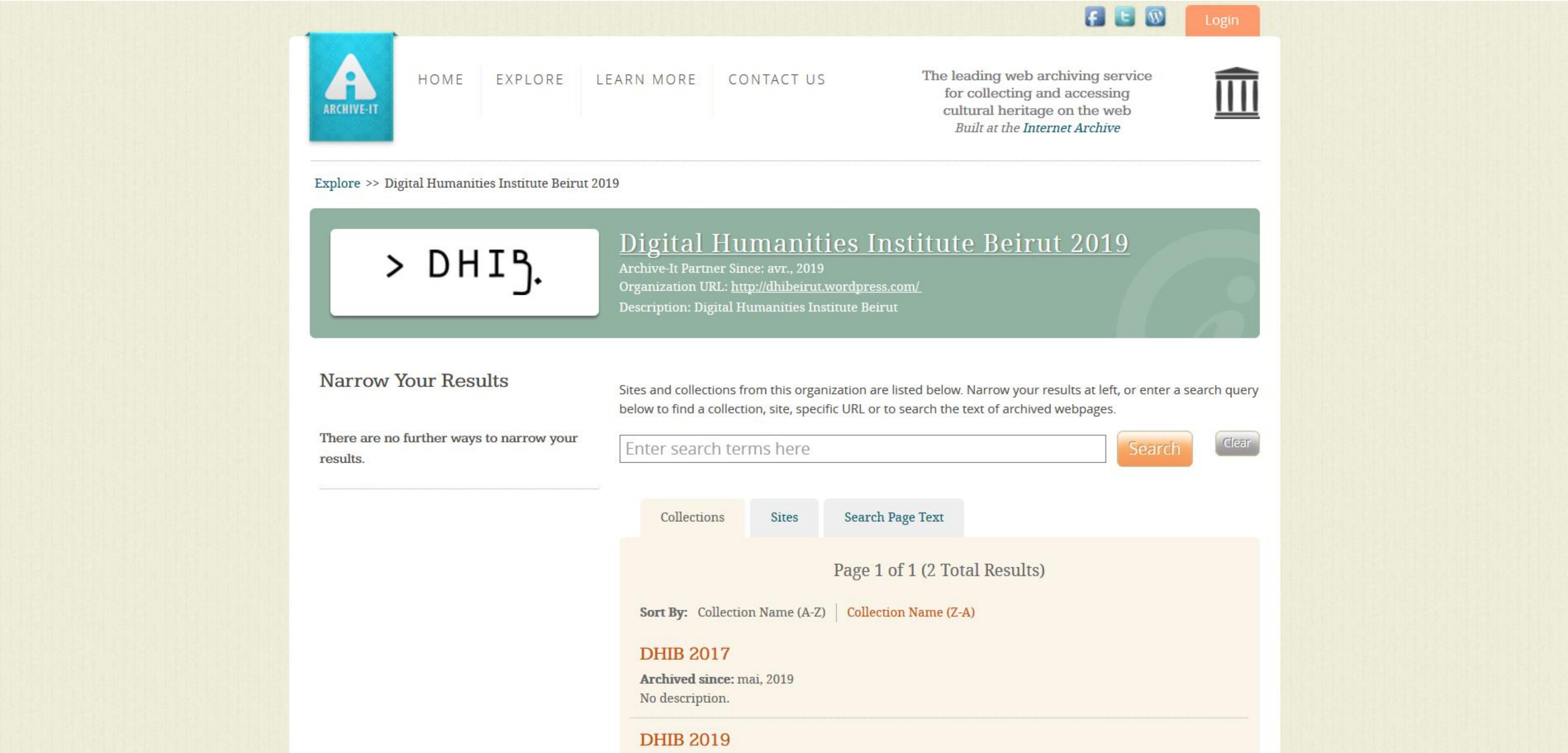
Jasmine Revolution - Tunisia 2011
By Internet Archive Global Events

A collection of websites, news coverage, and commentary surrounding the 2011 Jasmine Revolution in Tunisia. Our partners at

Explore Collecting Organizations [Find an Organization by Name] Search *Show All Organizations*

https://archive-it.org/

Archive-It: project public collection



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.

