Dorothée
Benhamou-Suesser

Guillaume Levrier

Closed data, open software

Building new ways into the French web archives

Solving the tension between **captive, long-term, institutional data sources**

and the need for

open-source software for scientific research.

Preserving the French web at the BnF

DADVSI law (2006) and decree : Mission to harvest, curate and give access to “signs, signals, writings, images, sounds or messages of every kind communicated to the public by electronic channels”

Processes and tools

- Open-source tools
- IIPC community

Harvest

Web crawler (bot) : Heritrix

Preservation

BnF digital long term preservation system based on OAIS framework

Access

Open Wayback

Collection’s scope

A mixed harvesting model combining 2 types of crawls

« broad crawl » or « national domain crawls »

- Once a year
- On a large sample of sites (5.9 M French domains in 2024)
- No selective process, lists of domain names from AFNIC, OVH
- 2000 to 2500 URL per domain

Thematic and curated crawls

- Selective (based on a network of contributors) and extension of printed collections at the BnF
- More frequent : daily, monthly, annually, etc.
- More in depth (up to 250 000 URL)
- Disciplinary (literature, sciences, history, etc.), thematic (environmental issues) or event-based (Elections, Covid-19, Olympics…)

Challenges to make these collections more open to scientific research

On-site access only

Massive data (52mds URL, 1,7 Pio), specific tools

Digital artefacts

"Unlike traditional institutional archives, the snapshots that comprised the archived Web are artifacts created by the archival process itself. (…)

These reconstructions of the live Web are never exact replicas (…) This makes an understanding of web archives especially important."

(Niels Brügger, Ian Milligan (eds), *The SAGE Handbook of Web History*, 2019)

Doing research involves building a **methodological** strategy to serve an **epistemological** purpose.

This, in turn, implies continuously asking oneself how we can create something that would deserve being called « scientific knowledge ».

The current vetting of this label is paradoxically **very rarely enforced**.

What is usually called « scientific knowledge » comes from [articles] published in [peer reviewed journals]. But the reviewers *de facto* seldom have access to either the data or the methods that were used to build the experiment that led to the presented results.

Peer review is about the **plausibility of a narrative** that would have led to the « results ».

Hence the need for tools that are:

- free
- open-source
- with a commented source code
- whose execution can be decentralized
- whose outputs are under the control of the user

In theory, respecting these principles helps ensuring the reproducibility of the work. Getting the same data in the same algorithm (often) yields the same (or comparable) results.

