Le Web comme source ou terrain en SHS : cadre théorique et méthodologique de l'archivage du web et de son analyse

Séminaire WebLab – Humathèque Condorcet "Le Web et les archives du Web pour la recherche en SHS", Médiathèque de la MMSH, Aix-en-Provence (hybride), jeudi 9 octobre 2025, 14h-16h

Maya Anderson-Gonzalez (Humathèque Condorcet)
Sophie Gebeil (TELEMMe, WebLab)
Jean-Christophe Peyssard (MMSH, WebLab)

TELEMMe
Temps, Espaces, Langages, Europe Méridionale, Méditerranée
UMR 7303

IUF institut universitaire de France

Atelier Visual Studies et humanités numériques TELEMMe

CAMPUS CONDORCET HUMATHÈQUE

Maison méditerranéenne des sciences humaines et sociales
UAR 3125

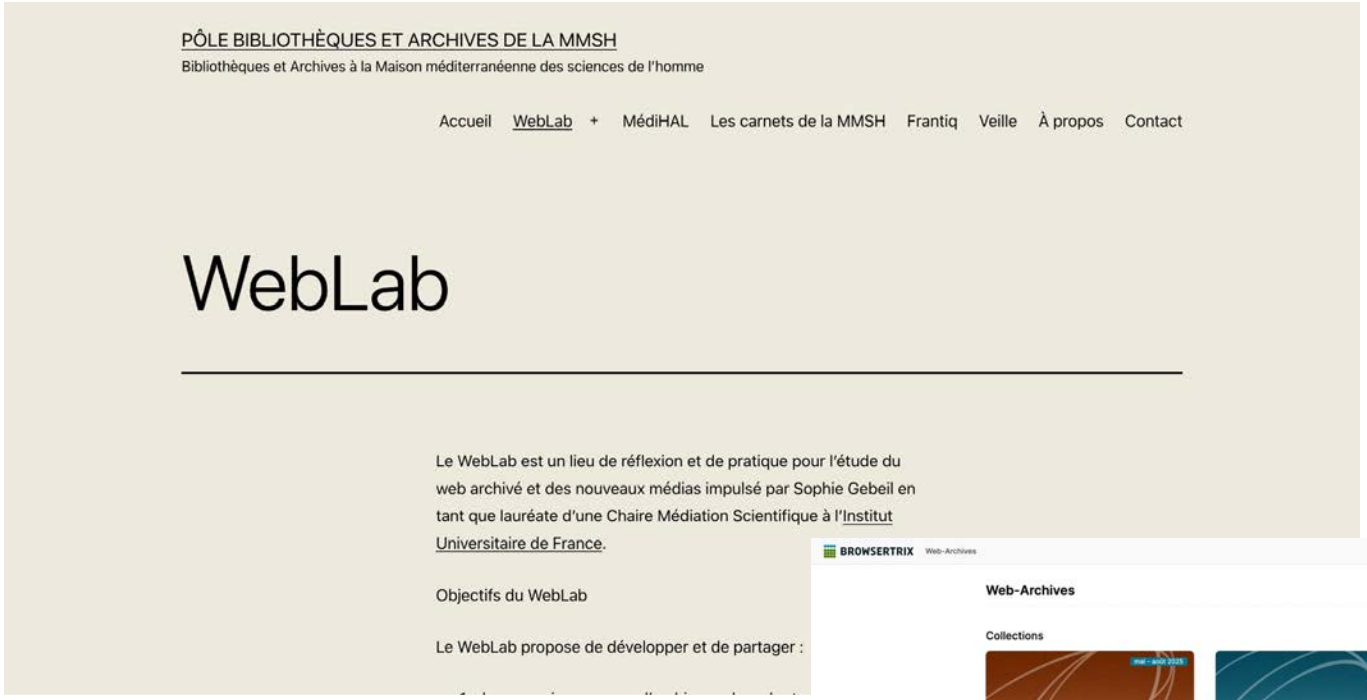
amU Aix Marseille Université

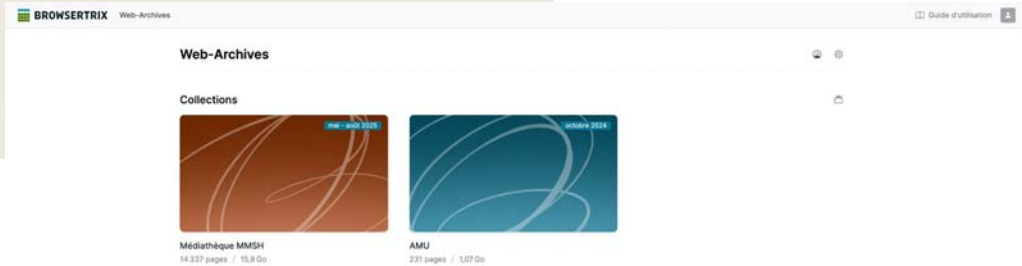
cnrs

Présentation des organisateurices

- Sophie GEBEIL, Maître de conférences en Histoire (TELEMME, AMU-CNRS), membre de l'IUF, VP déléguée Sciences humaines et Méditerranée pour AMU
- Maya ANDERSON-GONZALEZ, ingénieure d'étude à l'Humathèque Condorcet, chargée d'accompagnement pour les données du web au sein du service Accompagnement de projet et science ouverte (SAPSO)
- Jean-Christophe PEYSSARD, responsable de la Médiathèque de la MMSH

Présentation du WebLab





https://pba.mmsh.fr/?page_id=1465

https://archives-poc.cedre.univ-amu.fr/explore/web-archives

Présentation de l'Humathèque Condorcet

Le Campus Condorcet réunit onze établissements d'enseignement supérieur et de recherche.

- Le Centre national de la recherche scientifique
- L'École des hautes études en sciences sociales
- L'École nationale des chartes
- L'École pratique des hautes études
- La Fondation maison des sciences de l'homme
- L'Institut national d'études démographiques
- L'Université Paris 1 Panthéon-Sorbonne
- L'Université Sorbonne Nouvelle
- L'Université Paris 8 Vincennes-Saint-Denis
- L'Université Paris Nanterre
- L'Université Sorbonne Paris Nord

CAMPUS CONDORCET HUMATHÈQUE

L'Humathèque Condorcet du Campus Condorcet est le résultat d'une mutualisation des collections de 50 bibliothèques issues de 8 établissements.

Offre de service de l'Humathèque Condorcet

CONSTITUER DES CORPUS DE DONNÉES

L'Humathèque peut s'engager de manière **opérationnelle** en intervenant dans l'ensemble des étapes du cycle de vie des données.

Accompagnement méthodologique : définition des corpus, rédaction de guides de collecte des métadonnées adaptés aux corpus et entrepôts, etc.

Accompagnement technique : numérisation, aide à la saisie de métadonnées, dépôt de données et métadonnées dans des entrepôts à la pièce ou en masse via les APIs, etc.

https://www.humatheque-condorcet.fr/fr/pour-la-recherche/science-ouverte-et-donnees-de-la-recherche

Contexte Humathèque : étude et réalisation

Mission d'étude (2024-2025)

Panorama des données du web

mobilisées au sein des institutions résidentes et membres du Campus Condorcet

Mission de réalisation (2026-2027)

Prototyper un service

d'exploitation des données du web à l'Humathèque, suite à l'étude d' évaluation d'opportunité et de faisabilité

La collaboration HTQ <> WebLab

- Les problématiques de la recherche et de la conservation des données issues du web sont véritablement deux maillons d'une même chaîne de traitement de données.

- Les bibliothèques jouent un rôle central dans l'accompagnement des nouvelles pratiques de recherche fondées sur le web et ses données.

Le WebLab et l'Humathèque associent leur expertise pour s'engager dans la transformation des pratiques de recherche via :

- la création de connaissances ;
- la formation sur le sujet ;
- la sensibilisation ;
- le partage d'outils communs.

Ensemble, nous souhaitons également renforcer la place de nos communautés dans les réseaux internationaux liés aux archives du web (RESAW, IIPC).

La collaboration HTQ <> WebLab

Dans ce contexte, le WebLab et l'Humathèque Condorcet s'engagent à :

1. **Créer un écosystème (technique) commun** afin de susciter de nouveaux projets de coopération.
2. **Mener un travail collectif** sur la méthodologie d'accompagnement, de stabilisation et de pérennisation des corpus de données du web éventuellement constitués.
3. **Renforcer les liens entre les structures** existantes ayant une expertise autour des données du web et de leur analyse.
4. **Réfléchir ensemble à la constitution de corpus** de données du Web (vivant) à des fins d'expérimentation et d'analyse.

Le Web comme source ou terrain en SHS :

cadre théorique et méthodologique de l'archivage du web et de son analyse

Partie 1 : Institutions SHS et données du Web

Partie 2 : Le Web et ses archives

Partie 3 : Une communauté internationale structurée

Partie 4 : Défis pour les SHS

Partie 5 : Programme prévisionnel du séminaire du WebLab

Partie 1 :

Institutions SHS et données du Web

Un réseau de partenaires à remobiliser

Participation de l'Humathèque Condorcet au ResPaDON (2021-2023)

Réseau de Partenaires pour l'analyse et l'exploration de données numériques, *Améliorer l'exploitation des archives du web par les communautés de recherche : Les 15 préconisations du projet ResPaDon*, https://respadon.hypotheses.org/2768, 18 décembre 2023.

Des données du web “vivant” ?

« Par ‘données issues du Web’, nous entendons les **données construites à partir des traces des activités sociales en ligne** des internautes, que celles-ci soient fournies par les sites web qui organisent ces activités, ou qu’elles soient construites par le chercheur à partir des informations visibles sur le Web. »

BEUSCART Jean-Samuel, « Des données du Web pour faire de la sociologie… du Web ? » in *Big data et traçabilité numérique : Les sciences sociales face à la quantification massive des individus*, Pierre-Michel Menger et Simon Paye (éd.), Paris, Collège de France, « Conférences », 2017, p. 141-161.

Etat des lieux

Une tendance :
des pratiques et méthodes propres à chaque communauté disciplinaire en SHS + des équipes et services d'accompagnement en quête de structuration et de standardisation à travers le territoire.

Un objectif :
insérer les données du web dans une chaîne de traitement des données de la recherche selon un processus structuré et une logique itérative.

Des perspectives :
des acteurs institutionnels souhaitant se constituer en communautés de pratique articulant différents métiers et profils complémentaires de l'ESR.

Compétences métiers multiples et complémentaires

- ★ Expertise thématique et technique
- ★ Ingénierie documentaire et de recherche
- ★ Conception archivistique, patrimoniale & design

L'outillage : point de jonction méthodologique ?

- ***Web scraping***

  Rétroingénierie de pages web : identifier le code HTML qui nous intéresse et le « racler » pour obtenir des données brutes non-structurées.

- ***Web crawling***

  Indexation automatisée de contenus trouvable sur le web (robot).

- ***Web mining***

  Analyse approfondie de jeux de données massives.

https://hyphe.medialab.sciences-po.fr/

Panorama des institutions

Densification et diversification du paysage des acteurs de l'exploitation des données du web pour la recherche en SHS. Apparition d'infrastructures et de services dédiés au sein de structures documentaires :

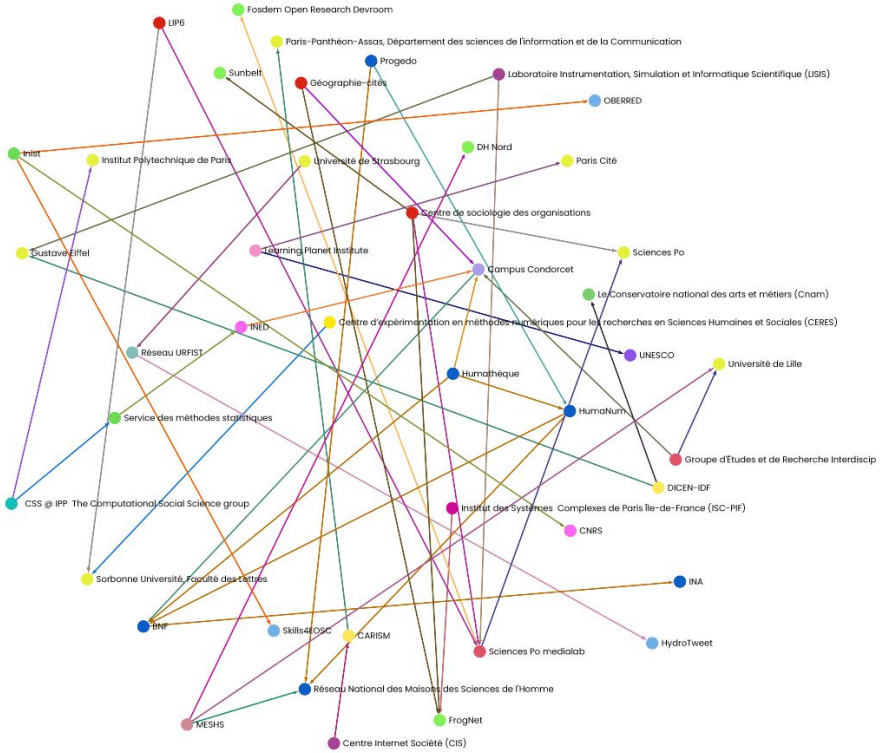
le medialab de Sciences Po ;

le BnF DataLab ;

