L’archivage du web français à l’ère de l’IA

Séminaire du weblab du 27 novembre 2025

Géraldine Camile
(BnF DataLab, Département de la découverte des collections et de l’accompagnement à la recherche)

Plan

- Panorama des collections des archives du web de la BnF
- Modalités de constitution des collections des archives du web à la BnF
- Les outils de consultation et de recherche de la BnF
- Identifier les services offerts par la BnF
- Impact de l'IA sur les archives du web, de la sélection à la recherche

(BnF

Panorama des collection à la BnF

(BnF

Fear of a digital dark age
La peur d'un web amnésique

Brewster Kahle lance la fondation Internet Archive en1996

Les12 membres fondateurs d'IIPC signe l'acte de naissance du consortium en 2003 à la BnF.

https://web.archive.org/
https://netpreserve.org/

Des archives du web qui remontent à 1996, avec rachat de collections à Internet Archive et expérimentations de collecte

La mission de dépôt légal de l'internet

La loi DADVSI (2006) et son décret d'application (2011), intégrés au Code du patrimoine, instituent le « dépôt légal du web ».

Dans la continuité des missions des institutions patrimoniales : « Rassembler, conserver le patrimoine culturel français » et y donner accès.

Deux établissements dépositaires : INA (sites de radio et télévision), BnF (le reste du web français).

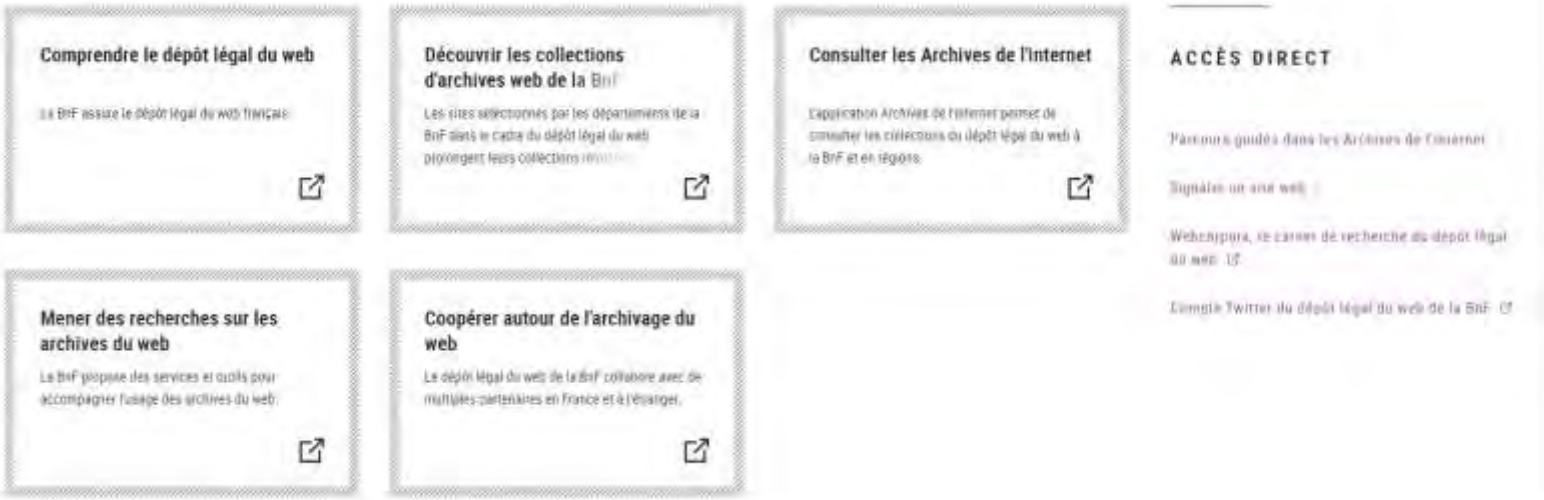
Les contenus restent soumis au droit d'auteur et les données collectées sont des données patrimoniales.

Des collections qui illustrent ce qu'est le web, dans toute sa diversité de formes et de contenus, à une époque donnée

