



Dorothee

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

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

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

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)*

Digital artefacts

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.

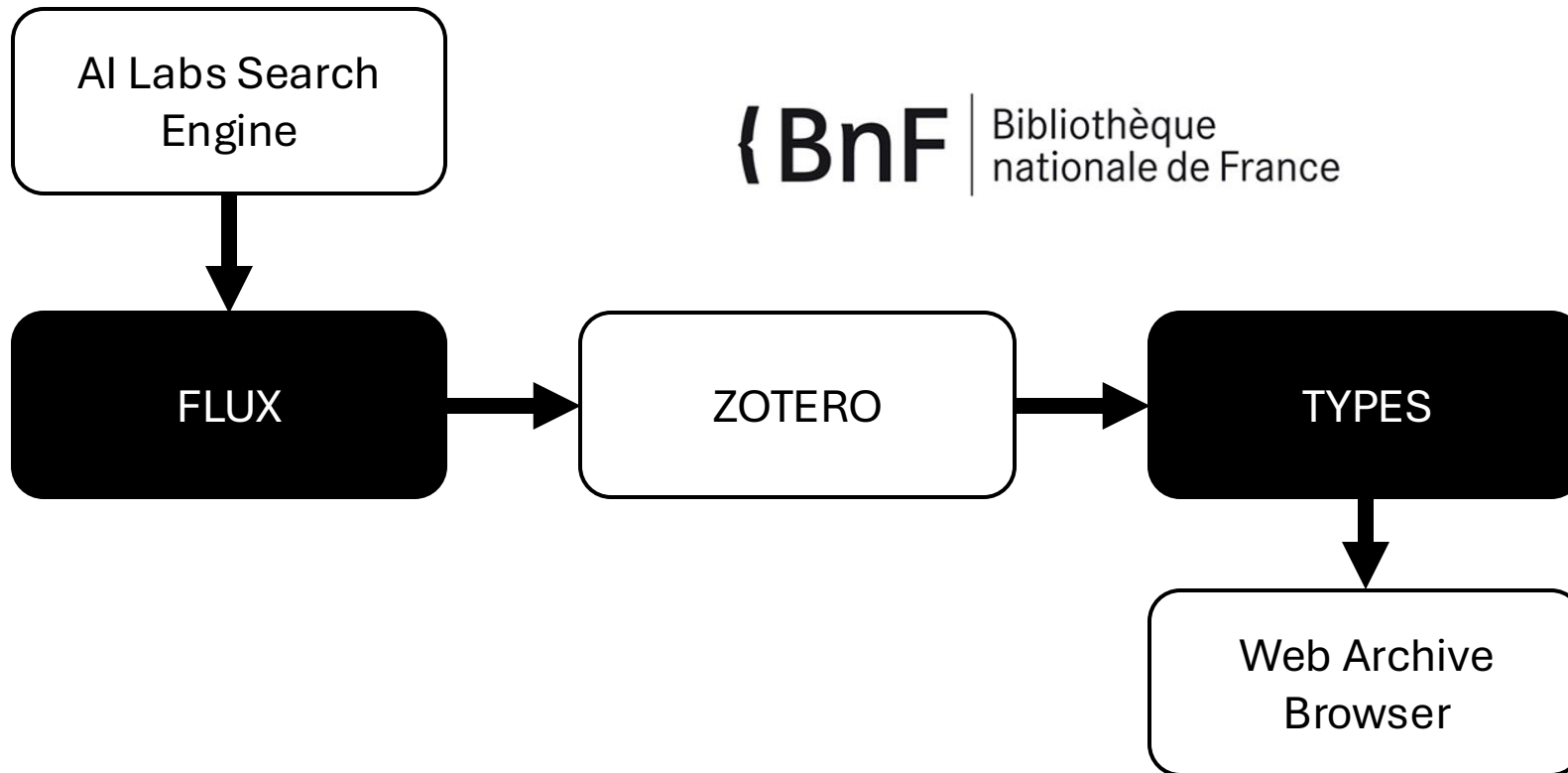
The purpose of PANDORÆ is to:

- Harvest datasets
- Standardize datasets
- Explore datasets



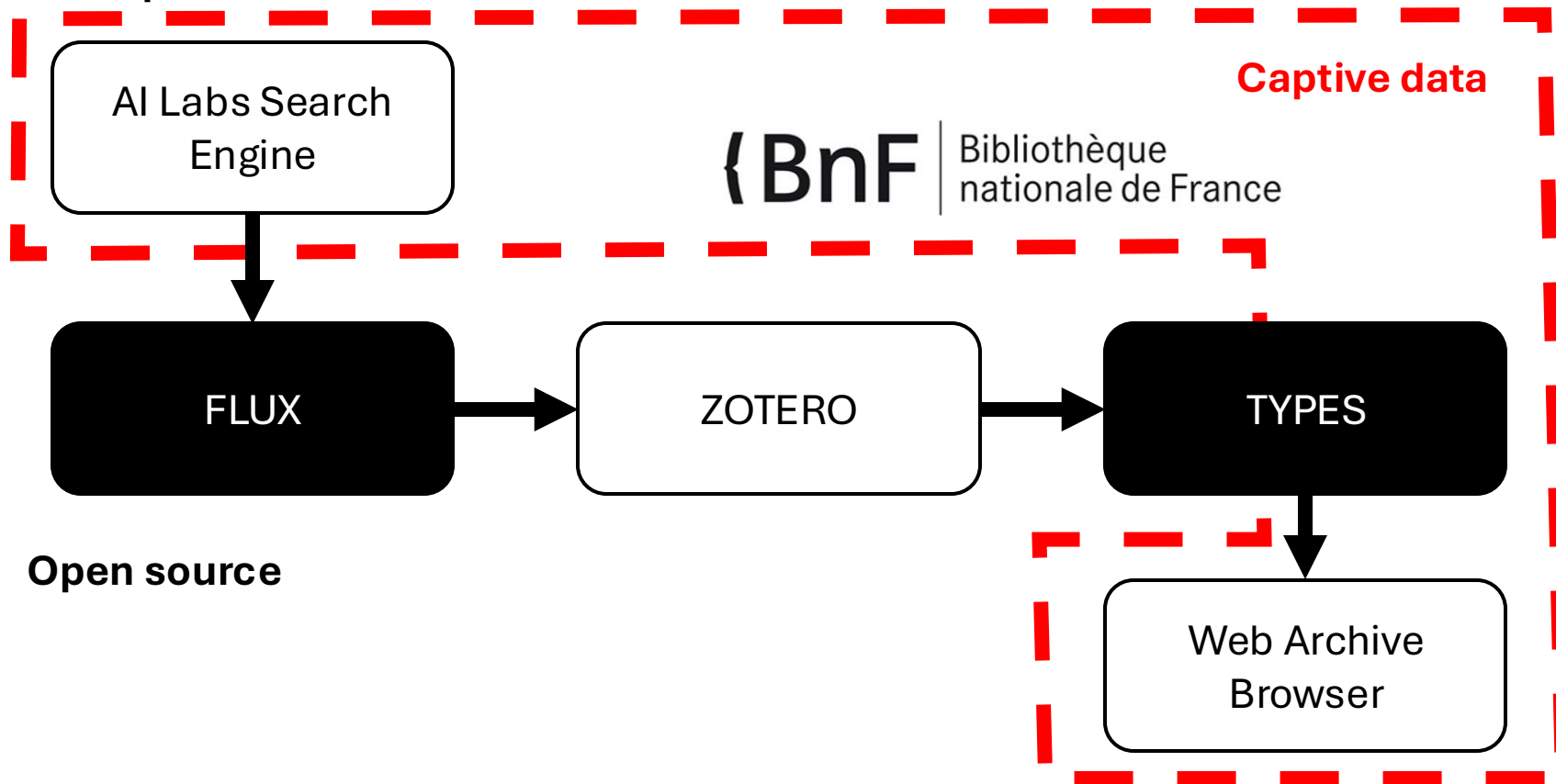
The purpose of PANDORÆ is to:

- Harvest datasets
- Standardize datasets
- Explore datasets



The purpose of PANDORÆ is to:

- Harvest datasets
- Standardize datasets
- Explore datasets



An open tool inserted in a captive context

{BnF Archives de l'internet Labs

Trouver l'archive d'un site, d'une page ou d'un fichier

» Recherche avancée

» Recherche experte

» Recherche n-gram

Recherche avancée (?)

