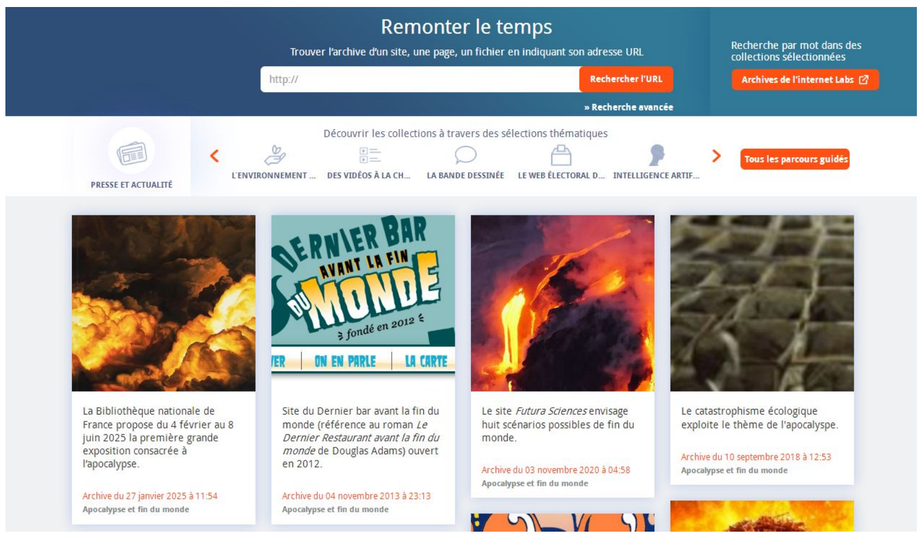
The BnF DataLab services and tools for researchers working on web archives

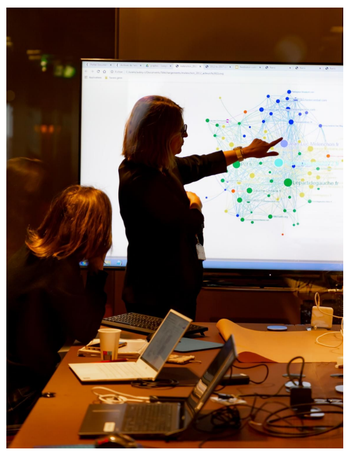
BnF Archives de l'internet



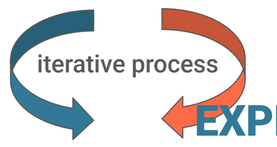
- From 1996 to now.
- 59 billion URLs, 2.5 PB of data.
- Annual broad crawls (6 million domains) and regular thematic, event, media crawls (70,000 websites).
- 6 collections (53 TB) available with full-text search.

The BnF DataLab: collections research and promotion

- A reading room, a technical infrastructure, a network of experts.
- A set of services: collections presentations, training and workshops, scientific events, on-demand crawls, full text indexing, support in metadata, text, image and link mining.
- Different hosting schemes: residence, partnership (associate researchers), call for projects.
- Building a community of researchers and web archivists.



Research stages and available tools



DEFINE CORPUS

URL de départ	Fréquence	Thème	Mots clés
https://13sens.com	2 fois par an	théâtres de province	théâtre / Obernai / cirque / musiqu
https://8nenverse.com	1 fois par an	danse	Huit renversé / danse contemporai
https://abcd2jon.org	2 fois par an	théâtres de province	ABC / Dijon / théâtre / danse / mus
http://aberrationmentalis.free.fr	1 fois par an	théâtres de province	non renseigné
http://abrideabattue.blogspot.com/search/label/spect...	2 fois par an	arts du spectacle	non renseigné
https://acthalia.com	2 fois par an	théâtres de province	Le Mans / Acthalia / théâtre
https://agatheprojetsante.wixsite.com/auquotidien	1 fois par mois	danse	danse / cirque / recherche en théât
https://alauda.fr	1 fois par an	théâtres de province	non renseigné
https://albatros-pernes.com	1 fois par an	théâtres de province	non renseigné
http://allegrotheatre.blogspot.com	1 fois par an	arts du spectacle	non renseigné
https://altigone.fr	2 fois par an	théâtres de province	Altigone / Saint-Orens-de-Gameville
http://amiduatheatre.dax.free.fr/pages/edito.php	1 fois par an	théâtres de province	ATP / théâtre / Dax
https://amis-theatre-biarritz.com	2 fois par an	théâtres de province	ATP / Biarritz / théâtre
https://ancre-bretagne.fr	1 fois par an	arts du spectacle	Ancre / réseau jeune public / assoc
https://annetheatrepassion.blogspot.com	1 fois par an	arts du spectacle	non renseigné
https://annibal.annibal-lacave.com	1 fois par an	spectacles de rue	Compagnie Annibal et ses éléphant
https://arcenciel.lievin.fr	1 fois par an	théâtres de province	Liévin / théâtre / danse musique
http://arc-en-cirque.asso.fr	1 fois par an	cirque	Arc en cirque / Chambéry / Savoie ,
http://archives.studiotheatre.fr	2 fois par an	théâtres d'Ile-de-France	théâtre / recherche / création / exp