Page 1 of 1 (2 Total Results)

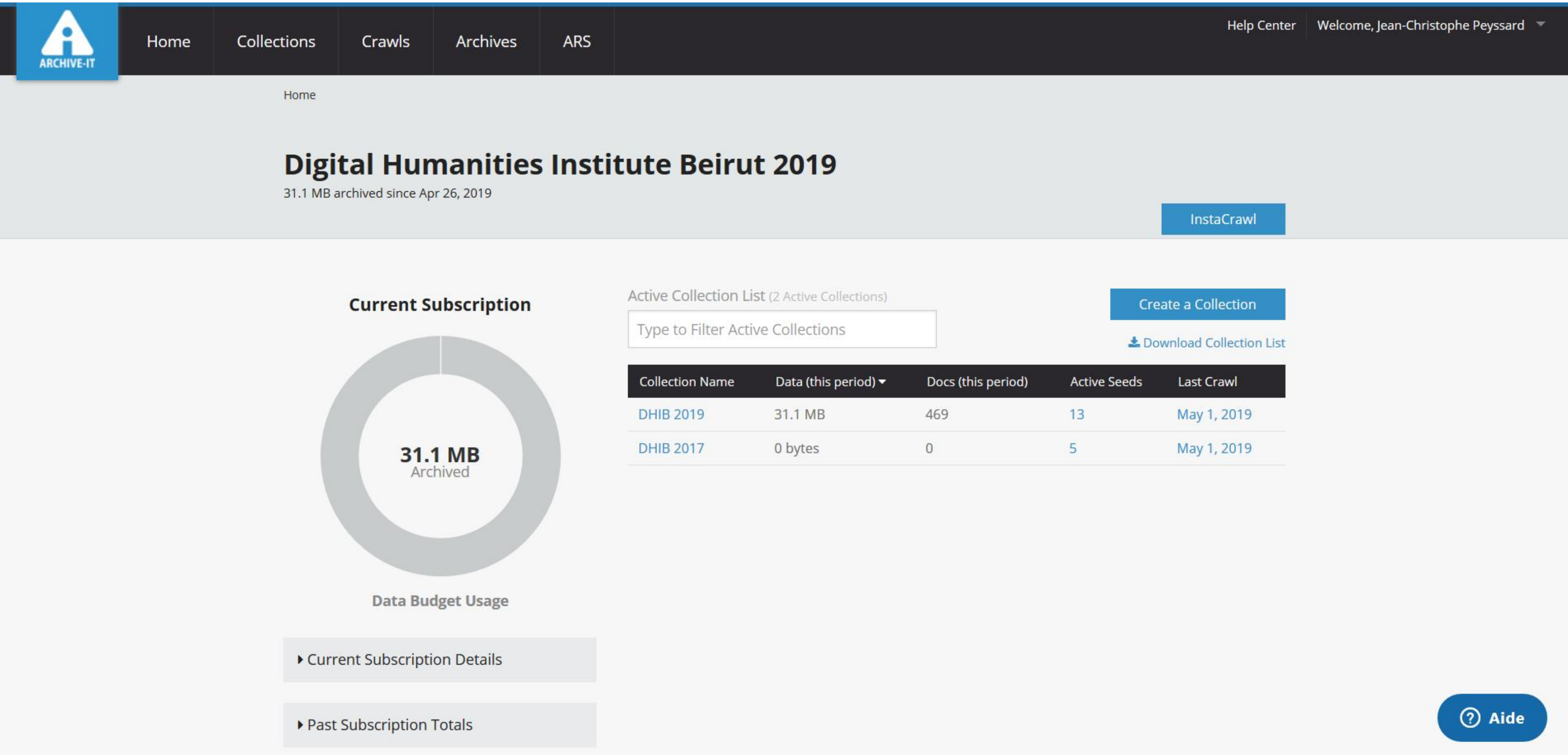
Sort By: Collection Name (A-Z) | Collection Name (Z-A)

DHIB 2017
Archived since: mai, 2019
No description.