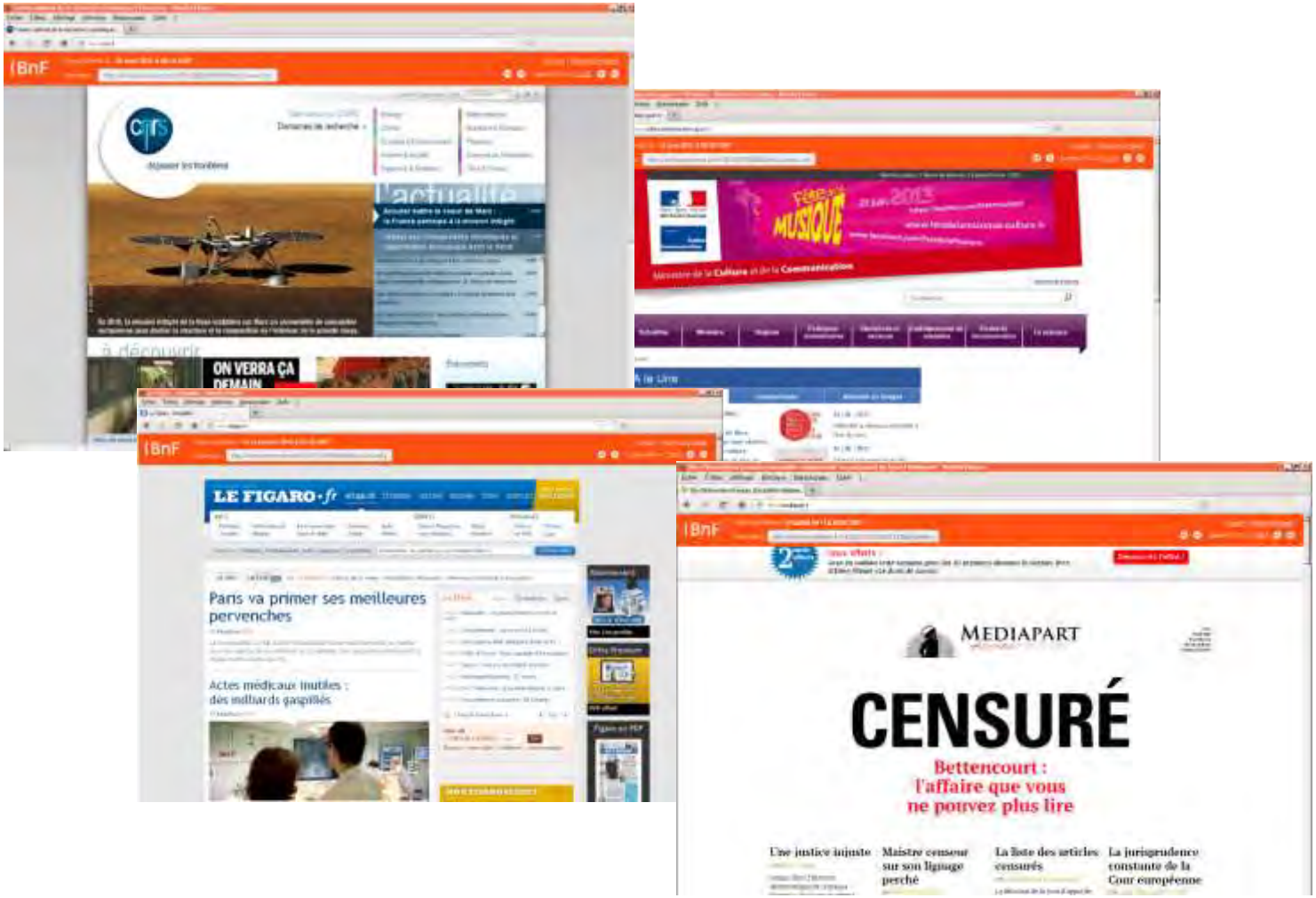



(BnF

https://www.bnf.fr/fr/depot-legal-du-web

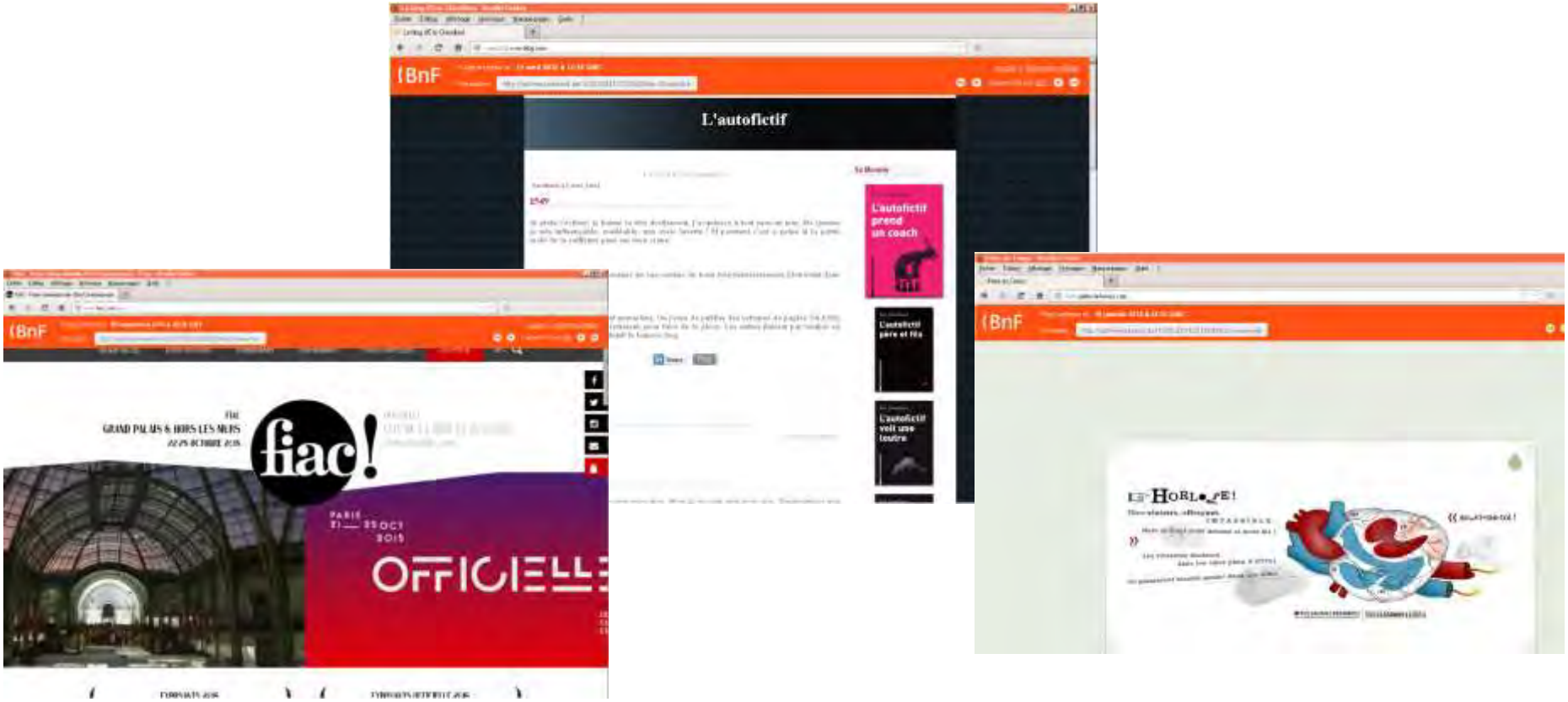


Des sites officiels, de référence, de médias

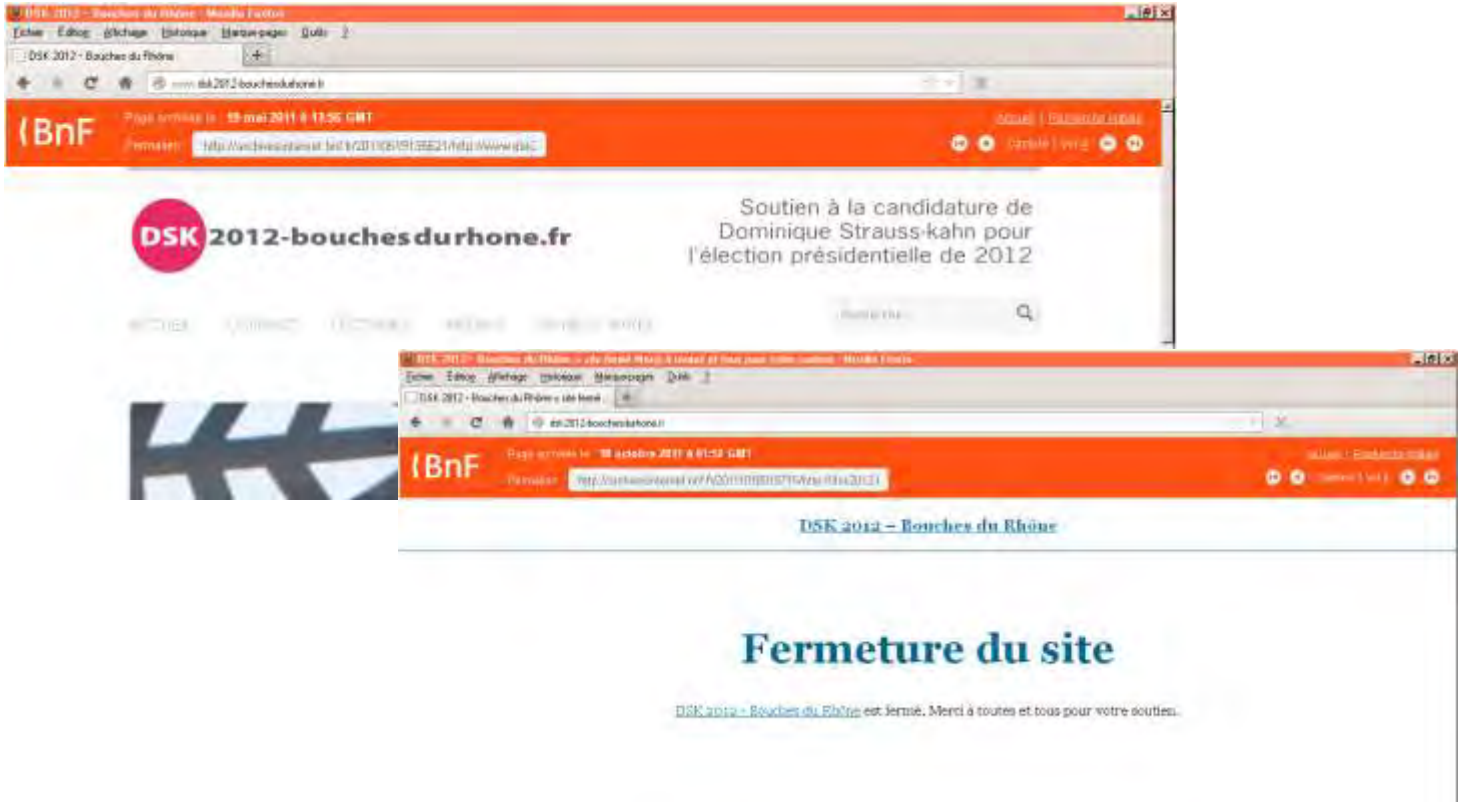


Des journaux en PDF, des livres numériques en EPUB

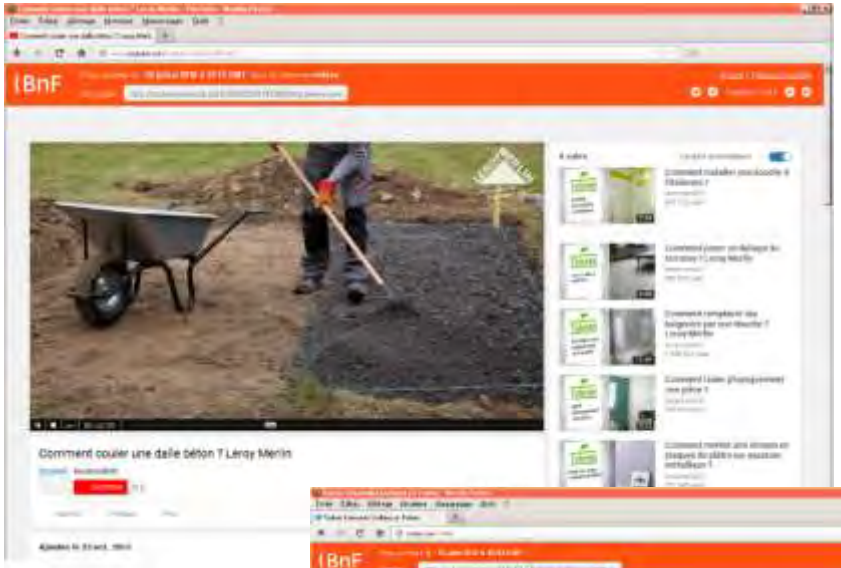
Des sites sur l’art, la littérature en ligne