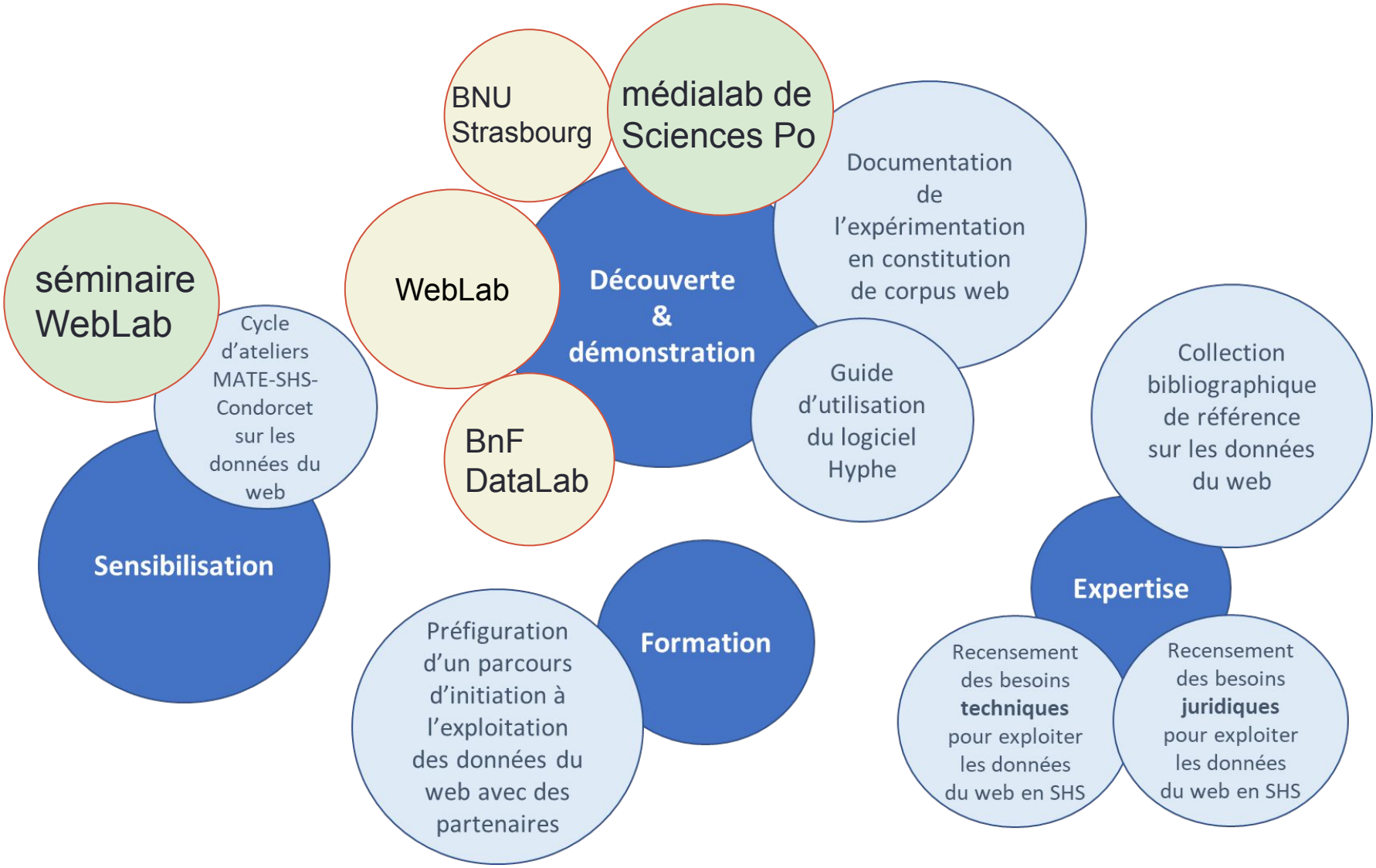
le Lab de l'INA ;

le Datalab de la BNU de Strasbourg ;

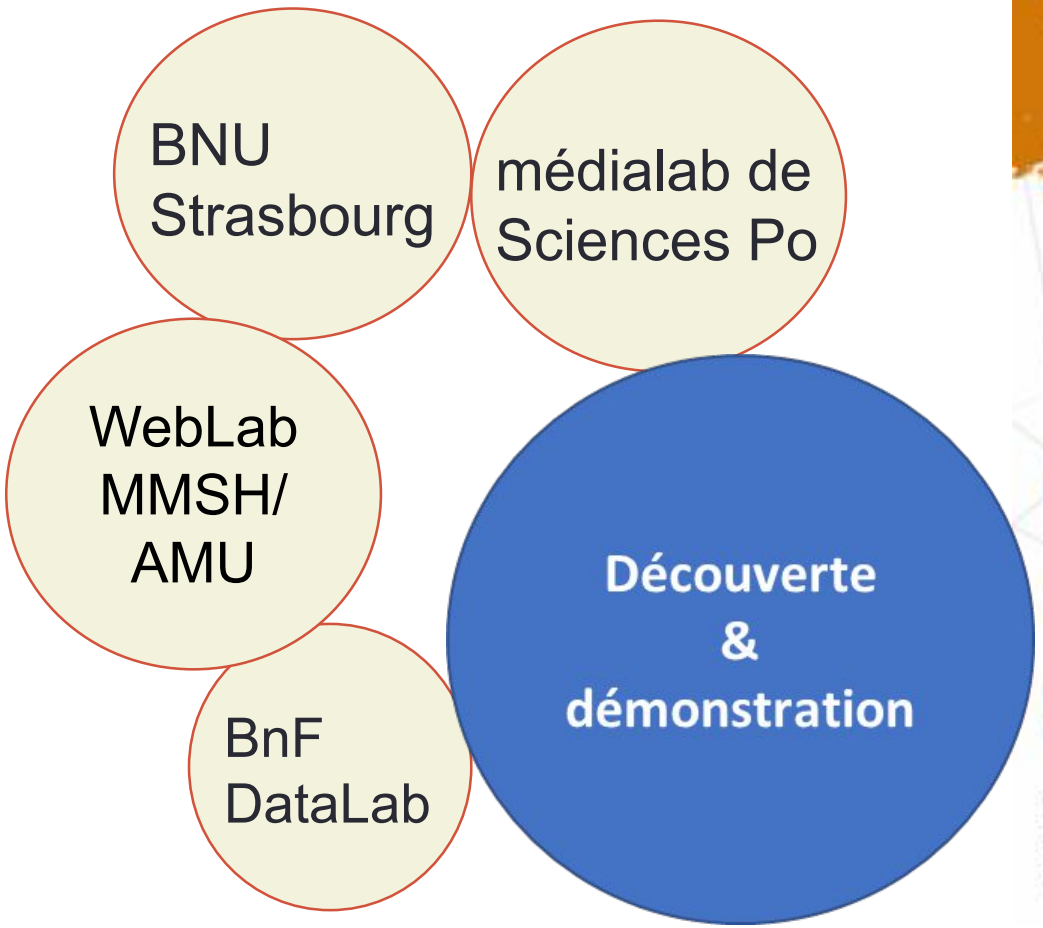
le WebLab AMU/MMSH.



Des actions à mener en partenariat



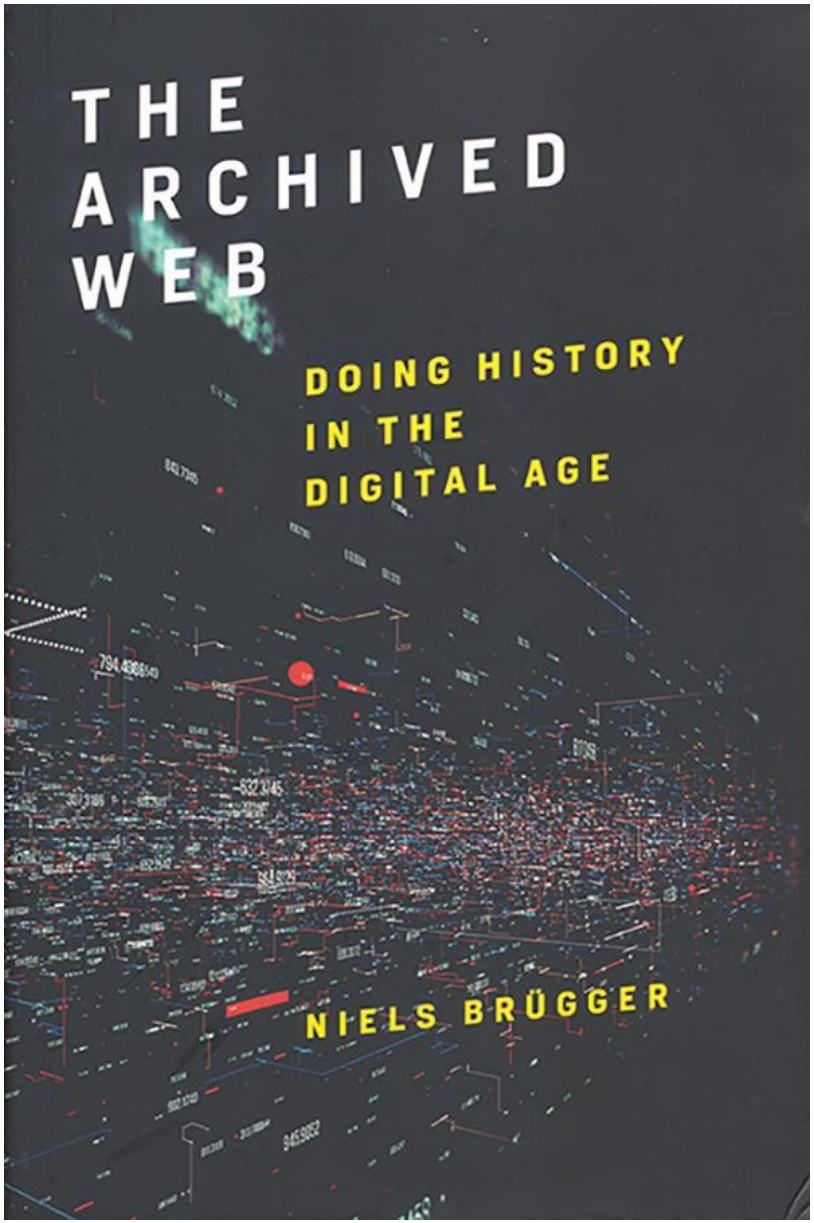
Prochain événement à l'Humathèque : 18-20 novembre 2025



Partie 2 :

Le Web et ses archives

Le manuel de référence



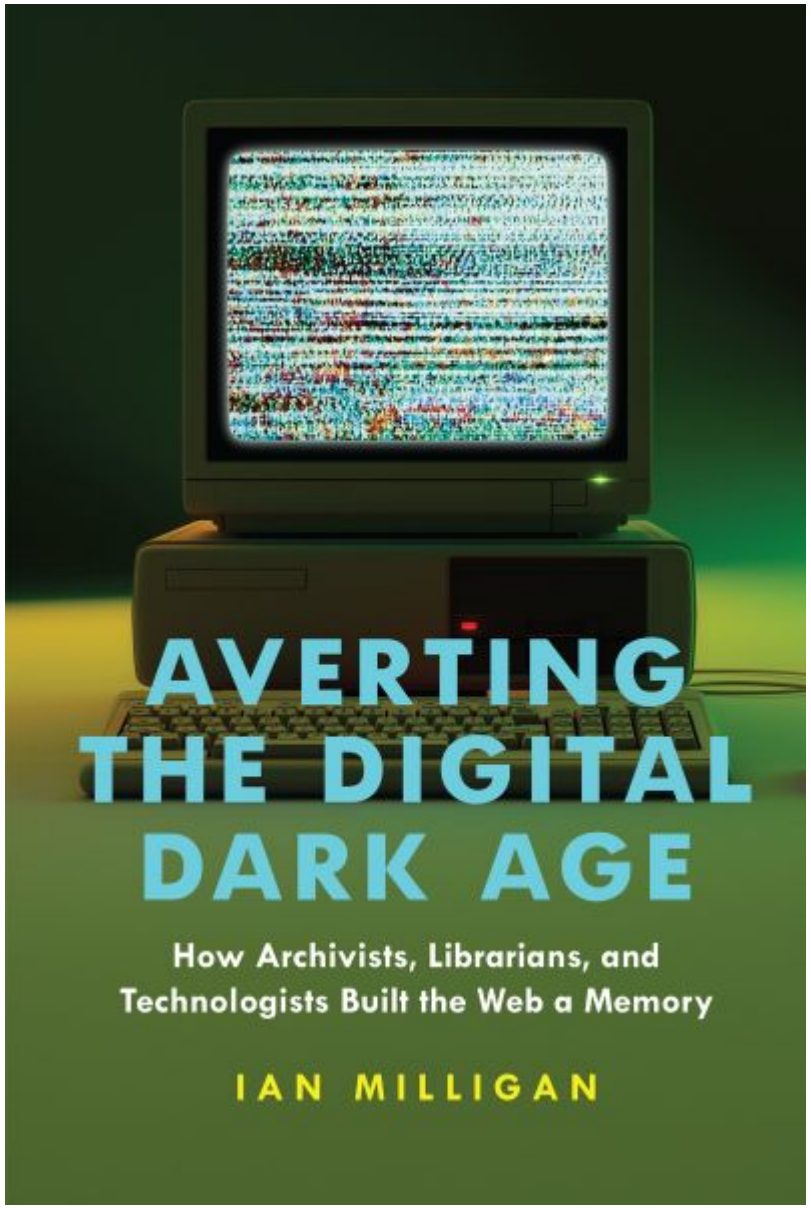
Brügger, Niels. 2018. *The Archived Web: Doing History in the Digital Age*. Cambridge, Massachusetts: The MIT Press. https://search.worldcat.org/fr/search?q=bn:9780262039024.