Trying to reach all available

data sources

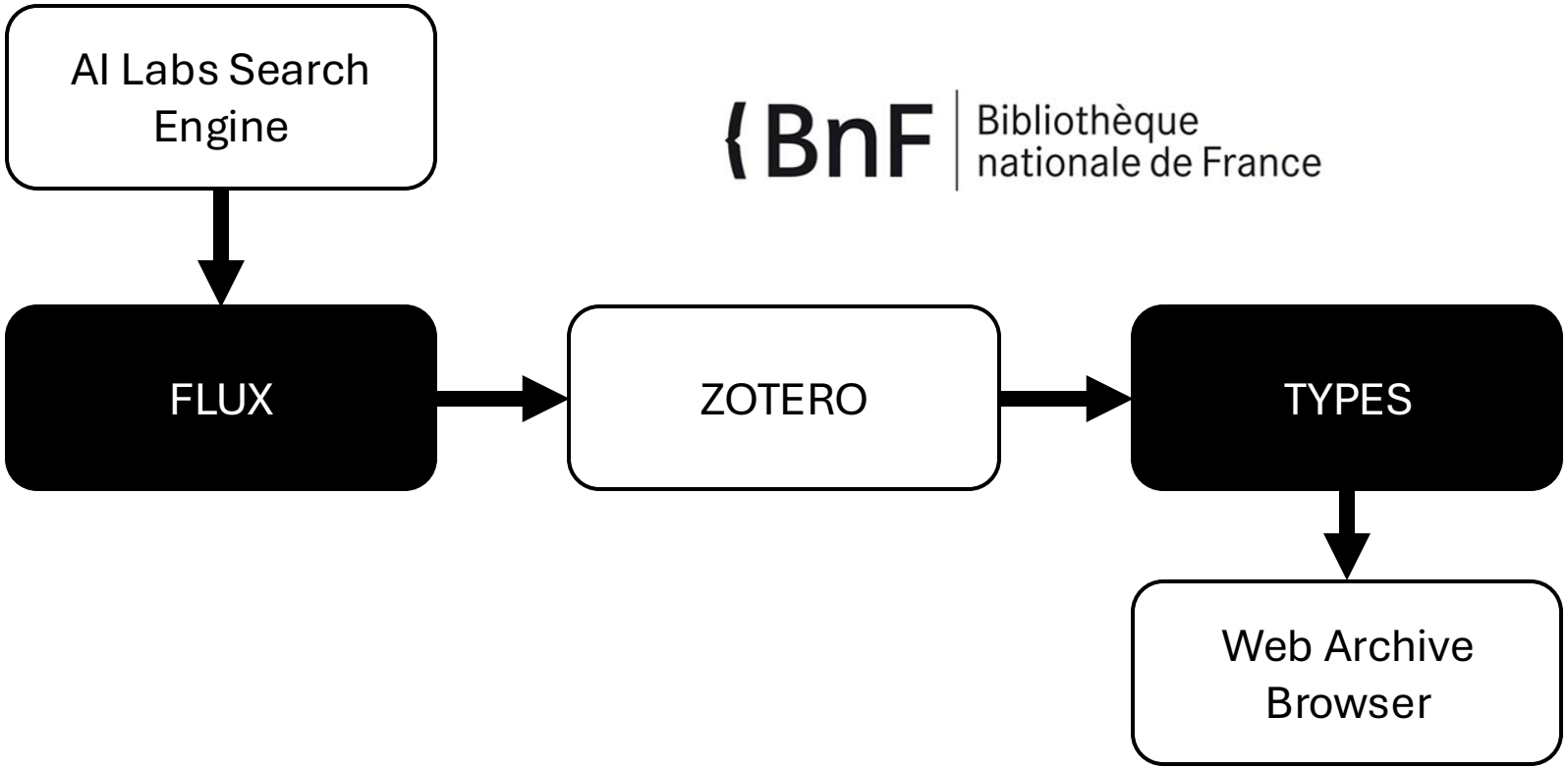
for one’s research.

The purpose of PANDORÆ is to:

- Harvest datasets
- Standardize datasets
- Explore datasets

The purpose of PANDORÆ is to:

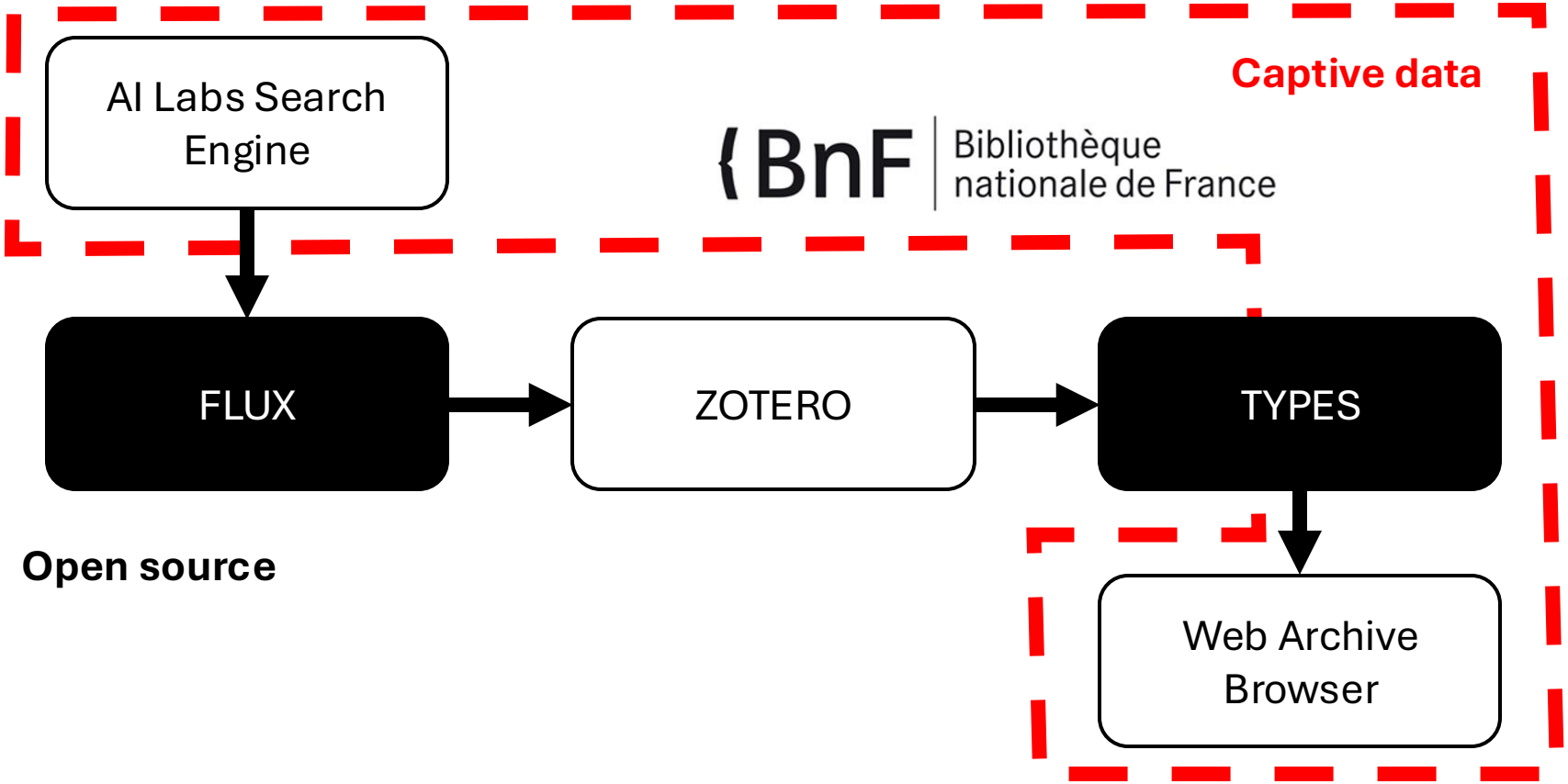
- Harvest datasets
- Standardize datasets
- Explore datasets



BnF Bibliothèque nationale de France

The purpose of PANDORÆ is to:

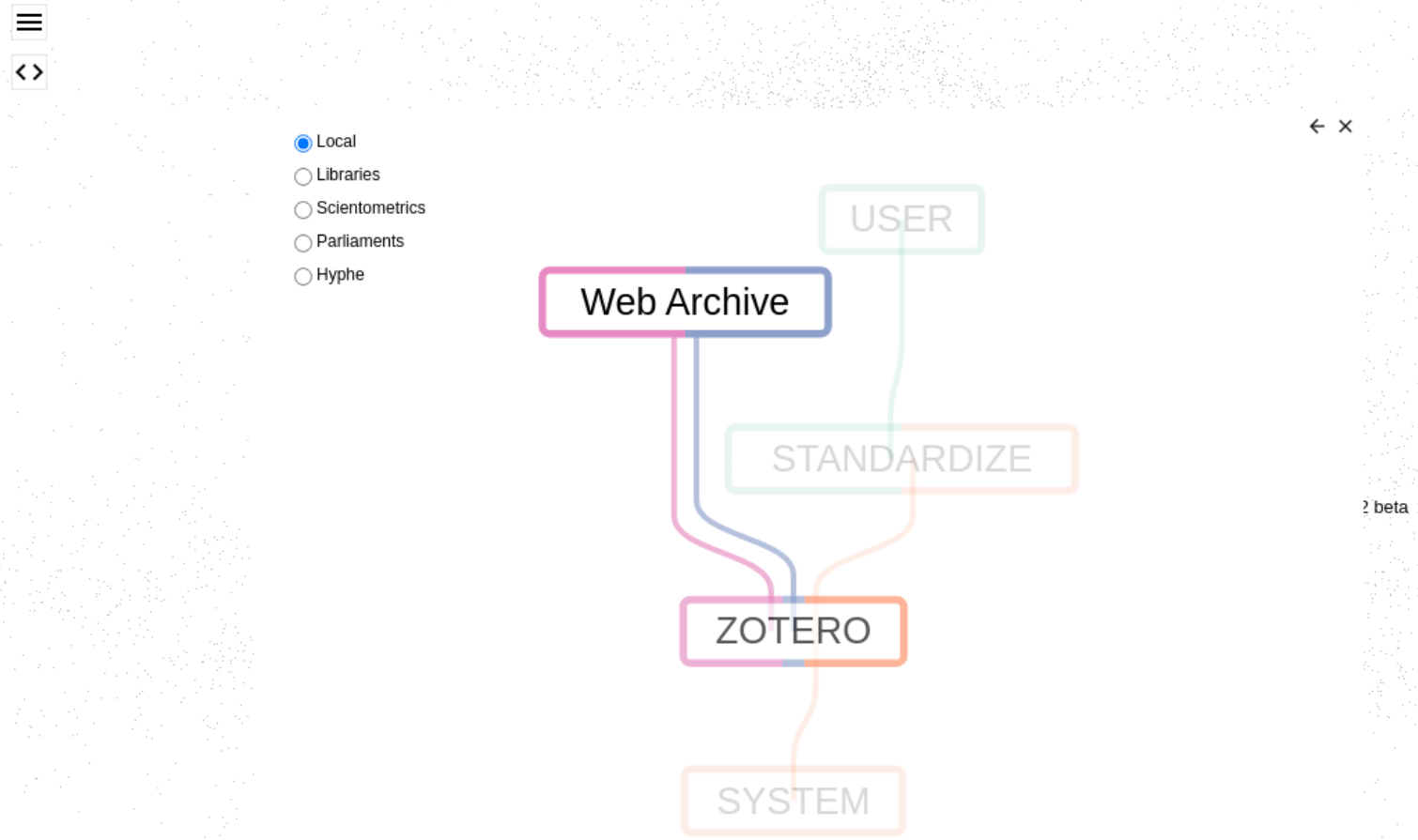
- Harvest datasets
- Standardize datasets
- Explore datasets



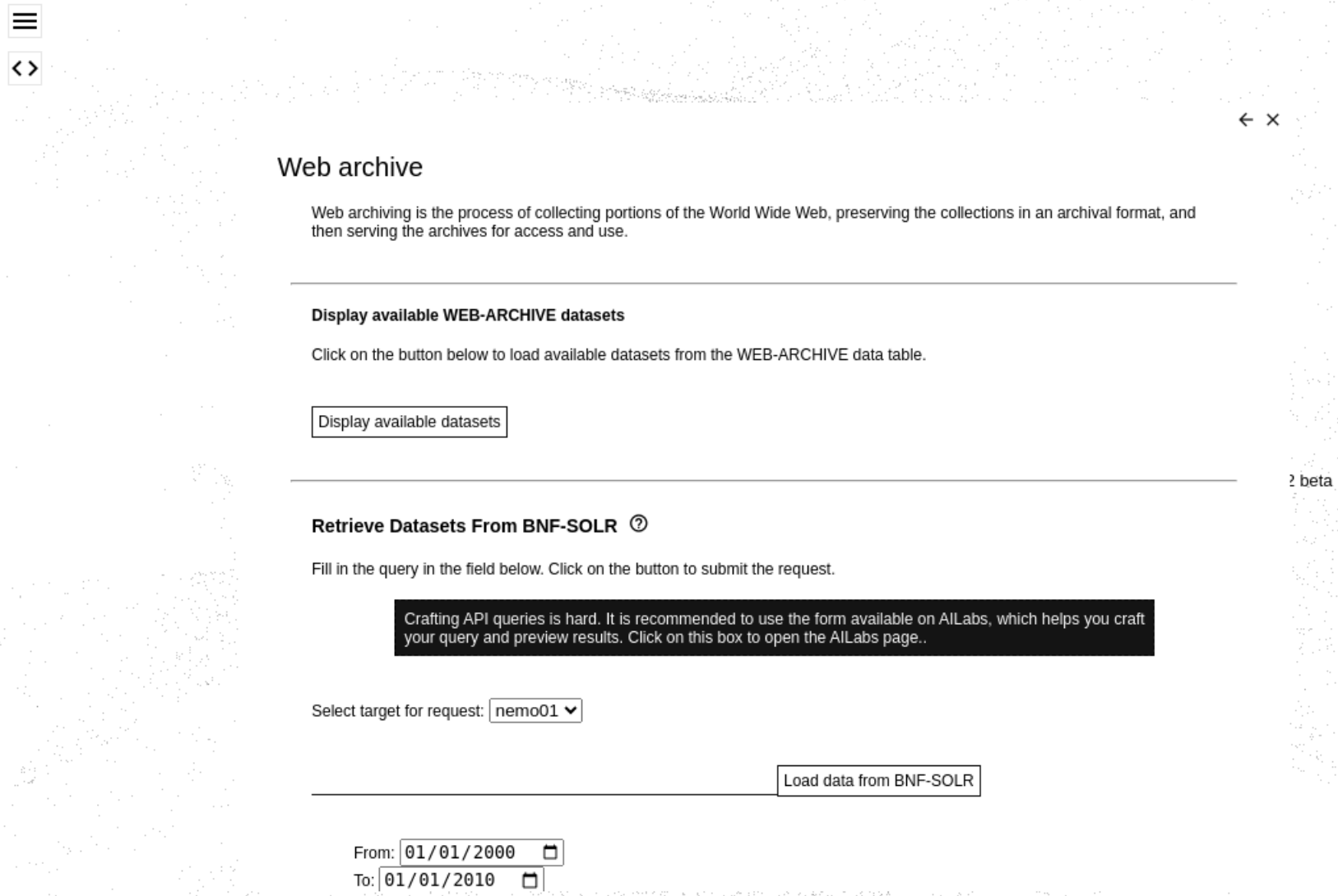
An open tool inserted in a captive context