Des sites liés à un événement



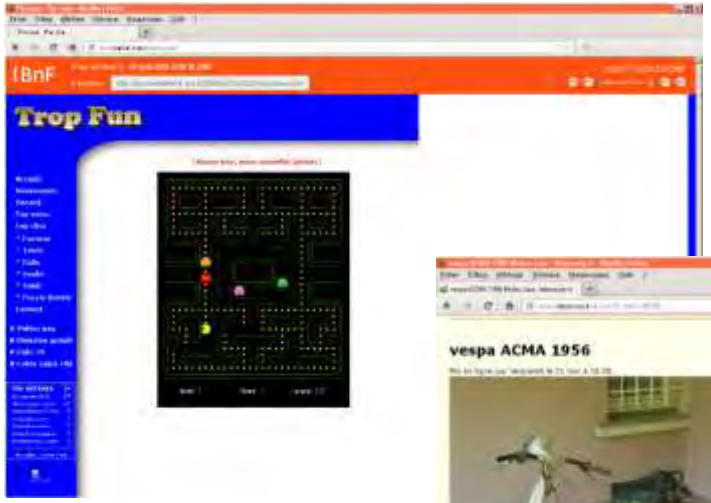
Réseaux sociaux, vidéo et audio



LA BNF ARCHIVE LES SKYBLOGS

Dans le cadre de sa mission de dépôt légal des sites web, la BnF a lancé une vaste opération de collecte de la plateforme Skyblog

… témoins de la diversité du web



des contenus qui évoluent dans le temps

http://solidarites-sante.gouv.fr/

Février 2020

Mai 2020

Qu’est-ce qu’une archive web ?

Niels Brügger

- La notion de ***reborn-digital material*** a été developpée par Niels Brügger pour qualifier l’archive web.
- « Le web archivé est fondamentalement différent de tous les autres types de médias existants, y compris le papier, la radio, la télévision et même le web vivant ».
- Il n’y a pas de copie de la source comme dans un acte de photographie, mais une **transformation** en archive.
- L’archive est le résultat des **modalités techniques** de constitution des collections et de la configuration des outils d’accès.

=> Comprendre l’archive, c’est comprendre les collectes.

Brügger, N. (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

(BnF

Modalités de constitution des collection à la BnF

(BnF

Les collections du Dépôt légal de l'internet

- Des captures de sites web, **restitués** à partir des fichiers collectés tels qu'ils étaient au moment des **crawls**
- Dans la continuité des autres collections de la Bibliothèque
- Dépôt légal : tous les types de site et toutes les formes de publication
- On retrouve des équivalents des publications sur support, et aussi de nouvelles formes spécifiques au web

(BnF

Deux modalités principales et complémentaires de constitution des collections

Collecte large

- Des échantillons de tous les sites du domaine français, une fois par an (listes fournies par l'AFNIC, OVH, etc.)

Collectes ciblées

- Dans la continuité des collections de la BnF
- Importance croissante des collectes projet : élections, COVID-19, actualité, enjeux environnementaux, intelligence artificielle, etc.
- Sélection qui obéit aux principes fondamentaux du dépôt legal : pas de jugement scientifique, moral, ni esthétique

BnF

BnF Collecte du web

Voir les collectes

FICHE DE SITE 97 357

Collecte : Mouvements sociaux (MOUV)
Responsable : Sophie GEBEIL
Création : Sophie GEBEIL, le 29/05/2015
Mise à jour : Sarah TOURNERIE, le 24/11/2023

Paramètres de collecte

État : Inactif
URL de départ : http://leschibanis.free.fr
URL supplémentaires : .
Type de collecte : ciblée
Budget : petit **Fréquence :** 1 fois par an **Profondeur :** hôte

Description du site

Thème : mémoires et patrimoines de l'immigration
Mots clés : chibanis / immigrés âgés / travailleurs étrangers maghrébins / photographie
Notes de contenu : Pb Flash (SG)
Site constitué de photographies sans commentaires.
Notes techniques : DLN - 20170628 - mise en collecte d'urgence pour parcours guidé (transfert MOUV => DDL) ; rajout URL supplémentaire (U
profondeur domaine.
- après collecte remettre en MOUV, responsable Sophie Gebeil, petit - 1 fois par an - hôte ; thème mémoires maghrébines
DLN - 20170711 - Transfert de DDL (demandes externes) à MOUV
DLN - 20190529 - suppression de l'URL http://www.chibanis.com (remplacée par l'ancienne URL supplémentaire)
PHS - 20221006 - 35 photos actuellement. Voir si le site évolue, et si des photos sont ajoutées d'ici un an, sinon l'inactiver.
PHS - 20232411 - site n'existe plus. Désactivé le 24/11/2023 ST

Outil de curation

- Vue documentaire par collecte
- Saisie des paramètres de collecte
- Thème, mot clé, description
- Exemple :

leschibanis.free.fr
fiche créée par Sophie Gebeil,
en 2015, inactif depuis 2023

(BnF

Les collectes ciblées :

quels principes de sélection ?

A défaut d'exhaustivité, un objectif de représentativité et d'échantillonnage

Suspension du jugement esthétique, moral, scientifique : culture populaire comme culture savante, *mainstream* ou longue traîne, tout est collectable

Capturer les actions, les savoirs, mais aussi les idées, les représentations qui circulent sur le web sur un sujet donné : rendre compte de la diversité du web, media et espace d'échange

Le robot de collecte Heritrix

Le robot part d'une liste d'URL de départ et exploite les liens dans les codes sources des sites.

Des paramètres de collecte sont ajoutés. Certains paramètres peuvent être modifiés en cours de collecte.

Le robot copie les fichiers dans des fichiers de sauvegarde WARC, format standard pour la conservation.

Ensemble à collecter → Jobs (urls de départ) →queue → Fichiers WARC

Mensuelle petite

Job 1
aqis.fr
anthropoblog.fr

Job 2
Inserm.fr
kezacovid.fr
le-bars.net/yoann/fr/
lesprit...