Deux références en français

Musiani, Francesca, Camille Paloque-Bergès, Valérie Schafer, and Benjamin G. Thierry. 2019. *Qu'est-ce qu'une archive du web ?* Encyclopédie numérique. Marseille: OpenEdition Press. https://doi.org/10.4000/books.oep.8713.

Gebeil, Sophie. 2021. *Website story: histoire, mémoires et archives du web*. Bry-sur-Marne, France: INA. https://www.cairn.info/website-story--9782869382824.htm.

Un récit et une analyse critique de la courte histoire des archives du Web



https://muse.jhu.edu/book/123276

Les médias numériques

- Matériaux numérisés / Digitized material

  Matériaux analogiques qui ont été numérisés : manuscrits numérisés, photographies analogiques, médias électroniques (émission de radio et télévision) ;
- Matériaux nativement numériques / Born-digital material

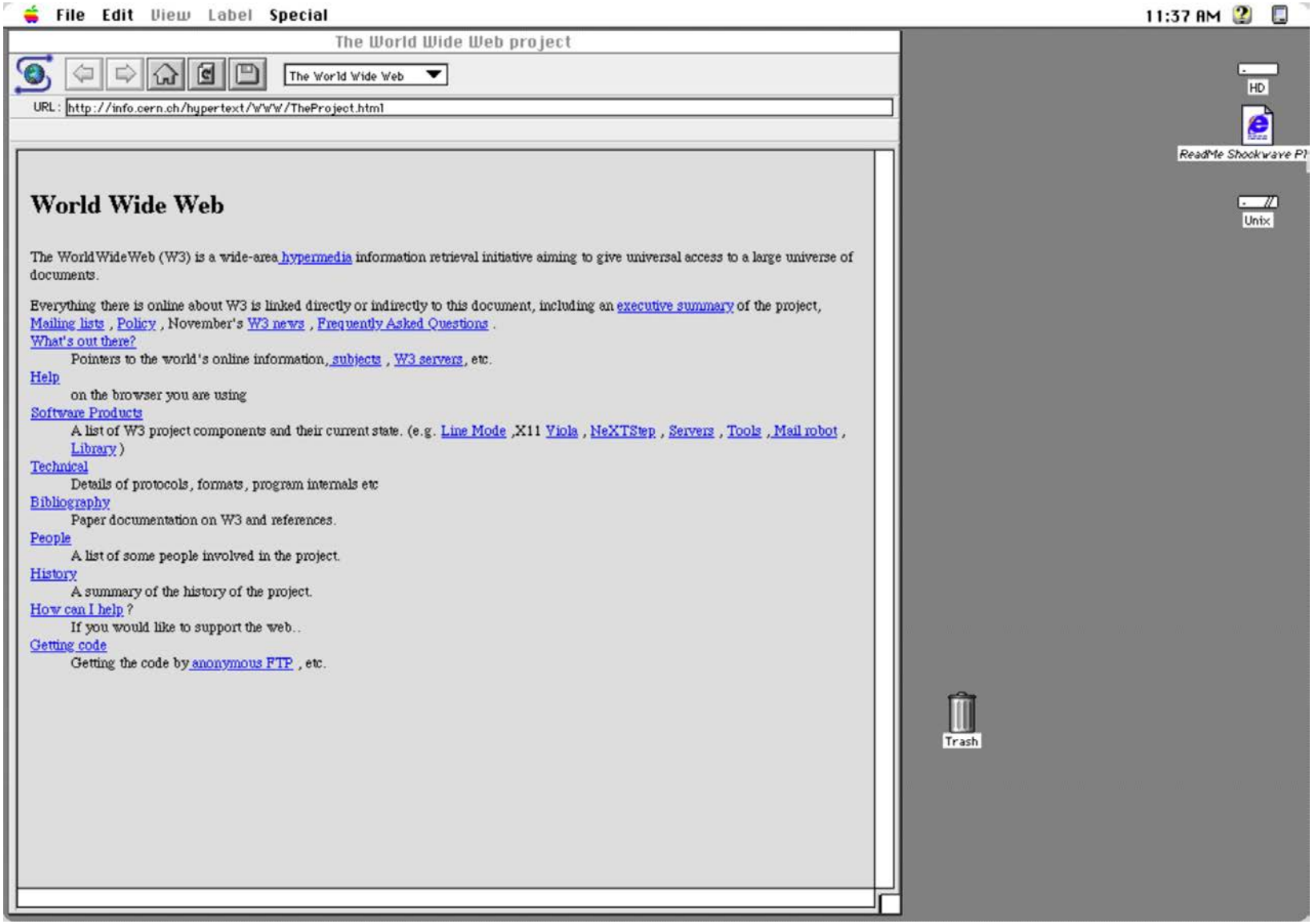
  Qui n'a jamais existé autrement que sous une forme numérique : un texte numérique, une image numérique, un son numérique, CD Rom, DVD, sites Web ;
- Matériaux numériques « régénérés » / reborn-digital material

  Matériaux nativement numériques collectés et conservés : jeux vidéos émulés, archives du Web : reborn digital heritage.

Brügger, Niels. 2016. « Digital Humanities in the 21st Century: Digital Material as a Driving Force ». *Digital Humanities Quarterly,* 10 (3). http://www.digitalhumanities.org/dhq/vol/10/3/000256/000256.html.

World Wide Web

Le premier site Web a été publié le 6 août 1991 par le physicien britannique Tim Berners-Lee au CERN en Suisse



http://info.cern.ch/hypertext/WWW/TheProject.html

URL : Uniform Resource Locator

La structure d'une URL sur le World Wide Web (www) :

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

Exemple :

https://distam.hypotheses.org/category/ecole-dete

https://en.wikipedia.org/wiki/URL
http://dac.au.dk/forskning/forskningsprogrammer p. 51

Le World Wide Web en chiffres

En janvier 2025, le Web indexé par les moteurs de recherche contiendrait 3,98 milliards de pages
https://www.worldwidewebsize.com

En août 2025, Netcraft recense 1 300 016 299 sites répartis sur 281 522 947 domaines et 13 797 308 serveurs
(« web-facing computers »)
https://www.netcraft.com/blog/august-2025-web-server-survey

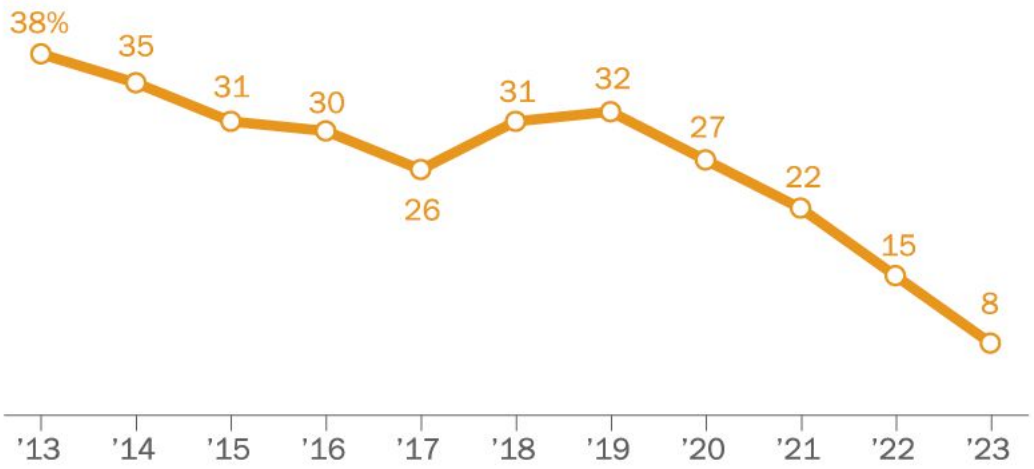
Vie et mort d'une page web

- La durée de vie moyenne d'une page web est de 44 jours ; 44% des sites web recensés en 1998, n'étaient plus trouvables en 1999 (Lyman, 2002 p. 38) ;
- 40% des contenus du web disparaît au cours d'une année, 40% sont modifiés, voilà pourquoi aujourd'hui, on peut seulement s'attendre à trouver 20% des contenus qui étaient disponibles il y a un an (Brügger, 2005, p. 15) ;
- En moyenne, une page web subira une modification ou disparaîtra, avant 100 jours (Kahle, 2015) ;
- En 2019, selon l'équipe de la Wayback Machine, la durée de vie moyenne d'une page web est de 92 jours ;

Vie et mort d'une page web

38% of webpages from 2013 are no longer accessible

% of links from each year that are no longer accessible as of October 2023



Source: Pew Research Center analysis of a random selection of URLs collected by the Common Crawl web repository (n=999,989) and checked using page and DNS response codes. Web pages defined as inaccessible if they returned a status code of 204, 400, 404, 410, 500, 501, 502, 503, 523 or did not return a valid status code.
"When Online Content Disappears"

PEW RESEARCH CENTER

Chapekis, Athena, Samuel Bestvater, Emma Remy, and Gonzalo Rivero. 2024. "When Online Content Disappears: 38% of Webpages That Existed in 2013 Are No Longer Accessible a Decade Later." Pew Research Center. https://www.pewresearch.org/data-labs/2024/05/17/when-online-content-disappears/.

Ramesh, Randeep, and Alex Hern. 2013. “Conservative Party Deletes Archive of Speeches from Internet.” *The Guardian*, November 13, 2013, sec. Politics. https://web.archive.org/web/20131113221027/https://www.theguardian.com/politics/2013/nov/13/conservative-party-archive-speeches-internet.

Liptak, Adam. 2013. “In Supreme Court Opinions, Web Links to Nowhere.” *The New York Times*, September 13, 2013, sec. Politics. https://web.archive.org/web/20200218163337/https://www.nytimes.com/2013/09/24/us/politics/in-supreme-court-opinions-clicks-that-lead-nowhere.html?hp&_r=3&#

Exemple de disparition

Roger Friedman's
SHOWBIZ★411
HOLLYWOOD TO THE HUDSON

Celebrity Fashion Media Music Politics

Paramount Global Erases Archives of MTV Website, Wipes Music, Culture History After 30 Plus Years

June 25, 2024 12:36 am | By **Roger Friedman**

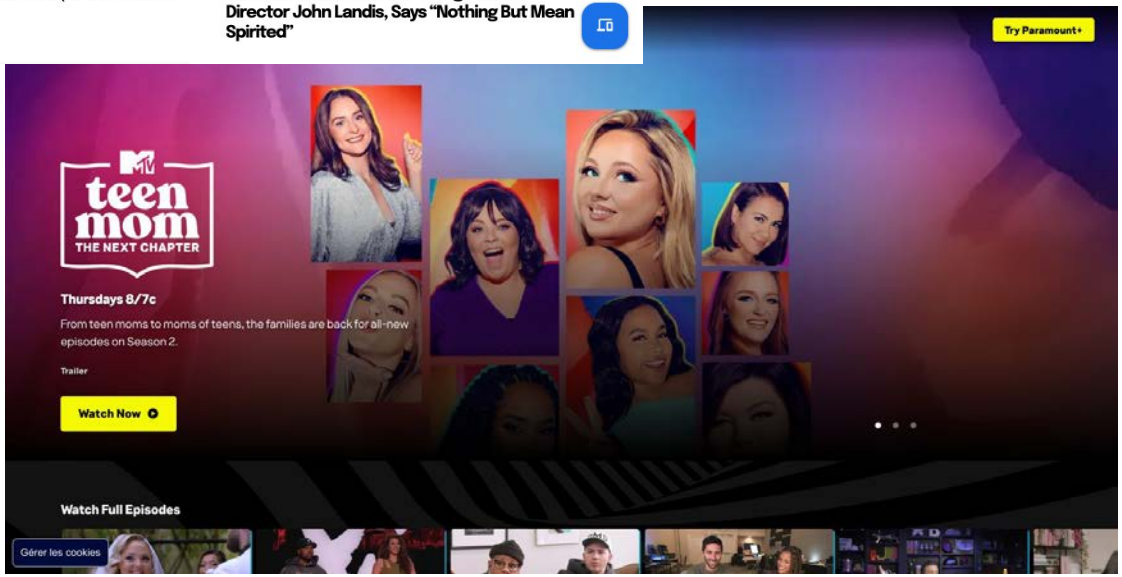
Share

MTV.com is gone. Kaput. Wiped off the face of the Earth.