DHIB 2019

https://archive-it.org/home/dhib

Archive-It: the archving back office user interface



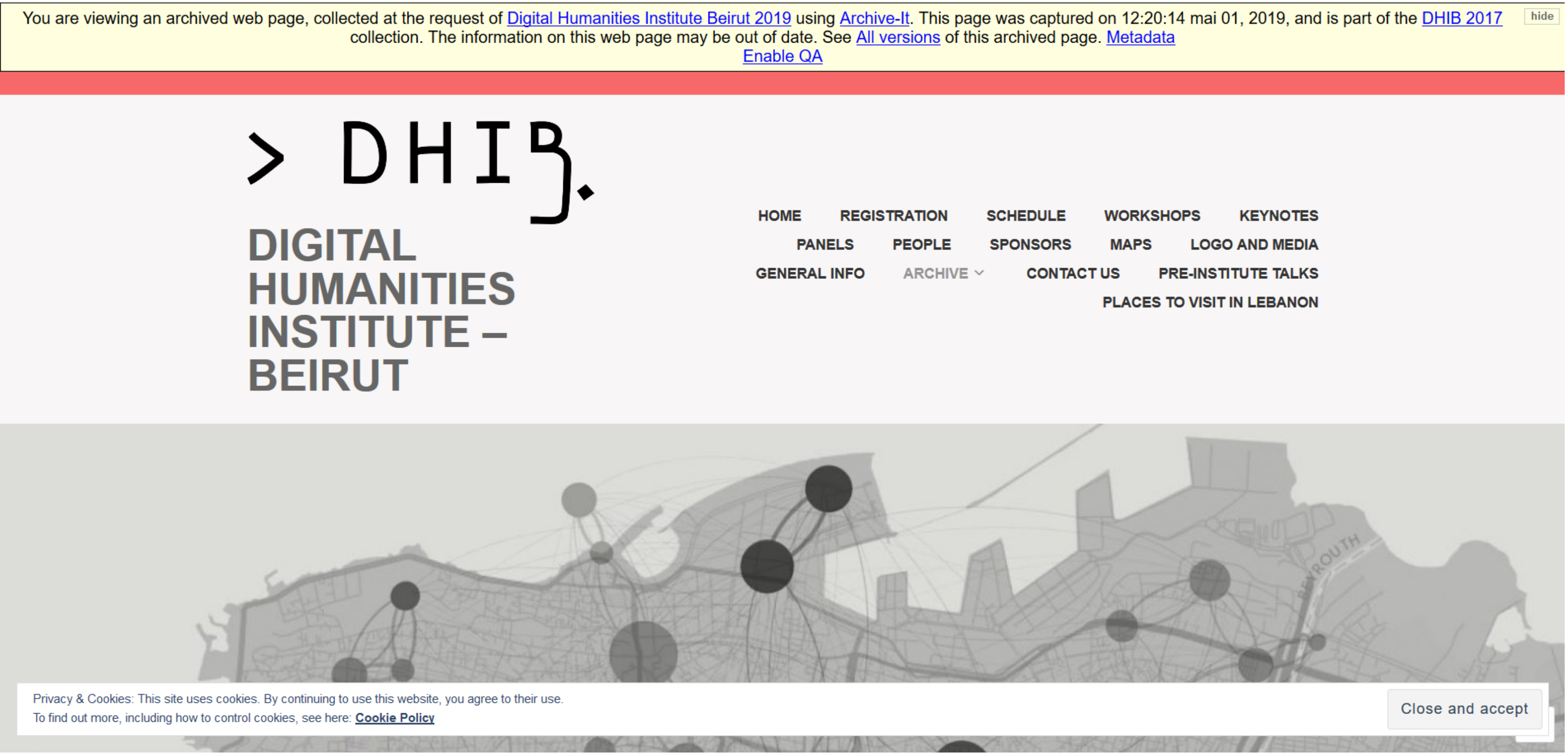
Digital Humanities Institute Beirut 2019