Job 3
topsante.com

ULRs de départ

http://presse.inserm.fr

http://www.kezacovid19.fr

WEB

Queue de collecte

https://kezacovid19.wordpress.com/2020/06/11/12-mutations/

https://kezacovid19.wordpress.com/2020/05/21/11-differents-tests-de-detection/

https://kezacovid19.wordpress.com/2020/05/07/10-virus-echappe-dun-laboratoire/

https://kezacovid19.wordpress.com/2020/05/02/08-echelle-logarithmique/

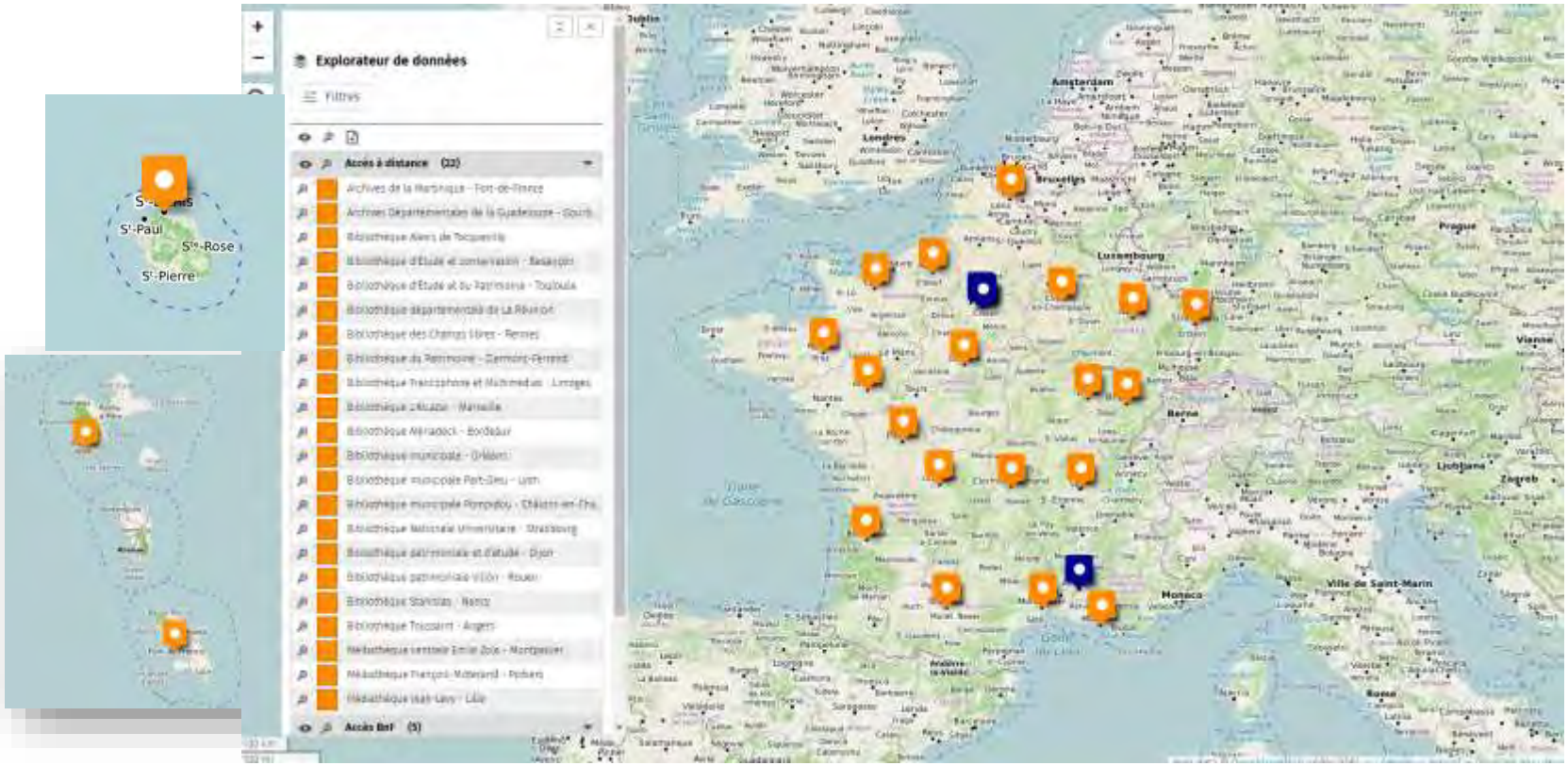
https://kezacovid19.wordpress.com/2020/05/02/07-sars-cov-2-et-notion-de-virus/