Parent company Paramount Global, formerly Viacom, has tossed twenty plus years of news archives. All that's left is a placeholder site for reality shows. The M in MTV – music — is gone, and so is all the reporting and all the journalism performed by music and political writers ever written. It's as if MTV never existed. (It's the same for VH1.com, all gone.)

Read more

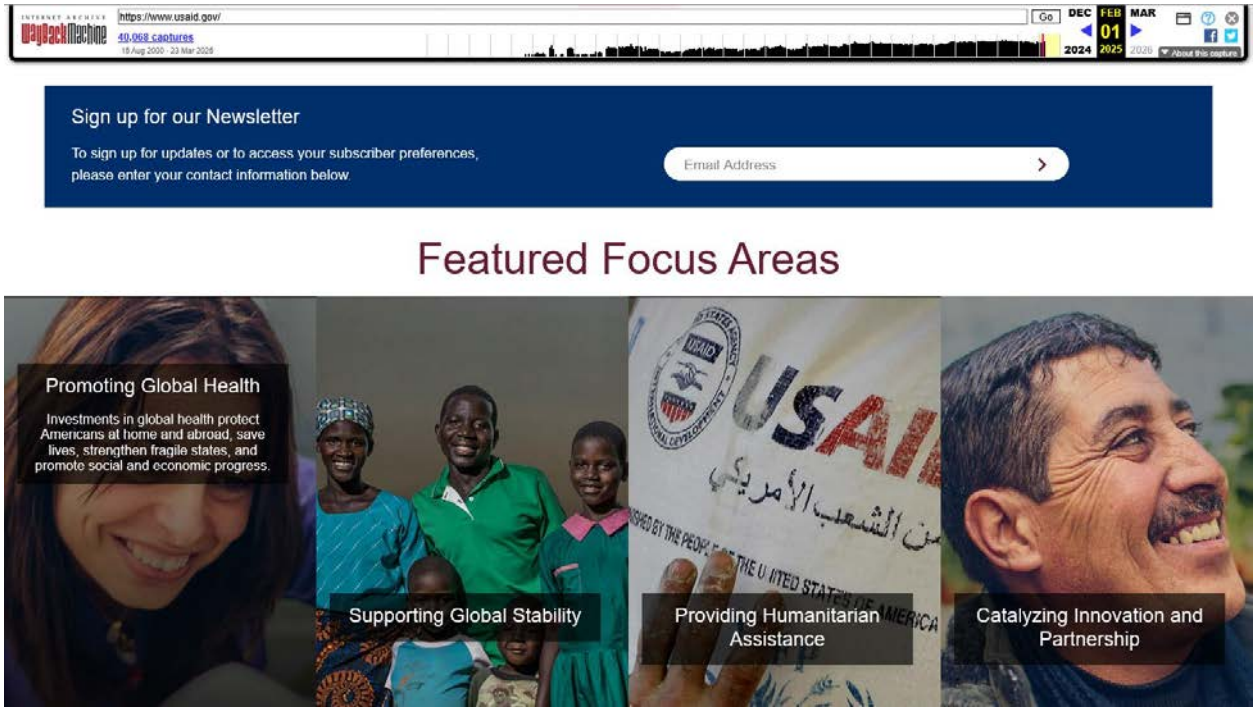
Celebrity
'70s Feud: Singer Songwriter Stephen Bishop Goes "On and On" About "Twilight Zone" Director John Landis, Says "Nothing But Mean Spirited"



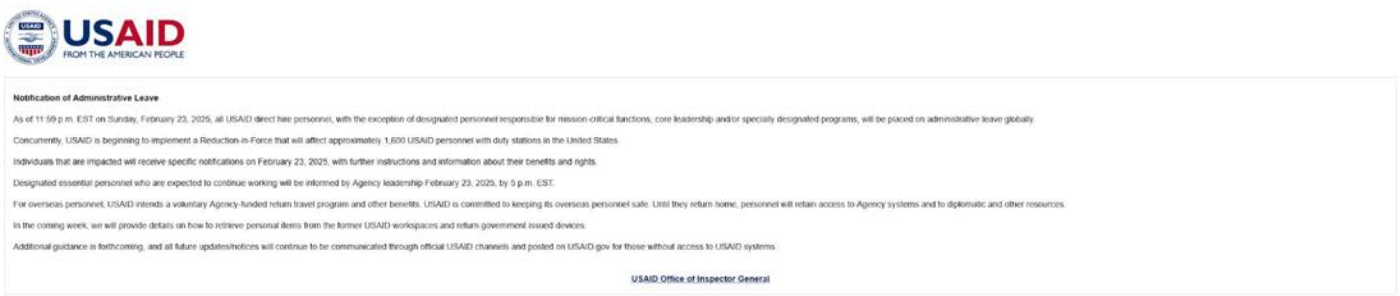
Friedman, Roger. 2024. "Paramount Erases Archives of MTV Website, Wipes Music, Culture History After 30 Plus Years." *Showbiz411* (blog). June 25, 2024. https://www.showbiz411.com/2024/06/25/paramount-shuts-down-mtv-website-wipes-history-after-20-plus-years.

Exemple de disparition

USAID



https://web.archive.org/web/20250201003843/https://www.usaid.gov/



https://www.usaid.gov/

Un problème pour la recherche et la communication scientifique

- Error 404, Broken links, Link rot, Reference rot, Infosuicide, digital ruins, content drift, zombie media,.. *Erreur 404, lien cassé, (…), suicide numérique, (…), site non maintenu ;*
- Fermetures volontaires et administratives (« take down »), fusions et acquisitions :

  Le 18 mars 2019, on apprend que *MySpace* a perdu les contenus de ses utilisateurs au cours d'un incident de migration de serveurs qui s'est mal déroulé. Plus de 50 millions de chansons et 12 années de production de contenus ont disparu pour toujours. Il n'y avait pas de sauvegarde (https://en.wikipedia.org/wiki/Myspace).
- Histoire :

  Séparation de la Yougoslavie (.yu - Serbie and Montenegro, .rs & .me) et dissolution de la Tchécoslovaquie (.cs – maintenant Republique Tchèque et Slovaquie, .cz & .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/)
- Neal, James G. 2014. "**The Integrity of Research Is at Risk**: Capturing and Preserving Web Sites and Web Documents and the Implications for Resource Sharing." In . Lyon, France. http://library.ifla.org/id/eprint/907.)

Pourquoi archiver le Web ?

Pourquoi archiver le web ?

- Conserver notre patrimoine numérique culturel et scientifique ;
- Stabiliser et conserver les contenus du web en tant qu'objet de recherche ;
- Administrer la preuve et citer les sources.

Nielsen, Janne. 2016. *Using Web Archives in Research: An Introduction*. Aarhus (Danemark): NetLab. https://digital.library.unt.edu/ark:/67531/metadc1743080/. p. 7.

Définition de l'archivage du web

- 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, Jinfang. 2012. "An Overview of Web Archiving." *D-Lib Magazine* 18 (3/4). https://doi.org/10.1045/march2012-niu1.)

Brève chronologie de l'archivage du web

1537 Création du dépôt légal en France (Espagne, 1619 ; Royaume-Uni, 1710)
1989 Le World Wide Web est inventé par Tim Berners-Lee
1996 Fondation d'Internet Archive par Brewster Kahle
1996 Début de l'archivage du web en Suède (domaine national .se) avec le projet Kulturarw3
1998 Lancement de Google
2001 Création de la Wayback Machine (Internet Archive)
2003 Publication de la charte de l'UNESCO sur la conservation du patrimoine culturel numérique
2003 Création du IIPC, International Internet Preservation Consortium à la bibliothèque nationale de France, avec 12 institutions partenaires
2005 Lancement de Youtube
2006 Lancement de Facebook et Twitter
2006 La bibliothèque nationale de France (BnF) et l'institut national de l'audiovisuel (Ina) sont chargés d'assurer le dépôt légal du web
2013 Création du projet EU Web Archive (https://op.europa.eu/fr/web/euwebarchive)
2025 1 000 milliards de pages web archivées par l'Internet Archive et accessibles via la Wayback Machine

Les stratégies & méthodes pour l'archivage du web

- Macro archivage ;
- Micro archivage ;
- Archivage sélectif et thématique ;
- Moissonnages par lots, à large spectres ou de *snapshots* ;
- Tentative d'exhaustivité d'archivage, par exemple à partir d'un nom de domaine national (.se en Suède) ;
- Archivage instiutionnel d'événements ou archivage en temps réel, ex l'incendie de Notre Dame en 2019 ;
- Coopération et partage d'archivage entre différentes institutions (BnF / Ina)
- …

Les différentes façons d'accéder aux archives du web

- Les archives du web en Finlande (depuis 2006): https://www.kansalliskirjasto.fi/en/collections-and-content-online#finnish-web-archive

  Les archives du web finlandais sont accessibles uniquement depuis les terminaux du dépôt legal du web à la bibliothèque nationale et dans un reseau de bibliothèques publiques sur le territoire.

- Au Portugal (depuis 2008) les archives du web sont accessibles en libre accès Arquivo.pt

- La Wayback Machine (depuis 2006) est accessible en libre accès https://archive.org/web/

Un large éventail d'archives du web

Comme pour les autres types d'archives, il est utile de connaître le contexte de production de l'archive pour mieux la comprendre et l'utiliser dans un travail académique. Lorsque vous consultez une archive du web, il s'agit d'une reconstruction et pas d'une simple copie.

- "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)

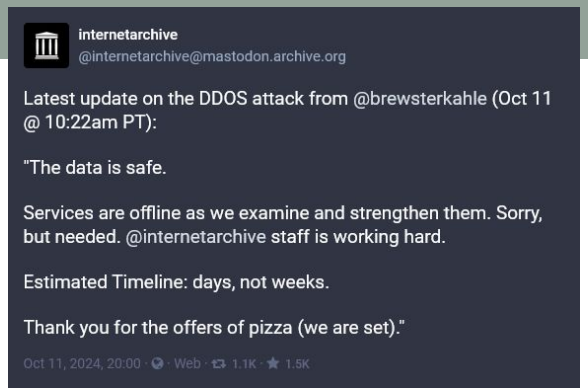
Les projets d'archivage du web ont besoins des compétences de différents acteurs et de leurs expertises : chercheur·e·s, archivistes, bibliothécaires, juristes, informaticien·ne·s, spécialistes de la données, … usagères et usagers, partenaires et membres de la société civile.

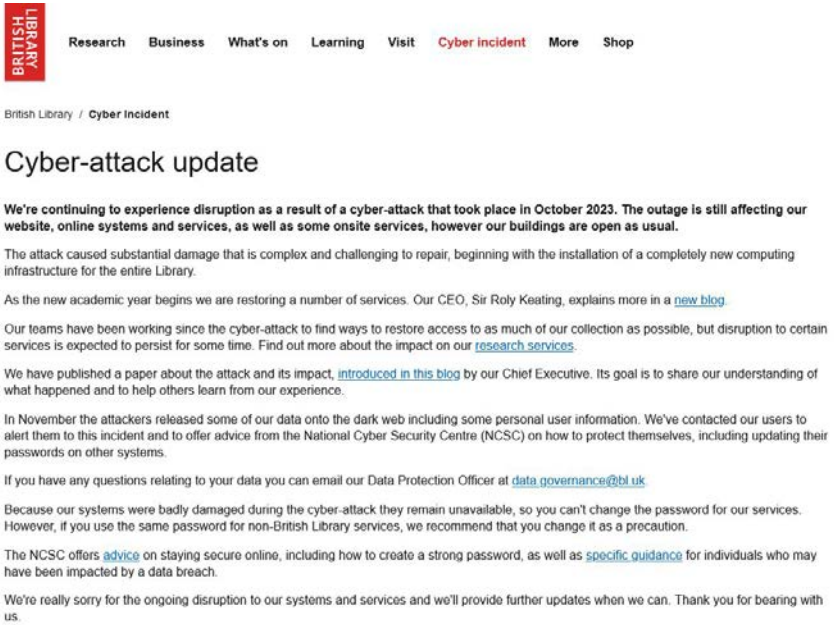
Questions éthiques et juridiques

Comme pour les autres types d'archives, il est nécessaire d'agir dans le cadre légale et éthique depuis le processus d'archivage et jusqu'à l'usage qui sera fait des archives du Web :

- Les documents contenus dans les archives Web sont protégés par les lois sur le droit d'auteur, comme ils le sont sur le Web "vivant".
- Il existe des "tensions entre les principes archivistiques de préservation des documents publics et les attentes des citoyens en matière de droit à l'oubli" (Bingam, 2018)
- Le traitement des données personnelles est soumis à des lois et plus encore à l'éthique du projet de recherche.
- règlement général sur la protection des données (RGPD) / GPDR / General Data Protection Regulation (GPDR)

Peur, incertitude et doute sur les archives du Web





Service Availability

Wayback Machine (provisional, read-only) service.

Other Internet Archive services are temporarily offline.