An open tool inserted in a captive context

FLUX – Harvesting and standardizing data with PANDORÆ



FLUX – Harvesting and standardizing data with PANDORÆ



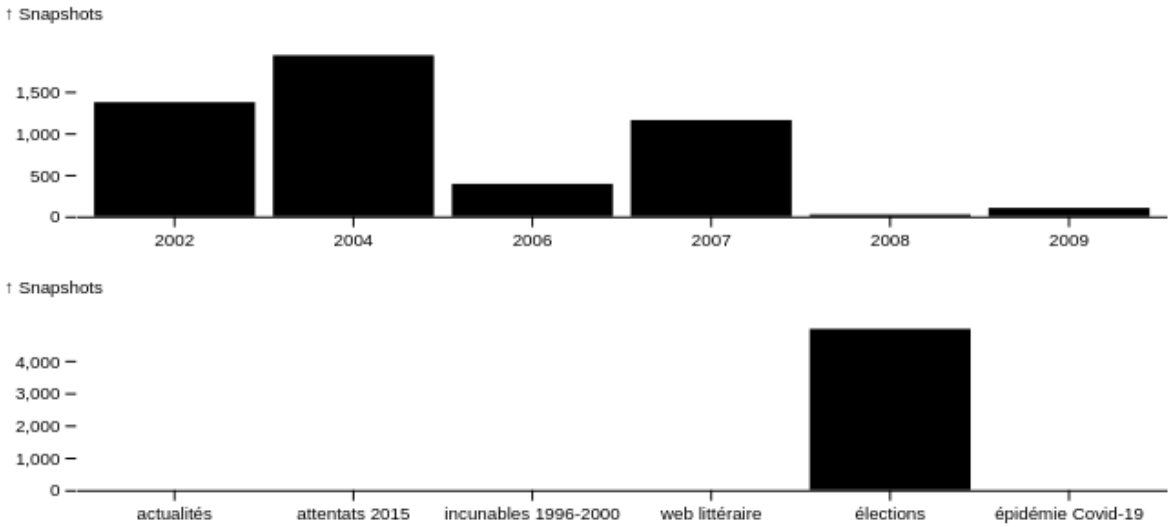
FLUX – Harvesting and standardizing data with PANDORÆ