https://kezacovid19.wordpress.com/2020/05/02/06-origine-du-virus/

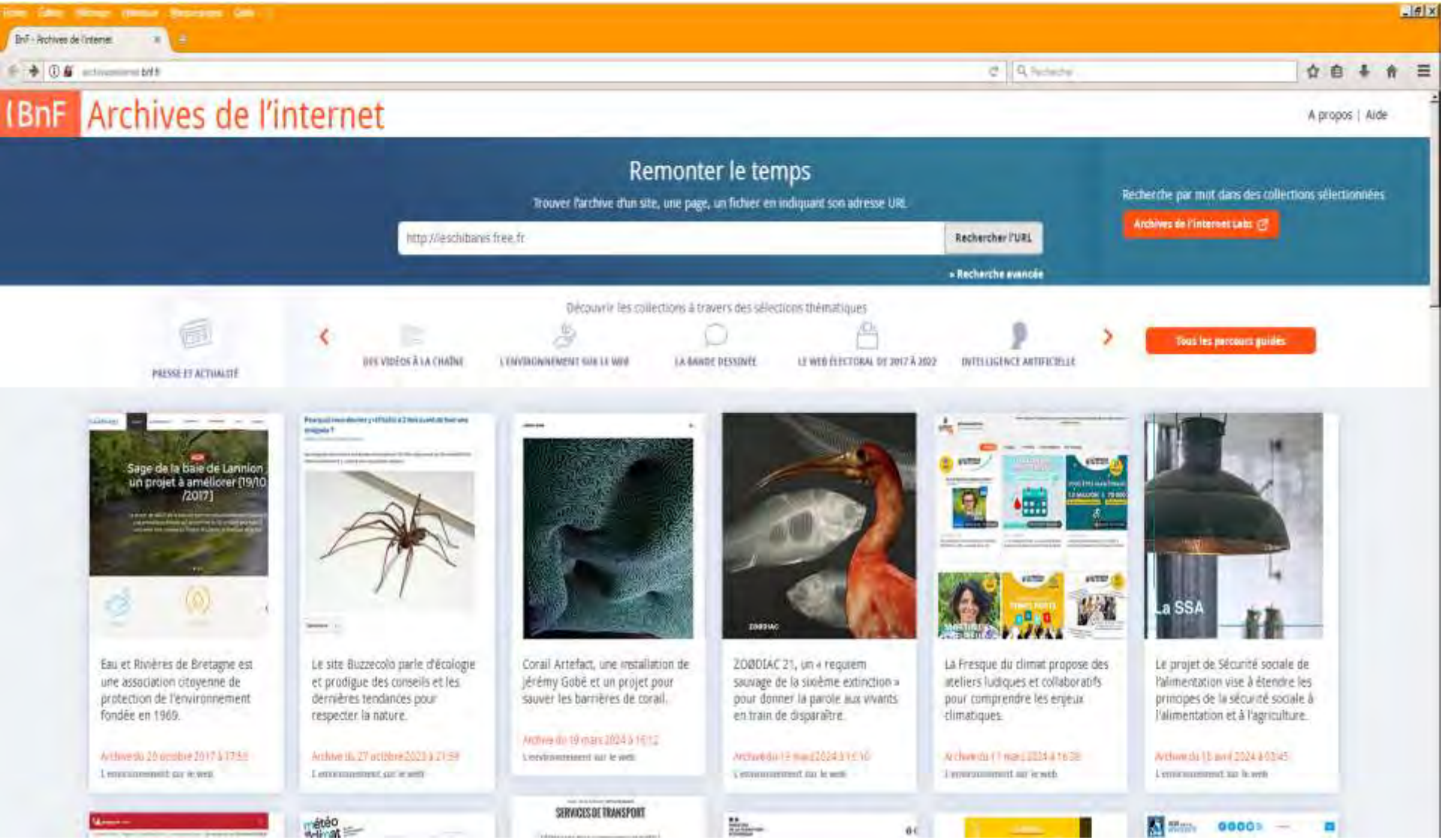
Les outils de consultation

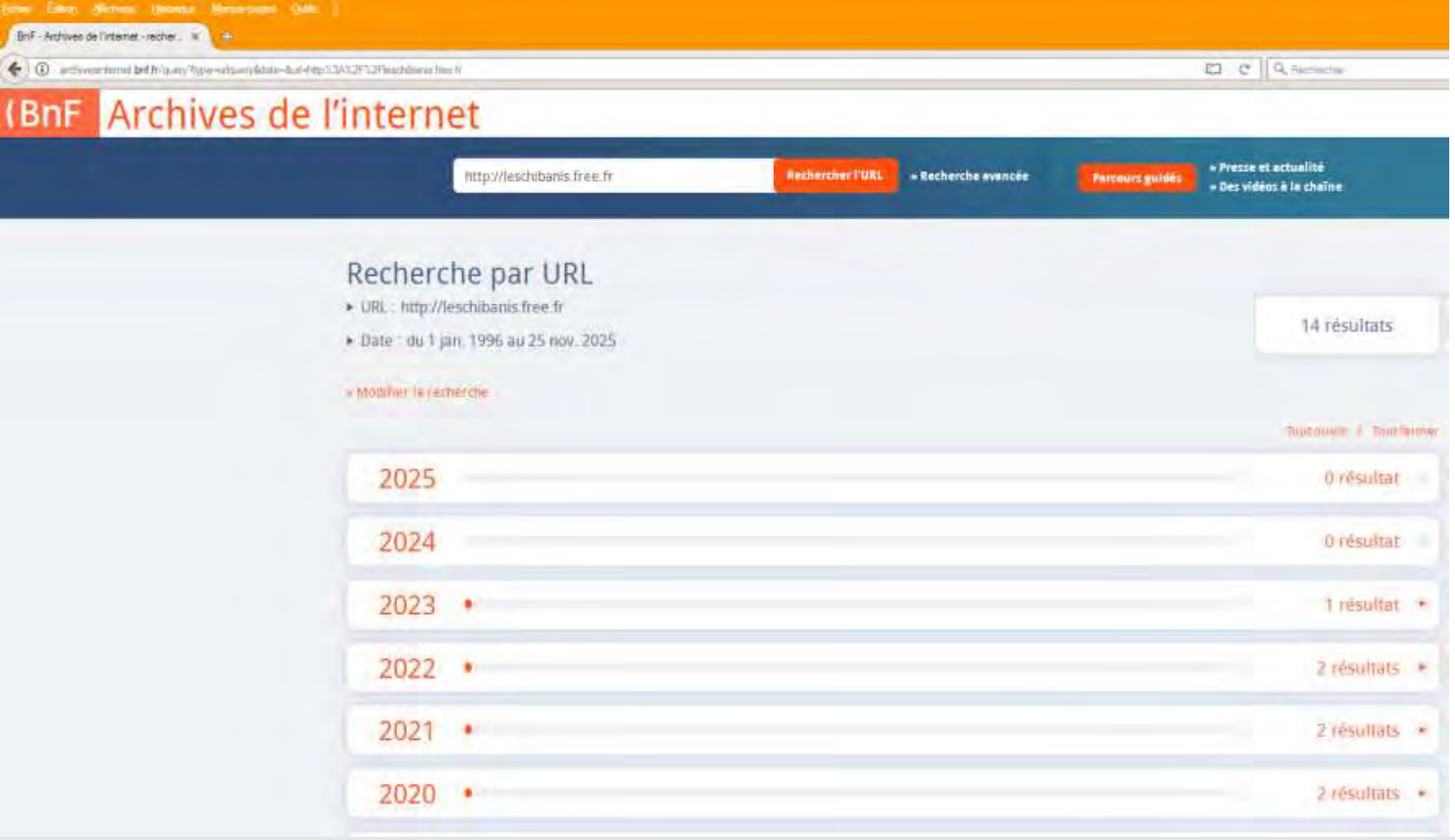
Accès à la BnF et dans 22 bibliothèques partenaires

réseau des bibliothèques de dépôt légal imprimeur (BDLI)



Les Archives de l’internet, une wayback machine





Rien n'indique :

- les origines techniques ou documentaires de la collecte
- ni la qualité de l'archive

Par exemple : Les 14 résultats présents ne sont pas pertinents car la vraie page est : http://leschibanis.free.fr/Menu1.html

(BnF

Petite philologie du web

- *Des sauts temporels lors de la navigation*
- *Une « page web » recomposée à partir d'éléments crawlés à différents moments*
- *Des pages inventées par l'extracteur Javascript du robot*
- *Des sites collectés à des fréquences différentes*
- *Des sites collectés profondément et d'autres superficiellement*
- *Des blocages du robot, des codes réponses http*
- *Des contenus « embed » qui laissent des traces fantômes*
- *L'absence de statistiques d'usage…*

Les Archives de l'internet restituent ce que le robot a collecté….

Citer une « Archive de l'internet »

- Permalien
  - Même construction pour Internet Archives que pour la Wayback Machine d'Internet Archives

http://archivesinternet.bnf.fr/20170701090810/http://leschibanis.free.fr/Menu1.html

Parcours guidés

Parcours sur un thème

- des sélections réalisées et des textes rédigés par des bibliothécaires et des partenaires institutionnels ou des chercheurs
- Un permalien est associé à chaque source décrite

Travail d'éditorialisation important.

Publication en ligne (PDF).

Forme d'entrée dans les collections.

Un parcours guidé sur l'intelligence artificielle

Application
« Archives de l'internet Labs »

Recherche plein texte

- Recherche simple
- Recherche avancée
- Recherche experte
- Recherche n-gram

Copie et utilisation des archives

- Possibilité de consulter des vidéos et d'écouter des enregistrements audio
  - prise casque sur les postes publics + casques audio (disponibles en banque de salle)
- Autorisation de copier/coller des extraits de textes à des fins de (courtes) citations
- Toute autre réutilisation des contenus nécessite l'autorisation des ayants droit
- Impossibilité de télécharger des contenus
- **Des utilisations plus avancées sont proposées au BnF DataLab**

Les services aux chercheurs

Les dispositifs d'accueil

- Chercheuses et chercheurs associés

Pour un projet de recherche en lien avec les collections de la BnF

- Accompagnements par le BnF DataLab

Pour un projet sur les collections numériques (numérisées ou nativement numériques) pouvant avoir une approche expérimentale

  - En répondant à un appel à projet DataLab (une fois par an)
  - En sollicitant le BnF DataLab

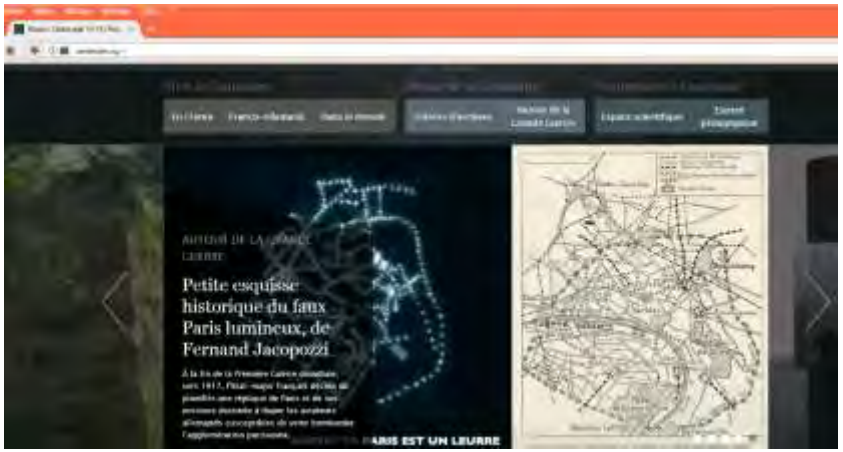
Le BnF DataLab travaille en partenariat avec HumaNum

Archives du web et recherche - collecte

Retrouver des sites disparus

- Approche archivistique « classique »

Exemple : site commémoratif centenaire.org



S'assurer de l'archivage des contenus étudiés

- Collecte dans le cadre de partenariats

Exemple : centre du féminisme d'Angers

Coproduire une collecte

- Dans le cadre d'un projet de recherche

Exemple : collecte sur les littératures francophones

Offre de services

- **Parcours guidé (médiation)**
- **Participation aux collectes courantes**
- **Collecte Corpus de recherche**