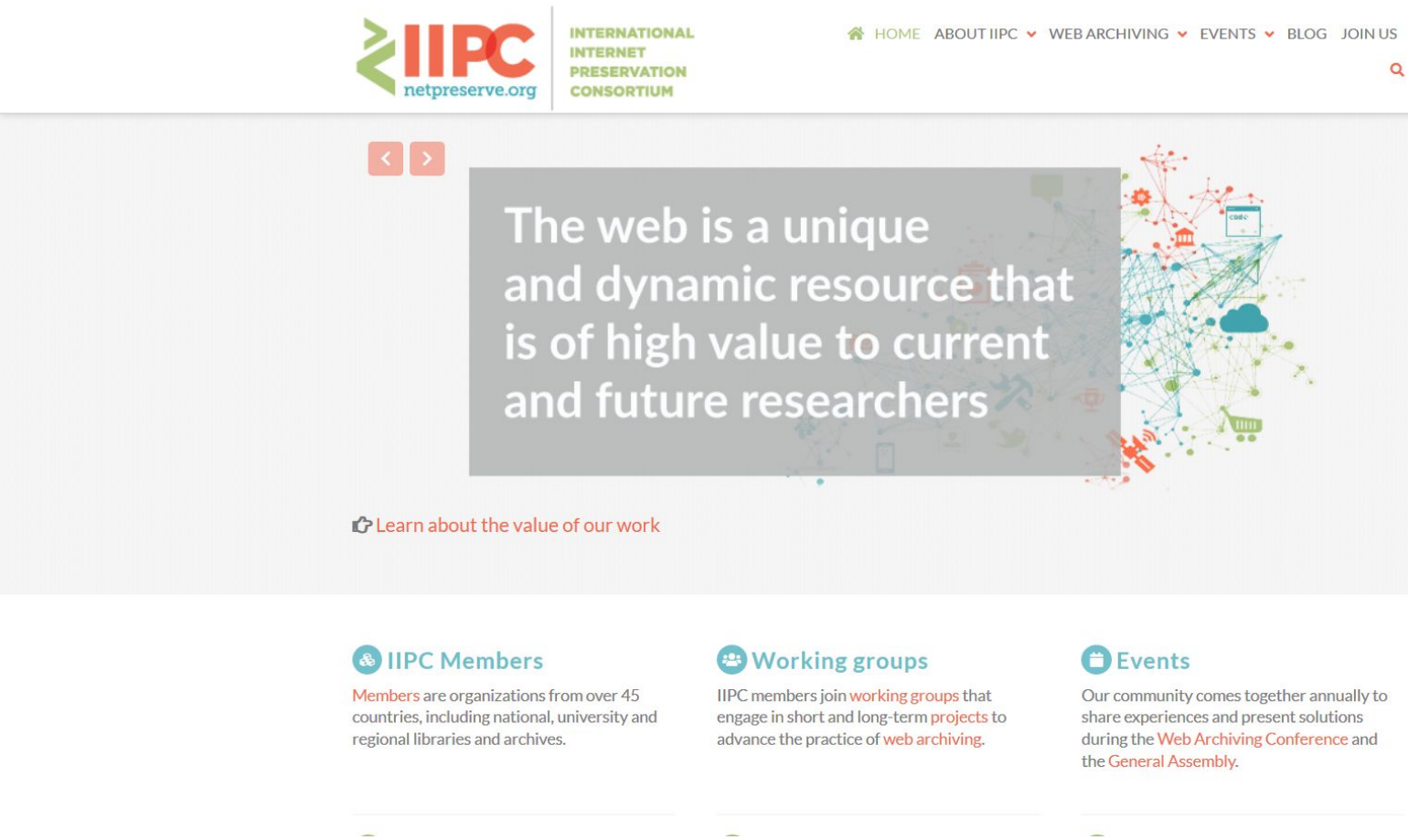
Please check our official accounts, including Twitter/X, Bluesky or Mastodon for the latest information.

We apologize for the inconvenience.

https://www.bl.uk/cyber-incident/

https://mastodon.archive.org/@internetarchive/113290094683712709

Partie 3 : une communauté internationale structurée



http://netpreserve.org/

La Wayback Machine

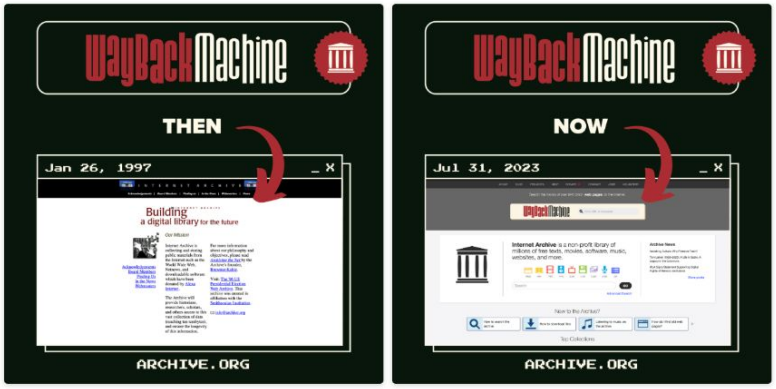
Internet Archive Blogs

A blog from the team at archive.org

Blog | Announcements | archive.org | About | Events | Developers | Donate

Calling All Libraries: Celebrate 1 Trillion Web Pages Archived with the Internet Archive

Posted on October 7, 2025 by Chris Freeland



The Internet Archive has released a new resource guide to help libraries join in commemorating a once-in-a-generation milestone: 1 trillion web pages archived by the Internet Archive and available for use via the Wayback Machine. This historic achievement represents a collective effort to preserve our shared digital history, and libraries have been at the heart of that work from the beginning. The resource guide provides background on why this milestone matters to libraries, highlights their role in safeguarding online knowledge, and offers ready-to-use materials to help libraries celebrate with their communities.

Celebrate with the Internet Archive this October: 1 trillion web pages archived

Search

Latest Posts

- Calling All Libraries: Celebrate 1 Trillion Web Pages Archived with the Internet Archive
- Happy College Radio Day! Explore Student Radio's Past and Present in the DLARC College Radio Collection
- Peer-to-Peer Fundraising at the Internet Archive
- Rural Libraries Receive Support from the Internet Archive to Preserve Community Stories
- Sir Tim Berners-Lee to Receive the 2025 Internet Archive Hero Award

Upcoming Events

OCT 9 7:30 pm - 8:30 pm Building and Preserving the

- Lancé en 2001
- 1000 milliards de pages web archivées (oct 2025)
- Archives remontant jusqu'en 1996

https://web.archive.org/

https://cc.au.dk/en/resaw

RESAW - Infrastructure de recherche pour l'étude des sources issues du web archivé

RESAW

- About
- Events
- RESAW timeline
- Participants
- Web Archives

Home

RESAW, a Research Infrastructure for the Study of Archived Web Materials, is a community established in 2012, aiming at promoting a collaborative European research infrastructure for the study of archived web materials.

Partie 4 : L’archivage du Web. Quels défis pour les SHS ?

Du Web comme source aux défis épistémologiques

https://presses-universitaires.univ-amu.fr/fabrique-memoires-limmigration-maghrebine-web-francais-1999-a-2014

L'archivage du Web comme défi historiographique : entre fragmentation et médiation (Projet IUF)
>> Questionnaire sur les usages https://madi.hypotheses.org/1436

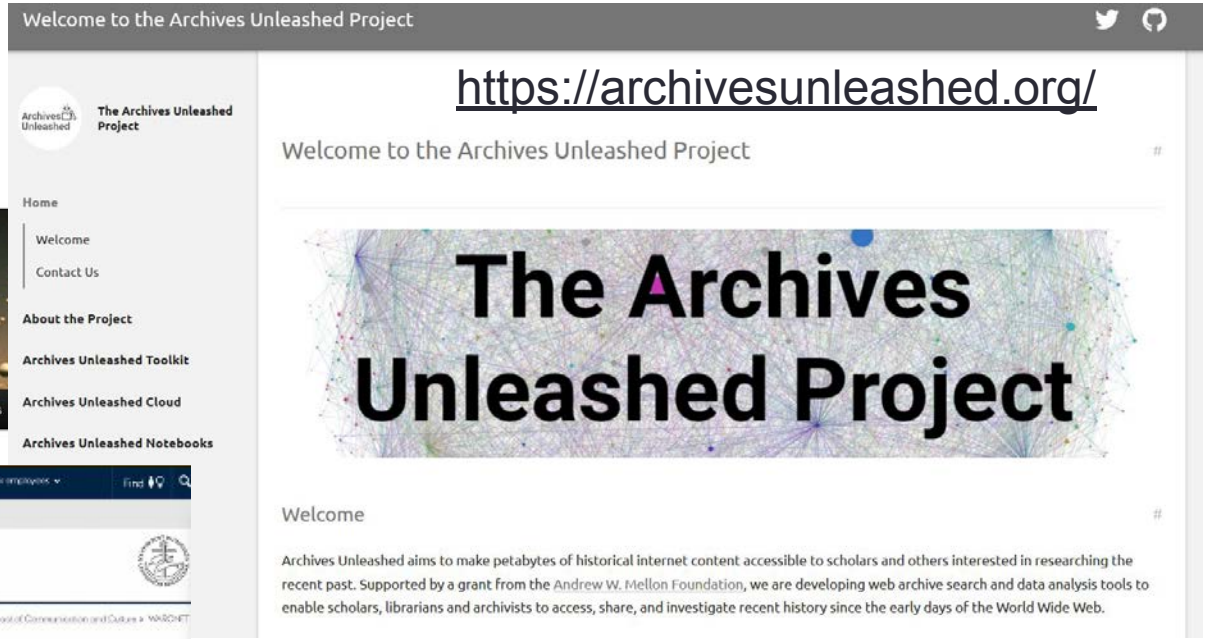
Parcours entre Histoire et SIC :
- Une première expérimentation en 2015 (thèse). Dépôt légal du Web français (BnF+INA), Internet Archive
- Réflexion méthodologique et épistémologique (2021) *Website Story, Histoire, mémoires et archives du Web*

Des travaux ancrés dans le domaine de recherche sur et avec les archives du web : international, interdisciplinaire, pluricatégoriel.

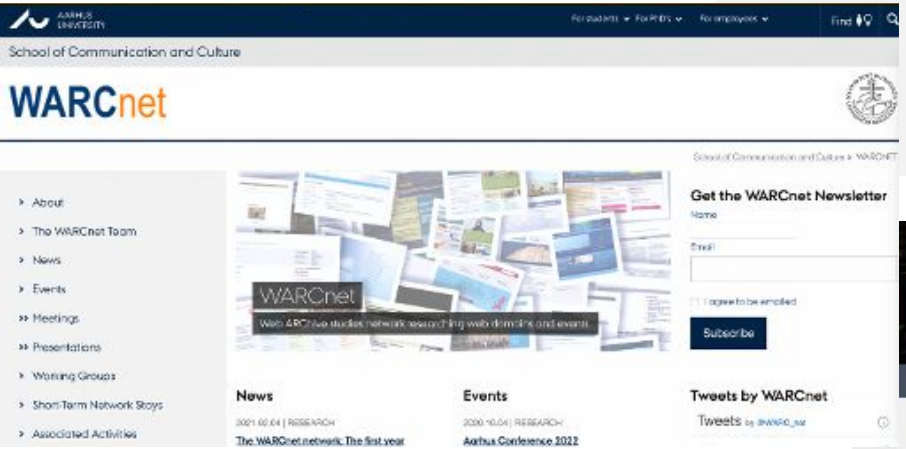
Principaux enjeux pour les SHS :
- Etat de l'art : besoin de développer des connaissances sur l'usage de l'archivage du web et ses implications épistémologiques
- Source / terrain / documents : accès aux documents nativement numériques et stabilisation des corpus issus du web vivant et des médias sociaux (citabilité)
- Web archivé, terrain de jeu des humanités numériques => méthodes computationnelles
- Mémoire et patrimoine nativement numériques : collectes, formats etc.