attentats 2015

web littéraire

Result profile for query **dolly AND clon*** on period 2000-01-01 to 2010-01-01

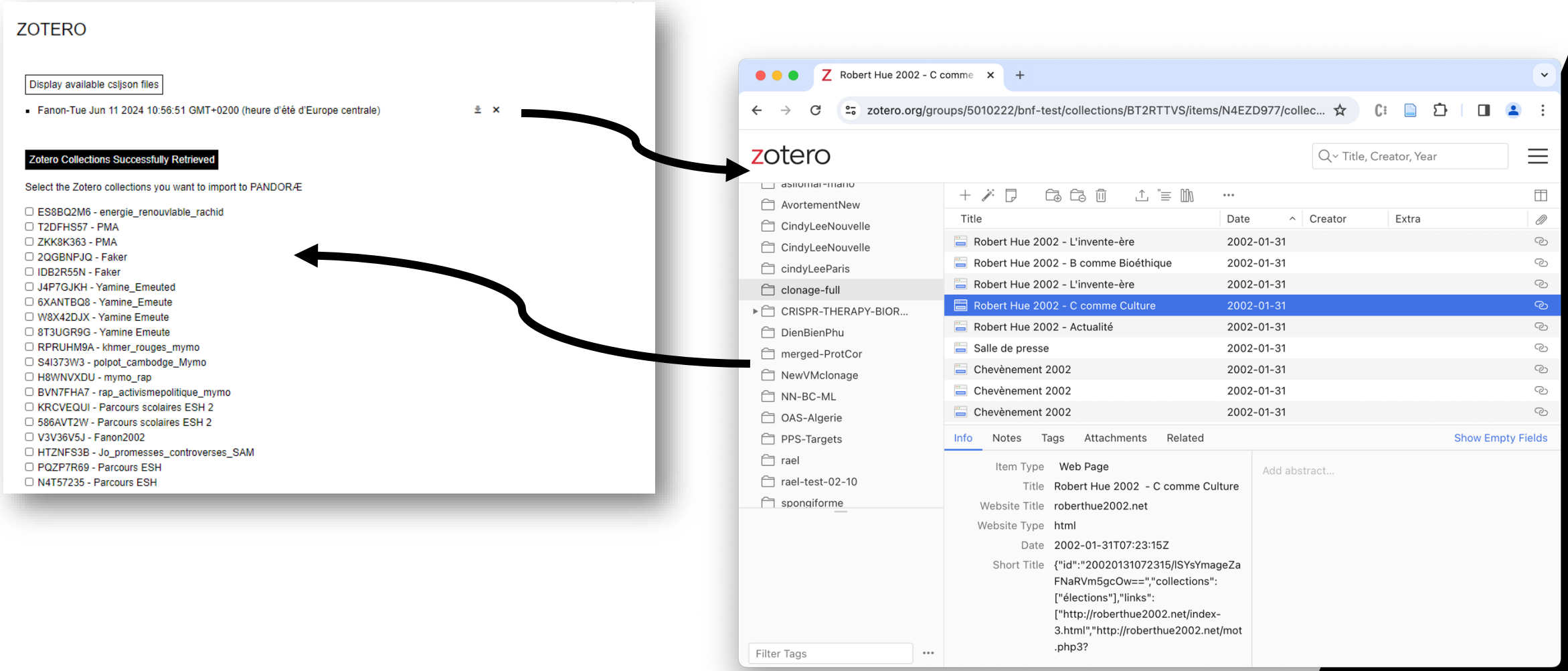
1064 unique documents found



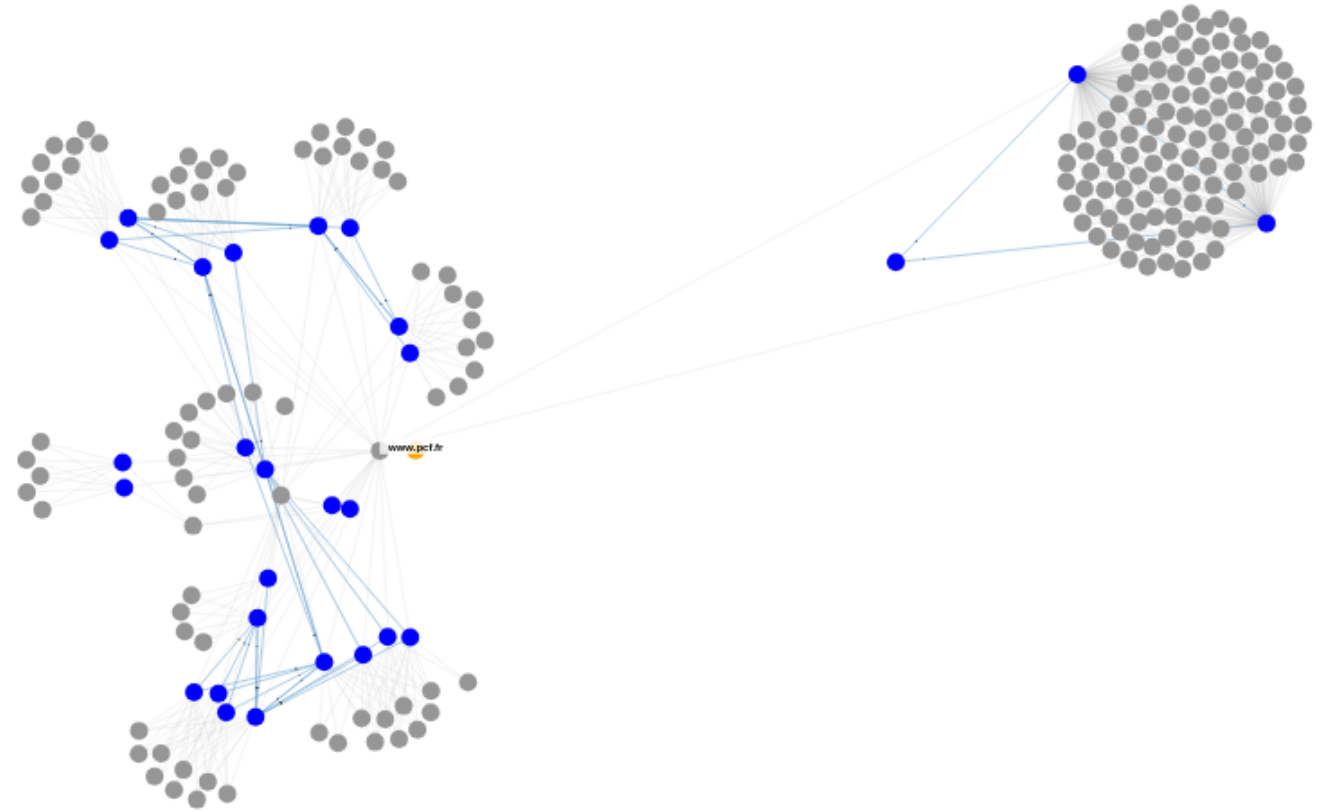
Captures found for the top 10 domains:
pcf.fr - 2145
assemblee-nationale.fr - 326
estvideo.com - 309
topchretien.com - 198
senat.fr - 194
e1789.com - 143
presidentielles.net - 127
ladepeche.com - 121

beta

ZOTERO – Saving and curating collections



TYPE – Explorer des corpus avec PANDORÆ



Ghost page (not in corpus)
Page capture
Host
capture <-> ghost link
capture -> other capture link

dollyANDclon*3-28_01_2025-14+29+54

dollyANDclon*3
Source : zotero
Date : 28/01/2025-14:29:54

For this corpus, this cluster contains:
- 27 captured pages available in the archive
- 200 pages "linked to pages" by the most recent available capture in the corpus, which might or might not be in the corpus

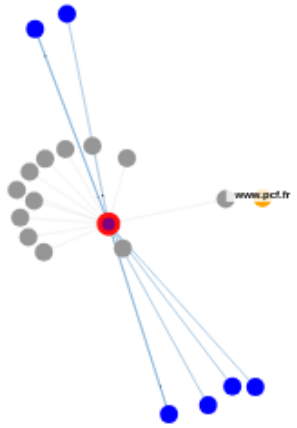
The captures in this cluster come from 1 domain(s):
- www.pcf.fr