Archives du web et recherche - data

Indexer des collections massives

- Indexation d'une collection d'intérêt
- Indexation dans le cadre d'un projet de recherche

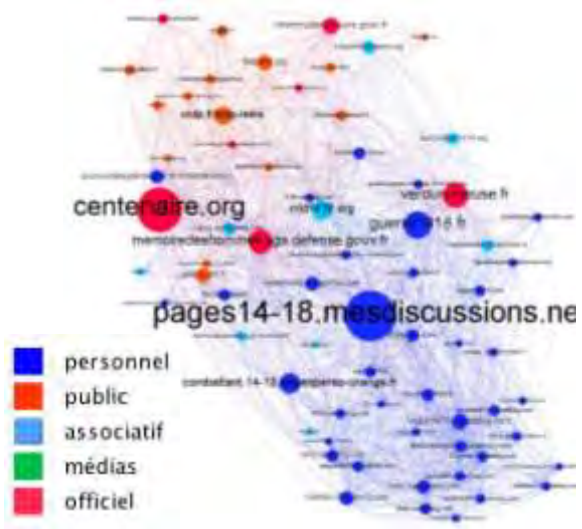
Déterminer quels fichiers doivent être indexés

Exemple : collection Attentats 2015

Fouille de données et création de corpus

- Requêtes et extractions à partir des :
  - métadonnées (URL, liens, types MIME)
  - index (recherche textuelle)
- mise en place d'un process et accueil d'un IGE/IGR

Exemple : cartographie du web de la Grande Guerre



Offre de services

- **Indexation**
- **Extraction de données et métadonnées**
- **Aide à la fouille de données**

Portail dédié du BnF DataLab

Conseils méthodologiques aux chercheurs

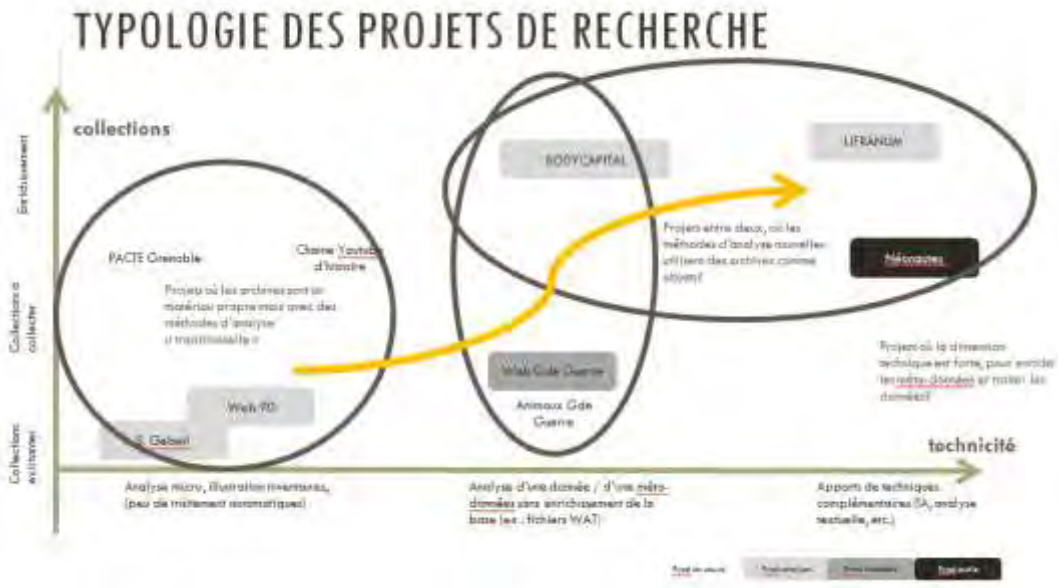
Les freins à l'usage des archives du web

- Une source difficile à appréhender et à citer
- Une consultation limitée aux emprises de la BnF et des bibliothèques partenaires
- Une faible proportion de collections indexées en plein texte

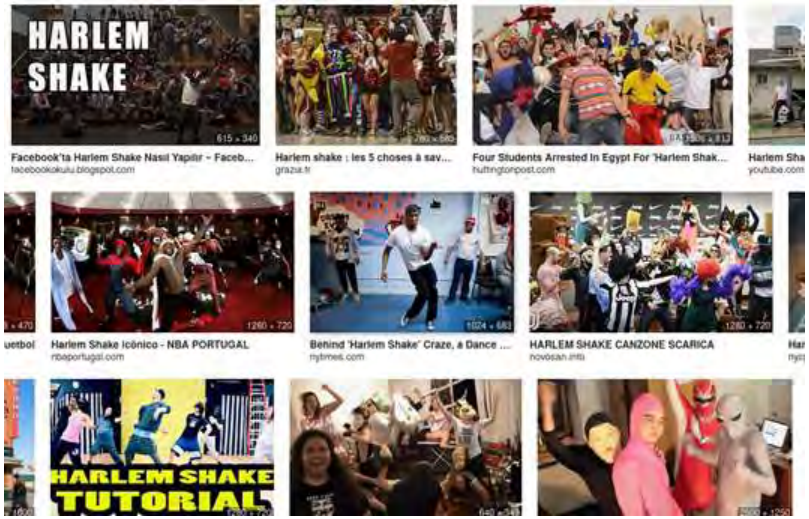
Métadonnées et données dérivées

❑ Créer des corpus par étape :

- des métadonnées pour appréhender l'archive
- mixer les approches quantitatives et qualitatives
- créer et exporter des données dérivées
- partager les corpus
- **abaisser la marche technique**
- **tester et expérimenter de nouveaux outils**
- **faciliter l'approche par la donnée**



Buzz-F Etude de la viralité en ligne



Dispositif d'accueil : projet lauréat AAP DataLab 2021

- **Le projet**

❑ Etude de la viralité :
Analyse des phénomène viraux sur le web et de leur diffusion sur le web français

❑ Convention avec l'Université du Luxembourg

❑ Equipe C2DH
- Valérie Schafer, historienne
- Fred Pailler, ingénieur d'étude, sociologue

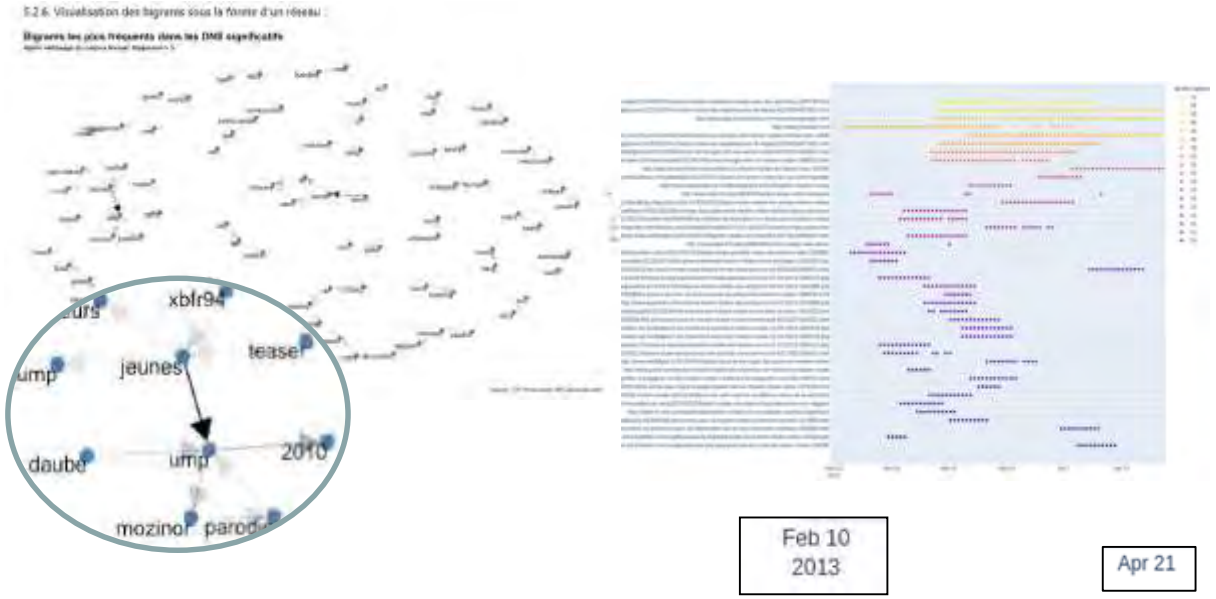
❑ Collecte large

Sites du domaine .fr et hébergés en France. Plusieurs millions de sites crawlés (2000 URL par site). Indexation limitée à l'URL et à la date de capture