De nouveaux champs de recherche, de nouveaux réseaux et projets

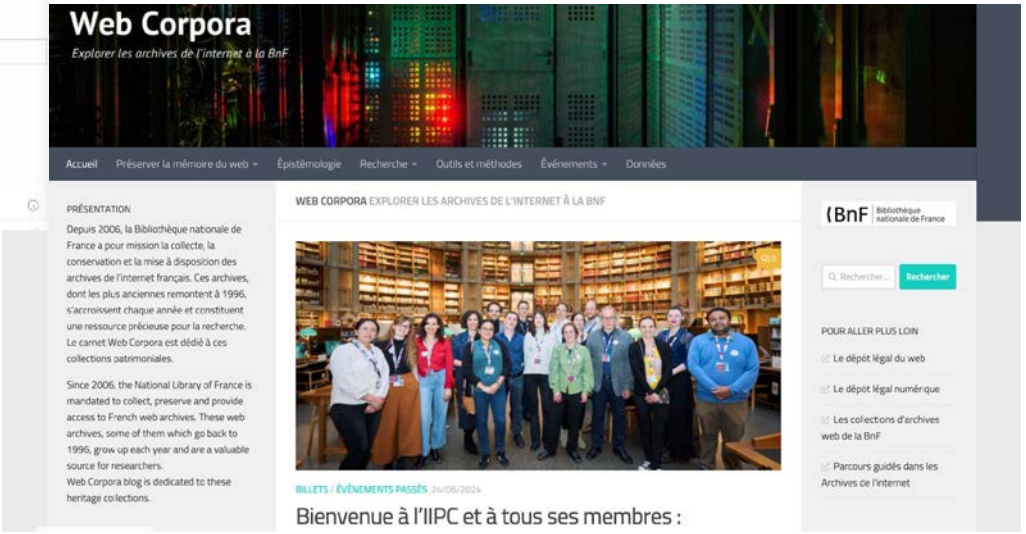
https://web90.hypotheses.org/



https://archivesunleashed.org/



https://cc.au.dk/en/warcnet (2021-2023)



https://webcorpora.hypotheses.org/

RESAW, a Research Infrastructure for the Study of Archived Web Materials

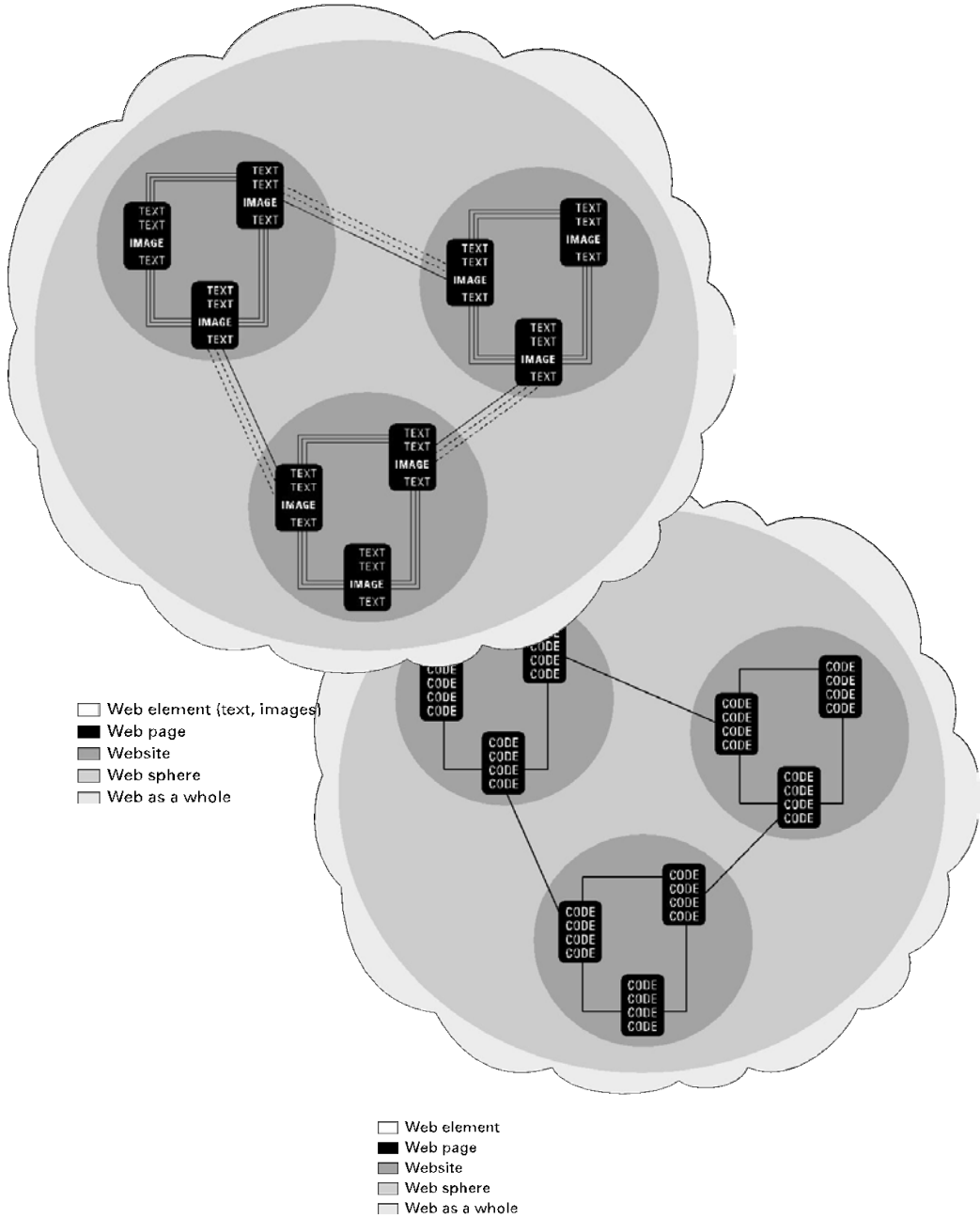
https://resaw2023.sciencesconf.org/

Given recent global crises, the imperative to preserve and analyze online content has never been more vital to enhancing our comprehension of contemporary changes. This book, the outcome of the 5th international RESAW conference that convened experts from fifty disciplines across seventeen countries in Marseille in June 2023, tackles the multifaceted challenges of web archiving. It underscores the dual roles of web archiving, as cultural heritage and as essential source material for researchers delving into contemporary events and the evolution of digital culture. Through twenty chapters, it explores the development of web archiving and examines how technical, cultural, geopolitical, societal, and environmental shifts impact its conception, study, and dissemination.

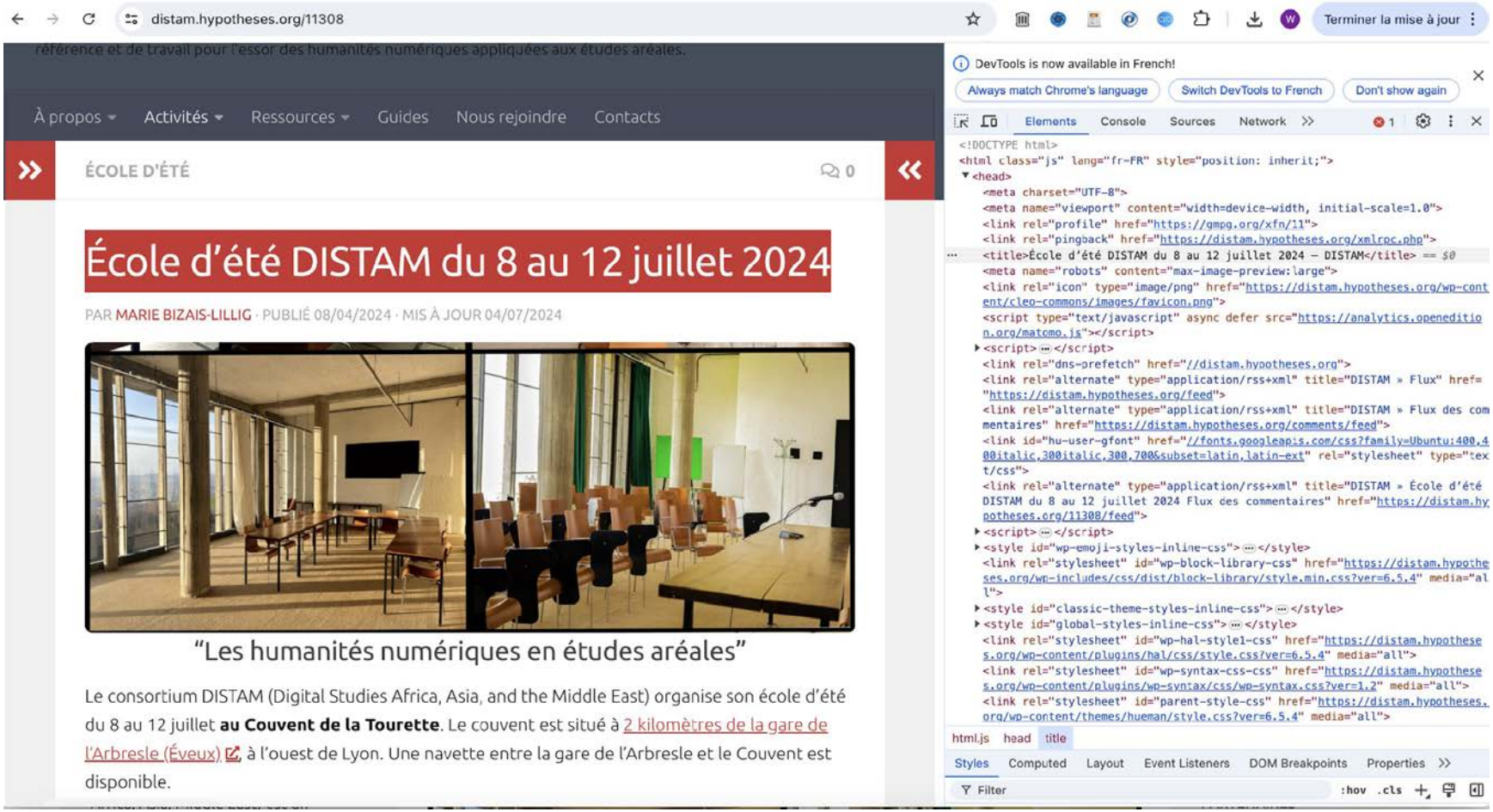
https://books.fupress.com/catalogue/exploring-the-archived-web-during-a-highly-transformative-age/14127

Les 5 couches analytiques du Web :

- Les éléments du Web
- Les pages web
- Les sites Web
- Les sphères du Web
- Le Web dans sa dimension globale

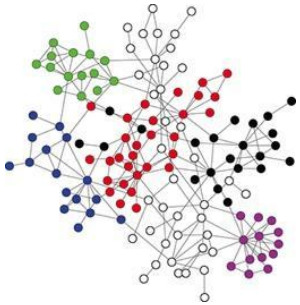


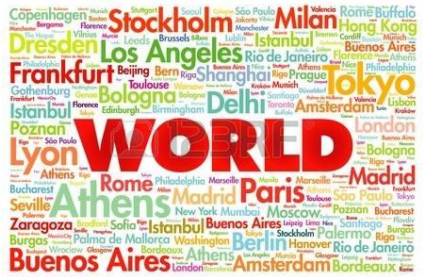
Deux couches de texte sur le Web : texte visible et texte invisible html



Ce qui est archivé est different de ce que l'on voit en ligne

Une archive du Web = une masse de fichiers interconnectés





Liens

Entités nommées

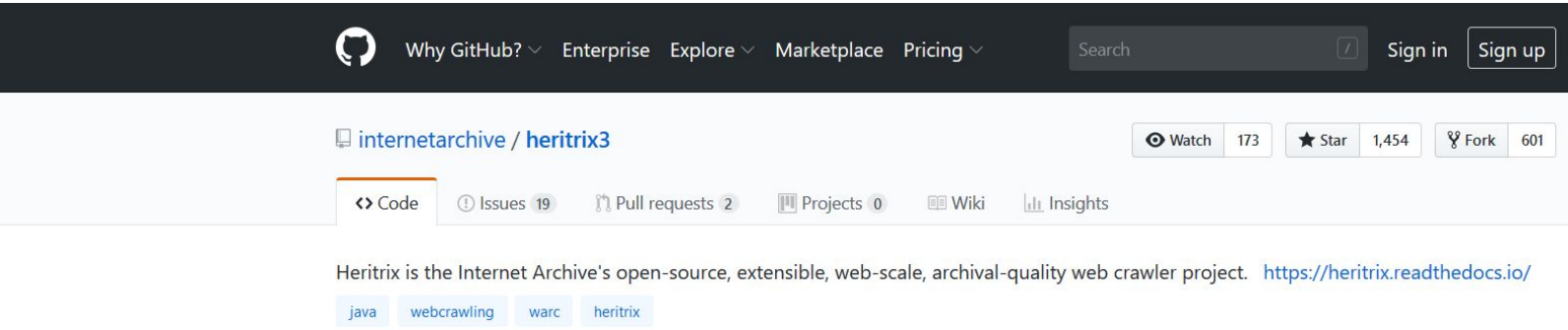
Mots

?