Mot(s) ou "expression" :

Collection(s) :

- toutes actualités attentats 2015 incunables 1996-2000 épidémie Covid-19
 web littéraire élections

Date de collecte :

Ou période :

de

à

Hôte, nom de domaine
ou extension :

Format de fichier :

Exclure le(s) mot(s) :

An open tool inserted in a captive context

Trouver l'archive d'un site, d'une page ou d'un fichier

Rechercher

» Recherche avancée

» Recherche experte

» Recherche n-gram

3 213 résultats

Rappel de la recherche :

Mot(s) ou \"expression\" : Fanon

Collection(s) : élections

Période : de 01/01/2002 à 31/12/2002

[modifier](#)

Trier par : Pertinence

← Page 1 →

10 résultats par page

Trier les valeurs des facettes : 0-9 | A-Z

Collection (1) inclure | exclure

+ élections (3 213)

Année (1) inclure | exclure

+ 2002 (3 213)

Nom de domaine (10+) inclure | exclure

+ chiensdegarde.org (294)

+ chevenement2002.com (348)

+ e1789.com (129)

+ lcr-rouge.org (75)

+ ladepêche.com (60)

+ assemblée-nationale.fr (46)

\"Forum généraliste - Re: reponse à Djembé\"

Archive du 26 mai 2002

Format : html

http://www.e1789.com/agora/view.php?site=e1789&bn=e1789_forumpublic&key=1022295317&first=1022368249&last=1022279623

[Voir les 3 captures](#)

le son magnanime du djembé sorcier et ensorcelant, pour offrir à l'Afrique ce qu'elle aura de meilleur à offrir au monde.Si Franz **Fanon** est mort pour Djembé, qui à force de taper à la sourdine le djembé, allant

\"Forum généraliste - Re: reponse à Djembé\"

Archive du 31 mai 2002

Format : html

http://www.e1789.com/agora/viewee3c.html?site=e1789&bn=e1789_forumpublic&key=1022295317&first=1022368249&last=1022279623

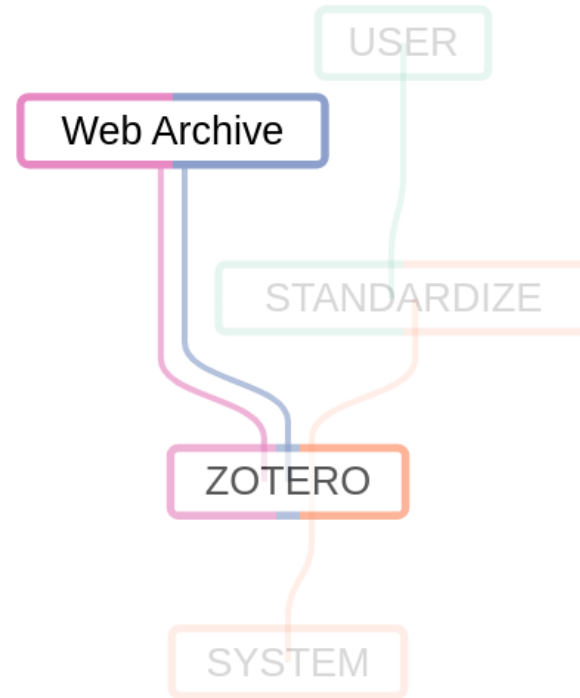
[Voir les 3 captures](#)

le son magnanime du djembé sorcier et ensorcelant, pour offrir à l'Afrique ce qu'elle aura de meilleur à offrir au monde.Si Franz **Fanon** est mort pour Djembé, qui à force de taper à la sourdine le djembé, allant

FLUX – Harvesting and standardizing data with PANDORÆ



- Local
- Libraries
- Scientometrics
- Parliaments
- Hyphe



← x

? beta

FLUX – Harvesting and standardizing data with PANDORÆ



Web archive

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.