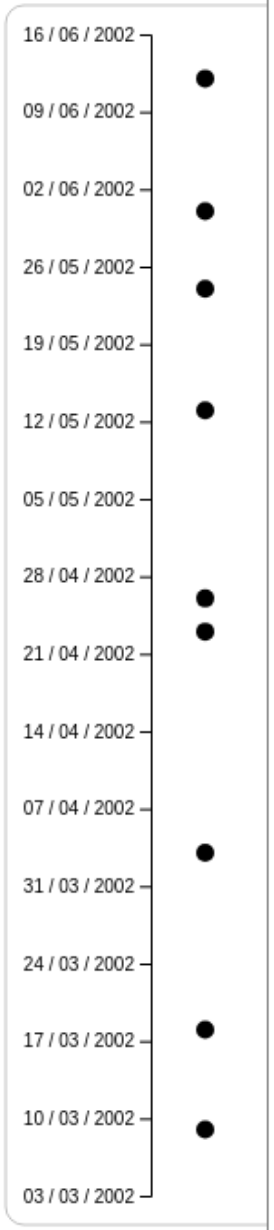
Hide solo ghosts

nemo01

- www.demlib.com (2)
- 2002.scrutins.net (3)
- www.academie-sciences.fr (4)
- www.les-verts.org (4)
- www.elysee.fr (5)
- membres.lycos.fr (10)
- presidentielles2002.citegay.fr (11)
- www.roger-mei.org (19)
- pluriel.free.fr (20)
- digipressetmp4.teaser.fr (20)
- www.d-l-c.org (22)
- www.mauvaisgenres.com (26)
- www.corriere.it (31)
- www.largeur.com (42)
- www.jean-francoismattei.com (65)
- www.noelmamere.eu.org (74)
- news.independent.co.uk (108)
- www.presidentielles.net (109)
- www.rajf.org (121)
- www.senat.fr (141)
- www.e1789.com (151)

TYPE – Exploring datasets with PANDORÆ





dollyANDclon*3
Source : zotero
Date : 28/01/2025-14:29:54

dolly | Copy permalink

se américaine « Géron corporation », pionnière du clonage par transfert de noyau de cellules adultes, a fusionné avec « Roslin BioMed's » qui a conçu Dolly. On voit ici la cohérence entre la demande de brevetage et l'autorisation du clonage thérapeutique. Dans un monde qui laisse l'Afrique à l'aband

url
http://www.pcf.fr/docs/indexa580.html?iddoc=101&iddos=25&idcol=22&cpt=40

id
20020513000000/GzR1YH0AVhauZyxlrQfQQQ==

resourcename
indexa580.html

wayback_date
20020513000000

title
PCF

content_first_bytes
3c 68 74 6d 6c 3e 20 0d 0a 3c 21 2d 2d 20 4d 69 72 72 6f 72 65 64 20 66 72 6f 6d 20 77 77 77 2e

ssdeep_hash_bs_768
ynHgTRZoU7jllgfH8oZl0xz5lvjSz+CZv+a90JU4S

ssdeep_hash_bs_1536
ynyRZoUt2/8oZmvSz+ib90JnS

version
1779471020176965600

content
Collège exécutif Particommuniste français exec
dossiers:Collège exécutifDéclara
(notes,etc.)DiscoursCommuniquésCommunistes
journalContenu du dossier Collège exécutif:préce
exécutif du 06 juin 2001Collège exécutif du 29 m

Ghost page (not in corpus)
Page capture
Host
capture <-> ghost link
capture -> other capture link

dollyANDclon*3-28_01_2025-14+29+54

PANDORÆ and the BnF hence try to solve the tension between closed data and open software by following what we tentatively call a « **one-way mirror** » model.

In this model, the web archive is like a suspect in an interrogation room that can only tell you **their truth on what happened in the past**.

1. You can harvest the data when you're in the room **but** you cannot have a write/erase access to it.

2. You can take parts of each record with you (the metadata) **but** cannot interrogate it when you're away.

3. You can come back for a highlight on a specific piece **but** cannot take the whole body out of the room.

Dorothée
Benhamou-Suesser

Guillaume Levrier

Closed data, open software

Building new ways into the French web archives