Web vivant

collecte

html + fichiers



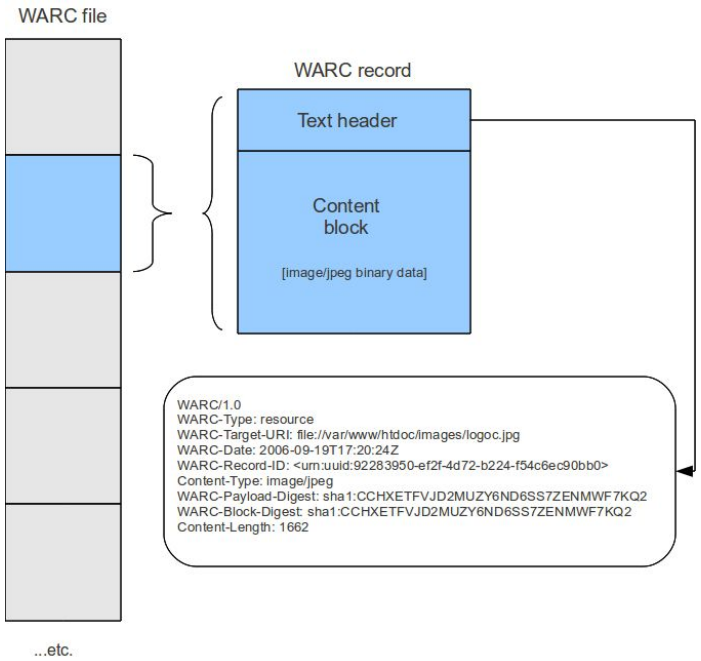
Heritrix is the Internet Archive's open-source, extensible, web-scale, archival-quality web crawler project

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

Un format standard et normalisé pour les archives du Web

- WARC file format = Web ARChive archive format
- WARC format international depuis 2009 (ISO 28500:2017)



https://en.wikipedia.org/wiki/Web_ARChive
https://wiki.archiveteam.org/index.php/The_WARC_Ecosystem
https://wiki.archivematica.org/File:WARCdiagram.png

Consulter le web archivé, un nouveau rapport à l'archive

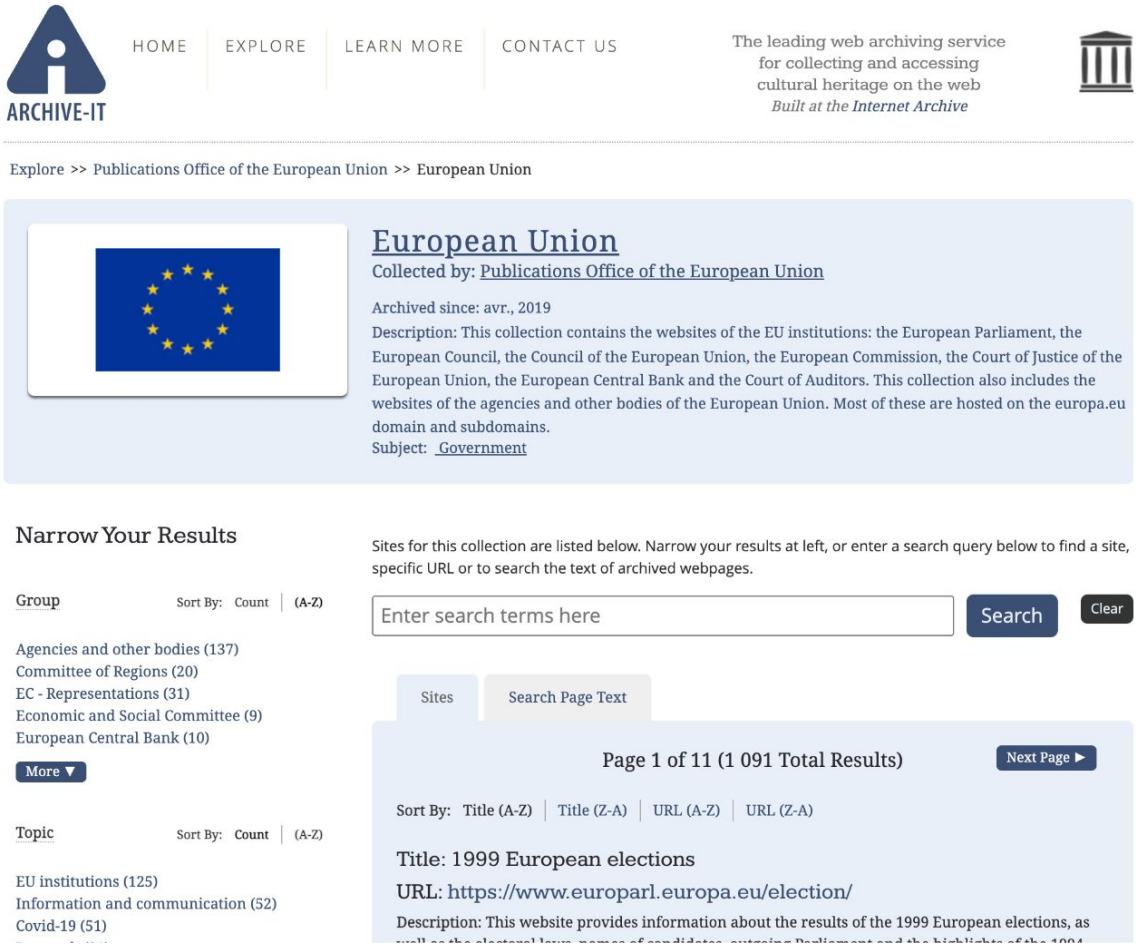
Un processus de *remediation* (Niels Brügger, 2005)

CNHI,
Page d'accueil,
Internet Archive,
05/10/2007

Webrecorder

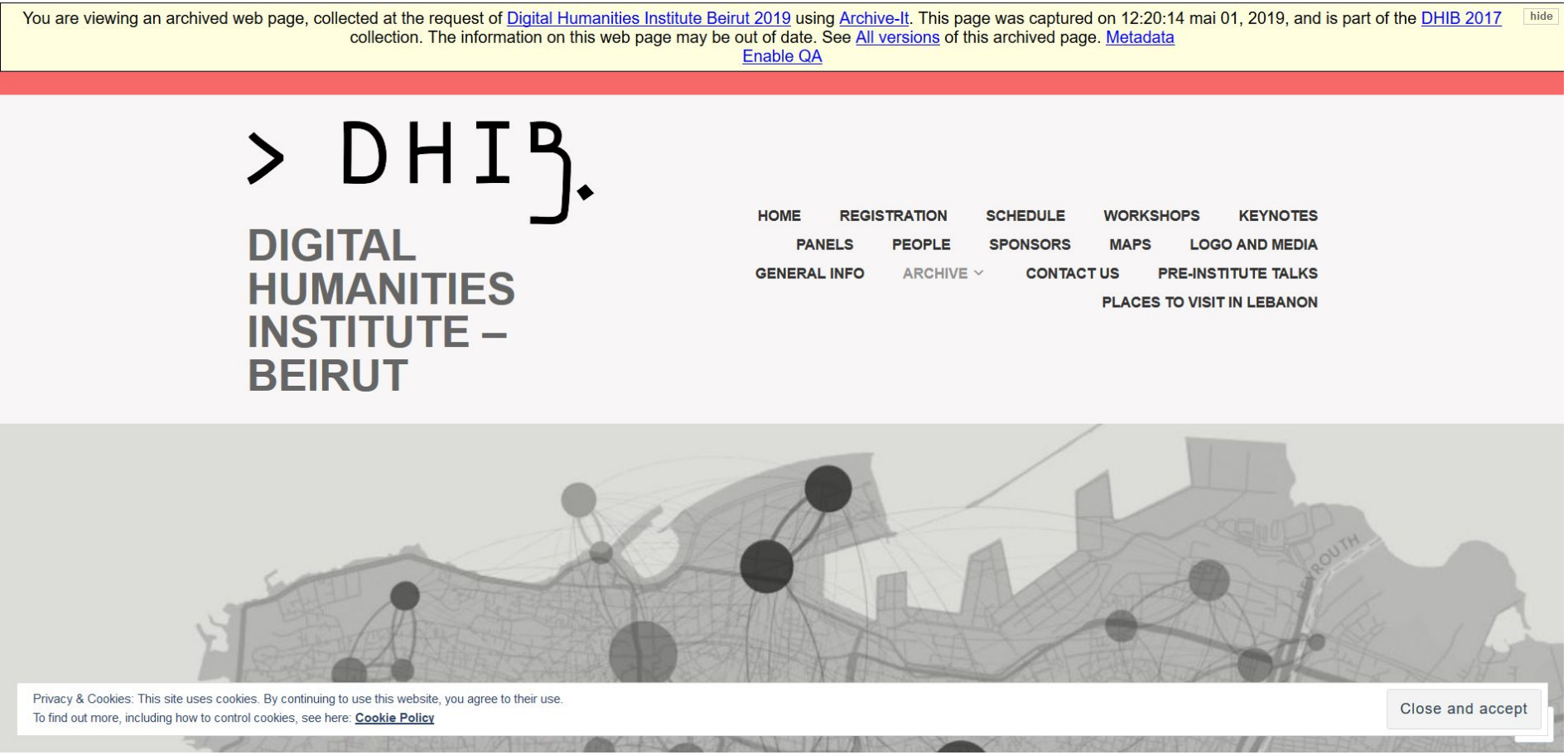
https://webrecorder.net/

Archive-It : des collections publiques pour les projets



https://archive-it.org/collections/12090

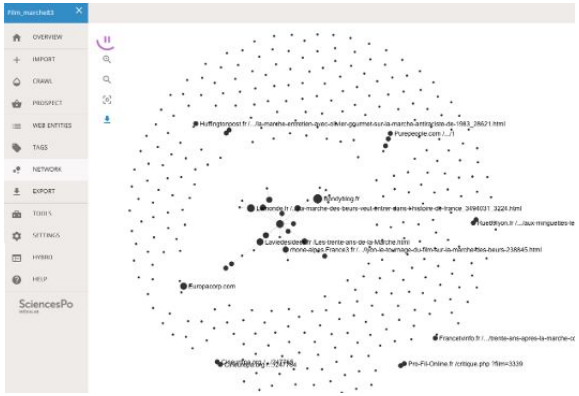
Archive-It: exemple d'une page archivée avec Archive-it

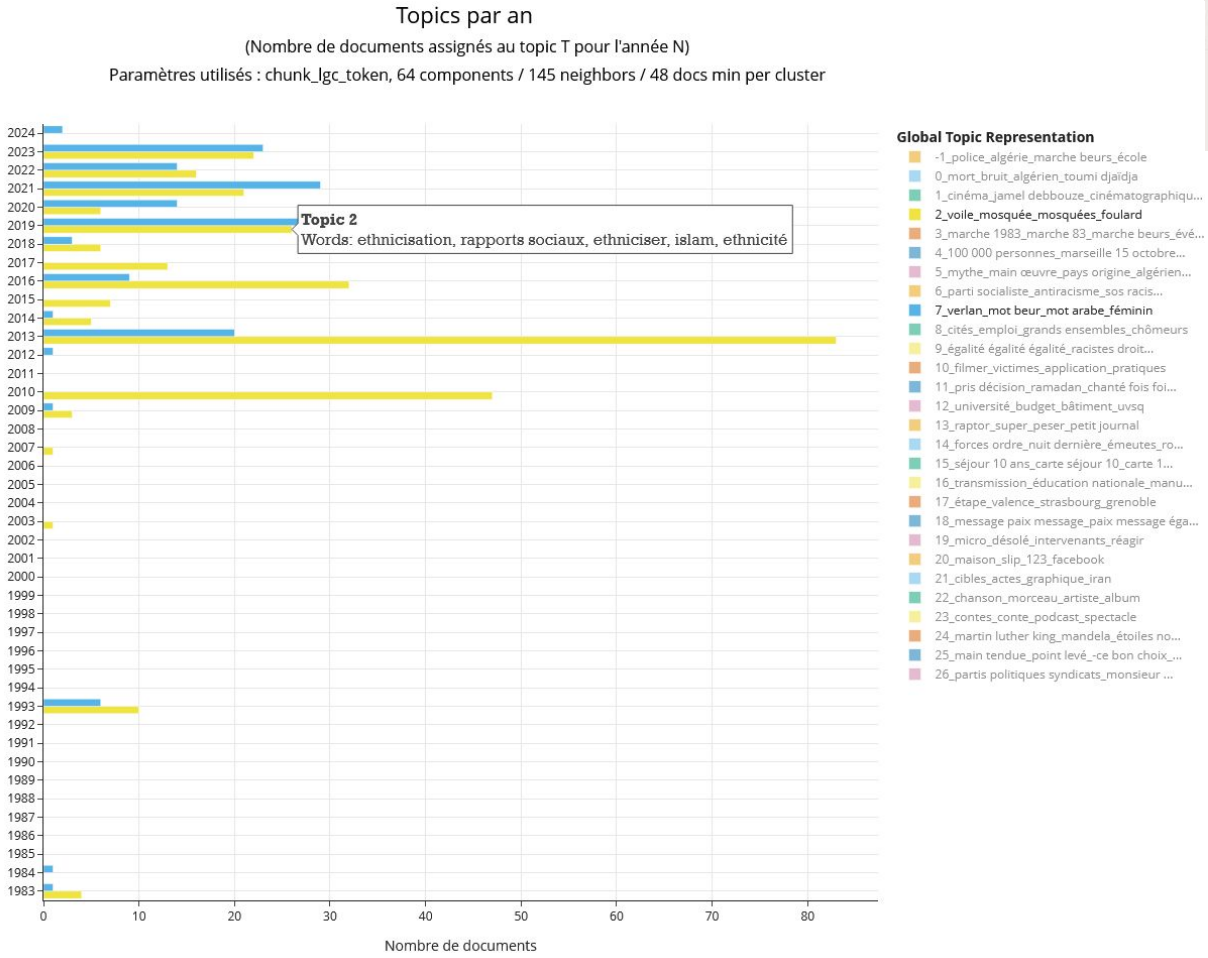


https://wayback.archive-it.org/12102/20190501122014/https://dhibeirut.wordpress.com/archive/dhi-b-2017/