❑ Collecte « Actualités » : recherche textuelle

Buzz-F datavisualisation

Etude de deux phénomènes : lip dub 2011 et Harlem shake 2013



- **La viralité : un phénomène difficile à appréhender**
- **dont l'empreinte se révèle dans les archives du web**

❑ Une approche des corpus par les données / métadonnées

❑ **Dépôt et datavisualisation**

Méthodologie et données sur GitLab Huma-num

- Tokenisation des URL

- Chronologie des traces numériques

- Création de sous-corpus vidéos

(BnF

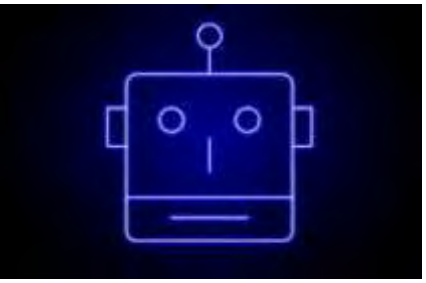
L’IA et les archives du web à la BnF

L’IA et la sélection

- Les correspondants peuvent utiliser des outils IA pour réaliser leur veille
- ***Une sélection thématique et un parcours guidé sur l’intelligence artificielle***

- Mais la sélection reste humaine avec le choix des paramètres de collecte

(BnF



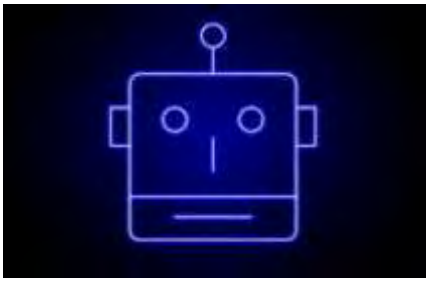
Impact des nouvelles technologies sur la collecte

Les évolutions technologiques comme les API, l'IA oblige à revoir les processus de collecte.

Par exemple pour des collectes de plateforme, les chercheurs utilisent des API pour récupérer le contenu.

La BnF pratique des formes hybrides comme pour sa collecte Youtube, puisqu'elle récupère les métadonnées de l'API en amont de son crawl. Cela permet de récupérer la liste des vidéos par chaine et d'évaluer le poids de la collecte. Une des conséquences est de devoir recréer une interface complète et artificielle de consultation.

(BnF



La collecte à l'heure de l'IA

- Quand des sites embarquent de l'IA, le robot ne fait pas de prompt et pour les sites marchands, le robot n'a pas de préférences de consommation.
- Collecte de vidéos, de témoignages sur l'IA

<html lang="fr-FR" data-build="prod-49b56bec084d41ac16d6a039a3add5463a0e2b9d" dir="ltr" class=""><head><meta charSet="UTF-8"/><meta name="viewport" content="width=device-width, initial-scale=1, viewport-fit=
import * as route0 from "https://cdn.oaistatic.com/assets/hbaqc5i44eqo3xfk.js";
import * as route1 from "https://cdn.oaistatic.com/assets/hknra01ttaxogo0e.js";
import * as route2 from "https://cdn.oaistatic.com/assets/iqxjdkm4sd4fiy9u.js";

window.__reactRouterRouteModules = {"root":route0,"routes/_conversation":route1,"routes/_conversation._index":route2};

import("https://cdn.oaistatic.com/assets/byj92clh16ig0xit.js");</script><!--$?--><template id="B:2"></template><!--/$--><script nonce="87e4eb65-965f-4a5d-b1e9-a40d8693f8fd" id="_R_">requestAnimationFrame(function(){$RT=per
$RC=function(a,b){if(b=document.getElementById(b))(a=document.getElementById(a))?(a.previousSibling.data="$~",$RB.push(a,b),2===$RB.length&&("number"!==typeof $RT?requestAnimationFrame($RV.bind(null,$RB)):(a=performance.no

<script nonce="87e4eb65-965f-4a5d-b1e9-a40d8693f8fd">window.ReactQueryError ??= class ReactQueryError extends Error {};
null==Promise.withResolvers&&(Promise.withResolvers=function(){let e,r;return{promise:new Promise((s,o)=>{e=s,r=o}),resolve:e,reject:r}});
window.__REACT_QUERY_CACHE__ ??= {}
window.__REACT_QUERY_CACHE__["[\"models\",\"{\\\"IIM\\\":false,\\\"isGizmo\\\":false}\"]"] = Promise.withResolvers();
window.__REACT_QUERY_CACHE__["[\"promptStarters\",8,null]"] = Promise.withResolvers();
</script>

<div hidden id="S:1"> <script nonce="87e4eb65-965f-4a5d-b1e9-a40d8693f8fd">$RC("B:1","S:1")</script>

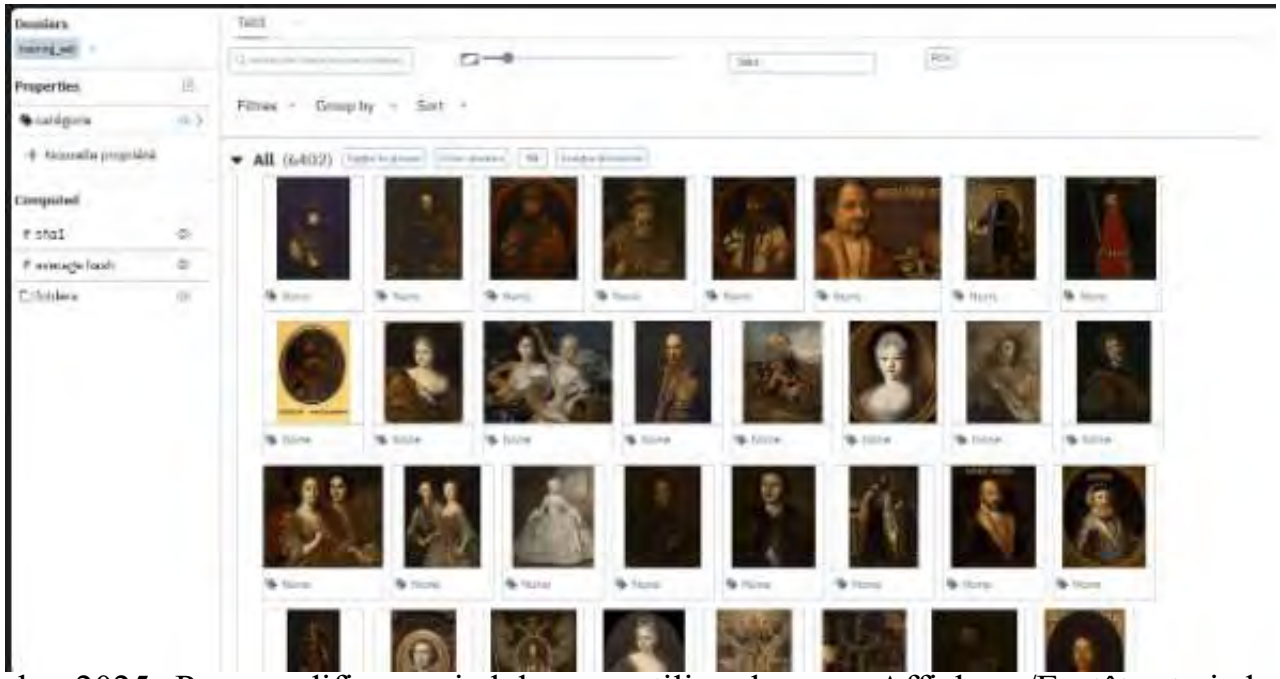
<script nonce="87e4eb65-965f-4a5d-b1e9-a40d8693f8fd">window.__REACT_QUERY_CACHE__["[\"promptStarters\",8,null]"].resolve({{items:[{id:"1b3b651a",title:"Créer une image",description:"pour ma présentation",oneliner:"Créer un
</script>