URL search

- BnF curated seeds
- Researcher curated seeds (URL check)
- CDX queries (URL pattern, date, format)

Full-text search

- Archives de l'internet Labs
- SolrWayback
- SolrAdmin (for complex requests)
- PANDORAE

PANDORÆ

PANDORÆ is a software built for scientific research. Its purpose is to enable its users to retrieve, normalize, and explore collections of documents from various established sources.

Hyperlink exploration

- Hyphe: define web entities (actors), crawl the web archive, discover new web entities through an iterative prospection and curation process

Live web

- On-demand crawl
- On-demand full-text indexing

Archived data

- Websites in tree structure (REPRO)
- Documents in flat structure (CORPUS)
- Filtered/Derivative WARCs (WARCD)

Metadata

- Zotero references with PANDORAE
- Collection, links, document metadata with Archives Unleashed Toolkit (AUT)
- Discovered URLs with Hyphe

Fragments

- Text snippets with SolrAdmin
- Snapshots with Chromium

All-in-one tools

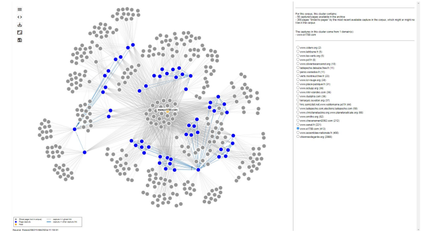
- SolrWayback
- PANDORAE

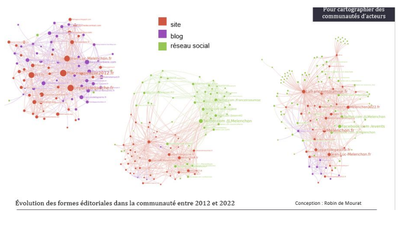
Text

- Archives Unleashed Toolkit
- JusText
- Gargantext
- Other tools used by Digital Humanities and for Natural Language Processing

Links

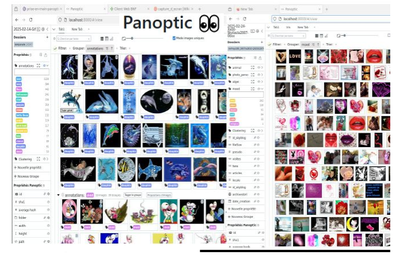
- Link graphs and web cartography with Hyphe





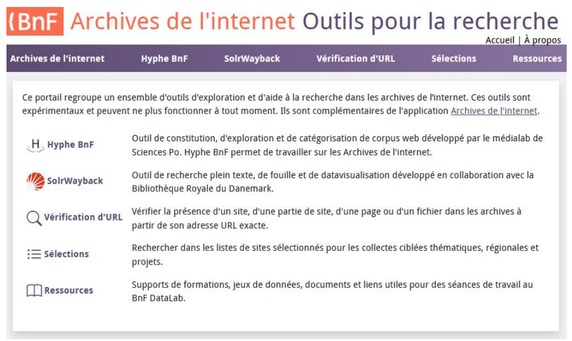
Images and videos

- Cluster and annotate with Panoptic



A "Research Portal"

A demonstrator with data, metadata, exploration and curation tools and educational material.



"These tools are experimental and may cease to function at any time." ⚠

A variety of tools and implementation models

- IIPC community tools (SolrWayback, Archives Unleashed Toolkit…)
- In-house developed tools focused on data and metadata extraction
- Tools developed by researchers to study the live web or online resources adapted to web archives (Hyphe, PANDORAE…)
- Generic tools used by DH or NLP researchers integrated into project pipeline (Gargantext, Panoptic, JusText…)

IIPC netpreserve.org INTERNATIONAL INTERNET PRESERVATION CONSORTIUM

IIPC Web Archiving Conference
April 9-10th 2025, Oslo

- BnF's Web Archives introduction: https://www.bnf.fr/fr/depot-legal-du-web
- BnF's DataLab services: https://www.bnf.fr/fr/bnf-datalab
- Contacts:
  Sara Aubry sara.aubry@bnf.fr
  Dorothée Benhamou-Suesser dorothee.benhamou-suesser@bnf.fr
- Photo credits: C. Maufroy / Sciences Po