31.1 MB archived since Apr 26, 2019

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

https://partner.archive-it.org/1562

Archive-It: example of an archived Web page



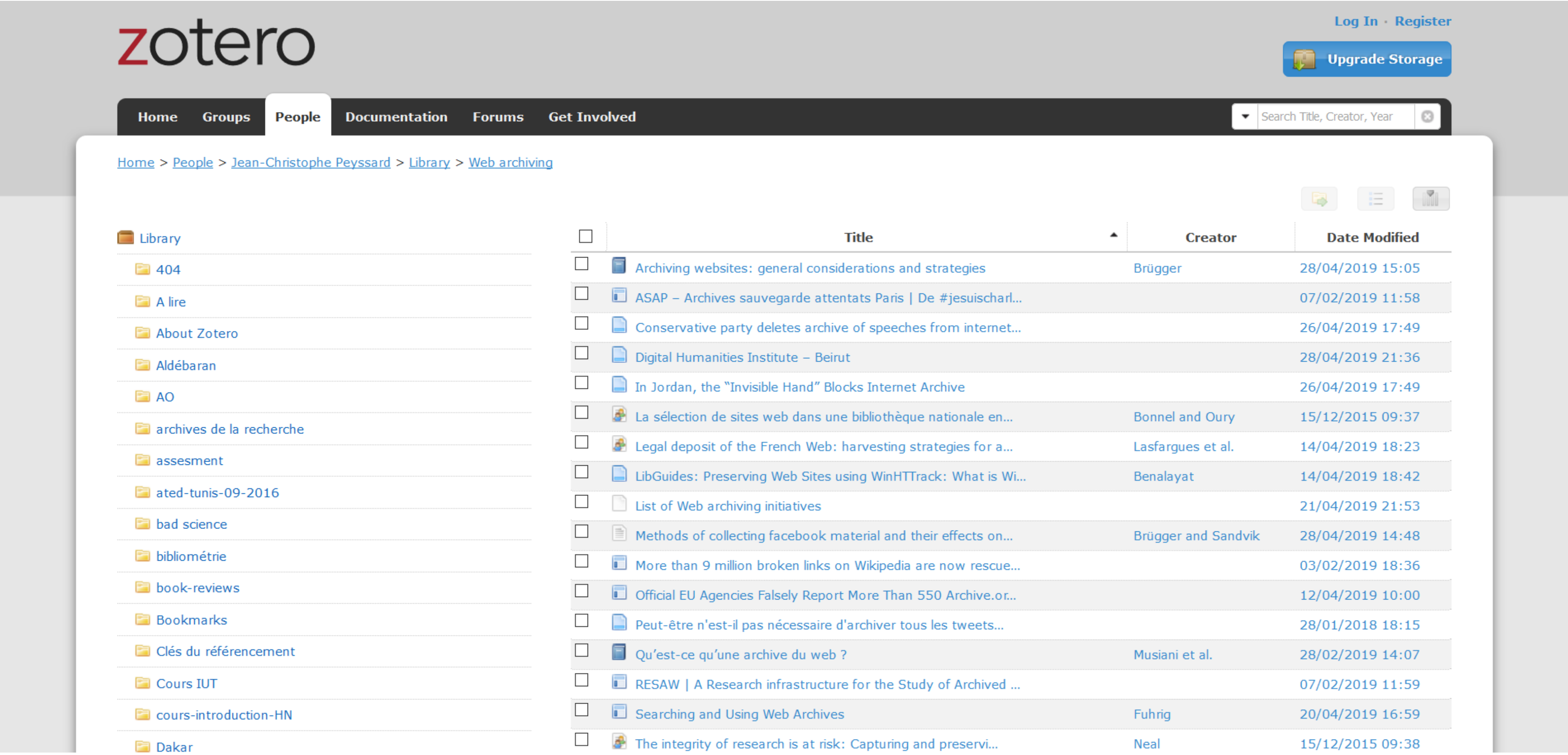
You are viewing an archived web page, collected at the request of Digital Humanities Institute Beirut 2019 using Archive-It. This page was captured on 12:20:14 mai 01, 2019, and is part of the DHIB 2017 collection. The information on this web page may be out of date. See All versions of this archived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

Enable QA

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**

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

Bibliography



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

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

https://www.zotero.org/peyssard/items/collectionKey/CNGX47Z3

Thank you / Merci / شكراً

> DHIB. 2019

jc.peyssard@ifporient.org

cnrs

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE
DE L'EUROPE ET DES
AFFAIRES ÉTRANGÈRES

UMIFRE
Unités mixtes des instituts
français de recherche à l'étranger

ifpo
Liban لبنان
T. palestiniens
فلسطين
Syrie سورية
الأردن Jordanie
Irak العراق