<script nonce="87e4eb65-965f-4a5d-b1e9-a40d8693f8fd">window.__REACT_QUERY_CACHE__["[\"models\",\"{\\\"IIM\\\":false,\\\"isGizmo\\\":false}\"]"].resolve((h=>({categories:[{color:"#000000",tagline:"",defaultModel:"auto",labe
</script>

<div hidden id="S:0"> <script nonce="87e4eb65-965f-4a5d-b1e9-a40d8693f8fd">$RC("B:0","S:0")</script></body></html>

La recherche dans les archives à l'heure de l'IA

- Des perspectives pour les images
  - Panoptic : outil de comparaison d'images, développé par le CERES



(BnF

Des perspectives pour les analyses textuelles

- Le projet AdaptMed : mise en place d'un système de génération automatique de reformulation pour les termes médicaux
  - Annotation des contenus collectés pendant le Covid pour constituer une base

Les limites à la recherche

- Les données des archives du web sont sous droit, ce qui limite les possibilités d'explorer les données avec des outils externes à la BnF

(BnF Intelligence artificielle

Feuille de route 2021-2026

À la croisée des chemins, la **Bibliothèque nationale de France** se constitue à la fois comme un observatoire privilégié, comme une instance réflexive et critique, et comme un moteur des évolutions liées à l'intelligence artificielle. Cette feuille de route présente une projection à 5 ans, fortement ancrée dans les réalisations actuelles et les projets en cours. Davantage qu'un achèvement, la démarche proposée est envisagée comme l'initialisation d'une dynamique qui sera appelée à s'étoffer et à évoluer.

Actions

1
Inscrire l'IA dans
la stratégie
de l'établissement

2
Organiser
la R&D

3
Acquérir
de **nouvelles compétences**,
indispensables
à l'émergence
et à la conduite
des projets IA

4
Préparer
l'infrastructure
et **les données**

5
Monter
un programme pluriannuel
sur l'IA
en partenariat avec
des acteurs clés

Jalons

Inscription de la stratégie IA dans le Contrat d'objectifs et de performance

Conférence internationale *Futurs Fantastiques*

Un premier projet IA en cours d'industrialisation

Catalogue de services pour l'infrastructure DataLab

Charte éthique numérique et tableau de bord des indicateurs associés

Premiers partenariats DataLab

Cycle de conférences « les débats de la science » sur l'IA

3 à 4 projets expérimentaux en cours

Au moins 10 agents ayant des compétences pointues en IA

Début du programme pluriannuel IA

Référentiel global pour l'élaboration, la mise en œuvre et le suivi des projets de R&D

15 à 20% du personnel sensibilisé à l'IA

Gestion prévisionnelle de l'évolution des métiers

Projet de service DSI

Cartographie des données et schéma de gouvernance

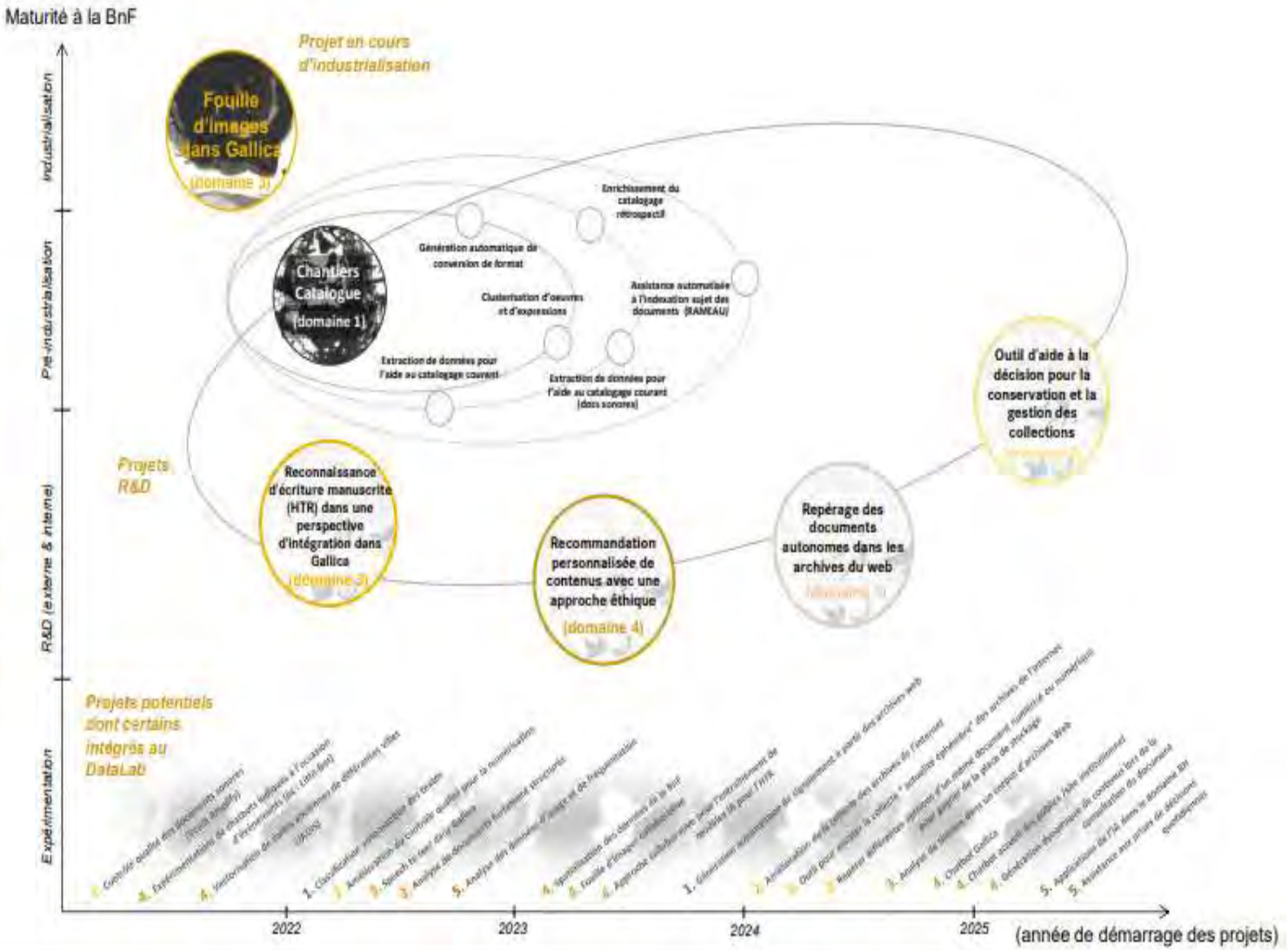
Bilan du réseau DataLab

5 à 10 projets réalisés dans le programme pluriannuel

Projets du programme pluriannuel

Ces projets relèvent de 5 domaines :

1. Aide au catalogage et signalement
2. Gestion des collections, des entrées à la conservation
3. Exploration, analyse des collections et amélioration de l'accès
4. Médiation, valorisation et éditorialisation des collections
5. Aide à la décision et au pilotage



Information et contact
www.bnf.fr/fr/feuille-de-route-ia
BIBLIOTHÈQUE NATIONALE DE FRANCE
JUIN 2022

Merci !

- Contacts : depot.legal.web@bnf.fr et datalab@bnf.fr

- Webcorpora et Carnet de recherche de la BnF
  https://webcorpora.hypotheses.org/ et https://bnf.hypotheses.org/

- Guide des archives du web
  https://www.bnf.fr/sites/default/files/2025-02/Guide_Archives_internet_BnF.pdf

- L'IA à la BnF : https://www.bnf.fr/fr/lintelligence-artificielle-la-bnf