L’archivage du Web, terrain de jeu des humanités numériques

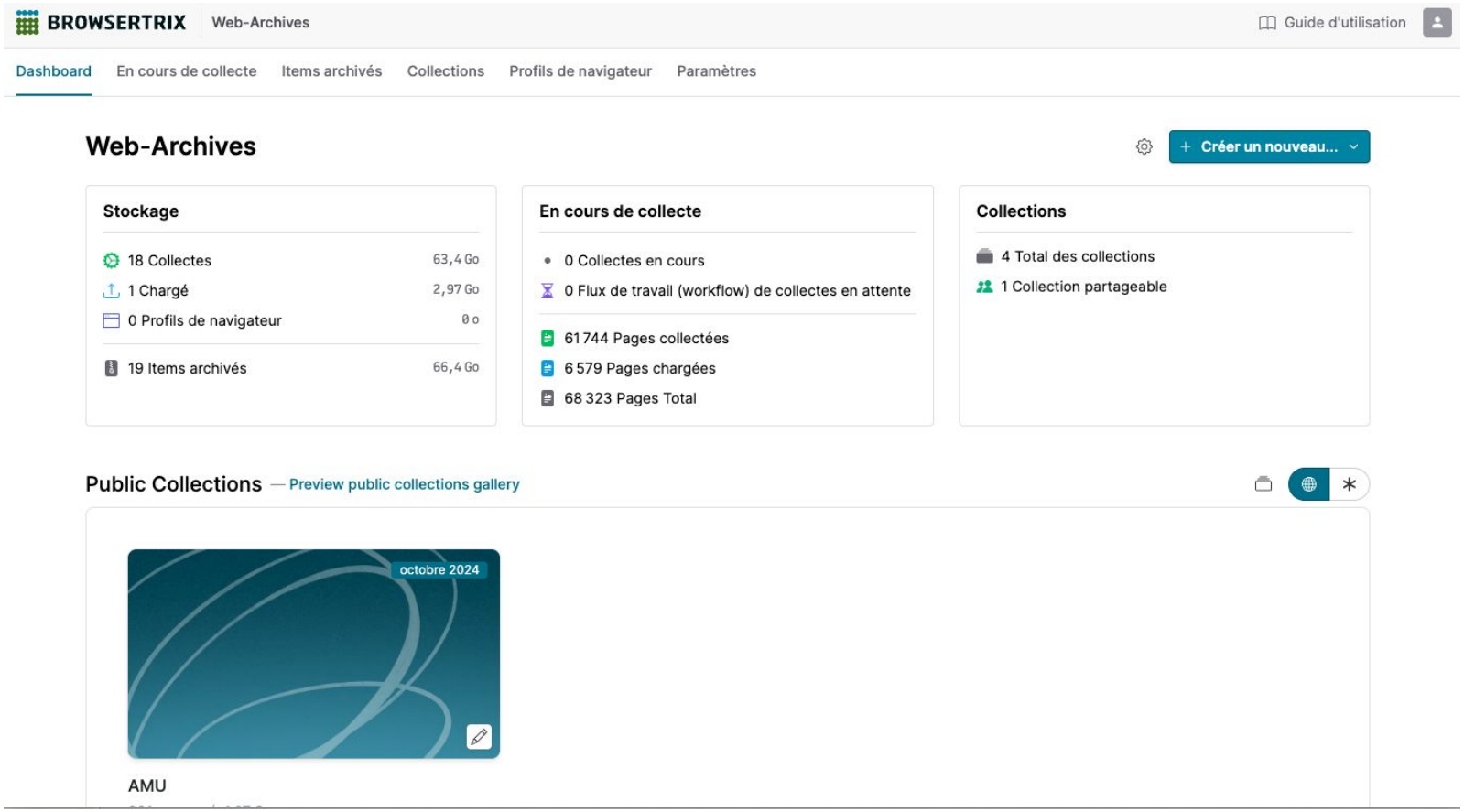
Quarante ans de médiatisation de la Marche de 1983 à la TV et en ligne

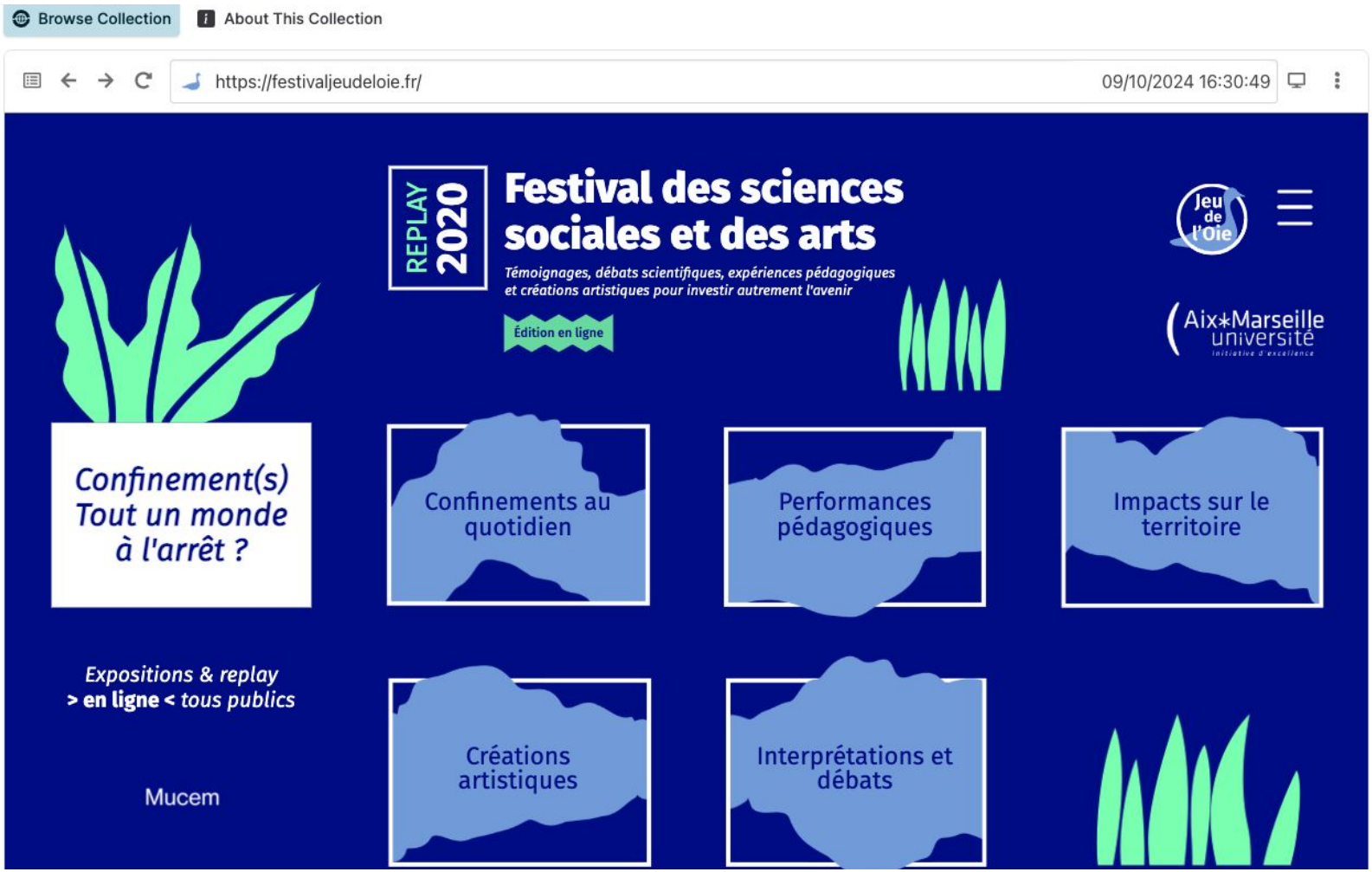




Projet Polyvocal Interpretation of a Contested Colonial Heritage (2022-2024). Collaboration avec Le Lab (INA)

Web Archiving Aix Marseille (POC)





https://archives-poc.cedre.univ-amu.fr/explore/web-archives/collections/amu#view=pages

Partie 5 : Vers une exploration collective du Web archivé

- Un domaine structuré au niveau international : avec des archivistes et des communautés académiques, une ingénierie spécialisée
- Un domaine en évolution constante : plateformisation, APIcalyps, défis de l'IA générative
- Une réflexion interdisciplinaire et pluri-catégorielle : séminaire avec des enjeux épistémologiques, patrimoniaux, digitaux, politiques de conservation

=> Collaboration WebLab et Humathèque

Séminaire 2025/2026 WebLab-Humathèque

Le Web et les archives du Web pour la recherche en SHS : savoirs, méthodes et outils pour la collecte, l'analyse et la pérennisation de corpus en ligne

Un site Web : https://pba.mmsh.fr/?page_id=1465

Une liste de diffusion : weblab@services.cnrs.fr

Une formation doctorale sur ADUM

Programme en cours de construction (7 séances)

Jeudi 27 nov. 14h-16h :

Institutions d'archivage et recherche en SHS : Le dépôt légal du web français (BnF et INA)

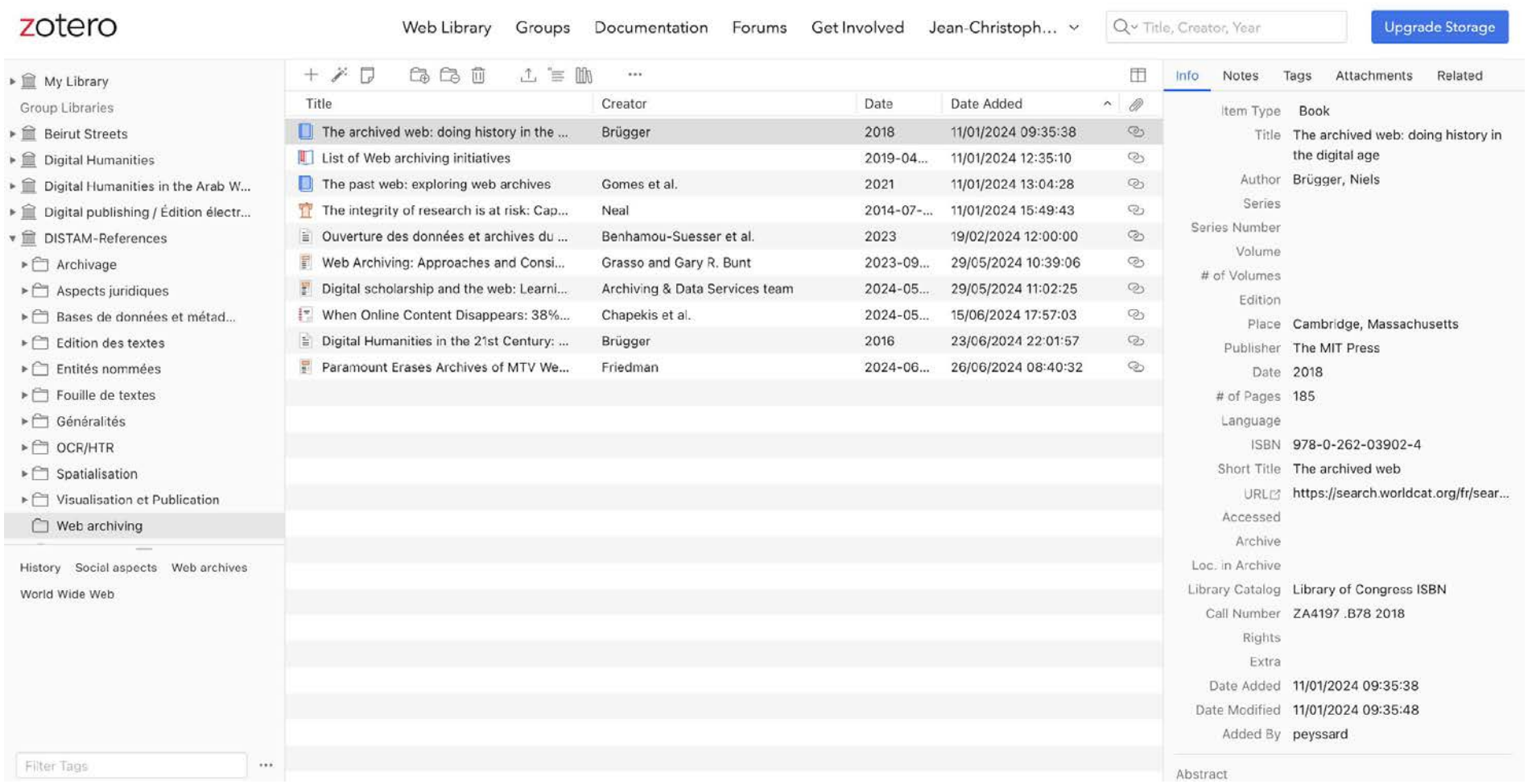
Mercredi 3 décembre 14/16h 2025

Stéphane Pouyllau (Huma-Num)

Thèmes à venir :

- Actualité du domaine
- Aspects éthiques et juridiques
- Enjeux méthodologiques : Hyphe, WAAM
- Analyse des usages

Bibliographie



https://www.zotero.org/groups/2908451/distam-references/collections/3ZYJKDFQ