Display available WEB-ARCHIVE datasets

Click on the button below to load available datasets from the WEB-ARCHIVE data table.

Display available datasets

? beta

Retrieve Datasets From BNF-SOLR [?](#)

Fill in the query in the field below. Click on the button to submit the request.

Crafting API queries is hard. It is recommended to use the form available on AILabs, which helps you craft your query and preview results. Click on this box to open the AILabs page..

Select target for request:

Load data from BNF-SOLR

From:

To:



FLUX – Harvesting and standardizing data with PANDORÆ

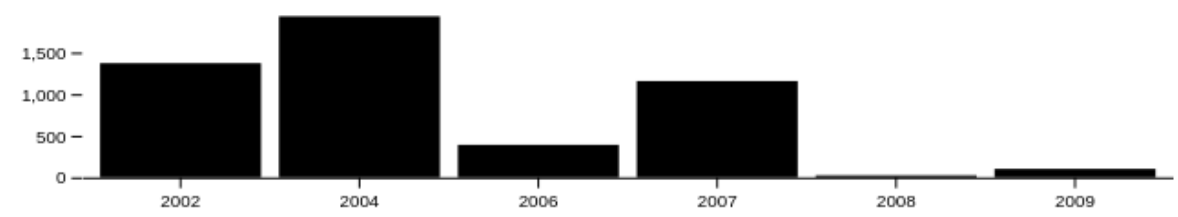
- attentats 2015
- web littéraire



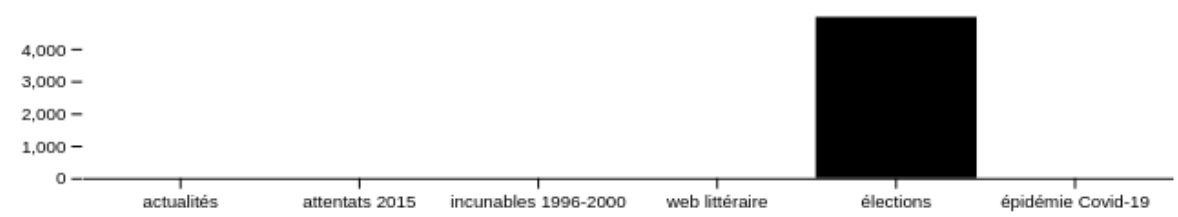
Result profile for query **dolly AND clon*** on period 2000-01-01 to 2010-01-01

1064 unique documents found

↑ Snapshots



↑ Snapshots



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

2 beta

ZOTERO – Saving and curating collections

ZOTERO

Display available csljson files

Fanon-Tue Jun 11 2024 10:56:51 GMT+0200 (heure d'été d'Europe centrale)

Zotero Collections Successfully Retrieved

Select the Zotero collections you want to import to PANDORÆ

- ES8BQ2M6 - energie_renouvable_rachid
- T2DFHS57 - PMA
- ZKK8K363 - PMA
- 2QGBNPJQ - Faker
- IDB2R55N - Faker
- J4P7GJKH - Yamine_Emeuted
- 6XANTBQ8 - Yamine_Emeute
- W8X42DJX - Yamine_Emeute
- 8T3UGR9G - Yamine_Emeute
- RPRUHM9A - khmer_rouges_mymo
- S4I373W3 - polpot_cambodge_Mymo
- H8WNVXDU - mymo_rap
- BVN7FHA7 - rap_activismepolitique_mymo
- KRCVEQUI - Parcours_scolaires_ESH_2
- 586AVT2W - Parcours_scolaires_ESH_2
- V3V36V5J - Fanon2002
- HTZNFS3B - Jo_promesses_controverses_SAM
- PQZP7R69 - Parcours_ESH
- N4T57235 - Parcours_ESH

Robert Hue 2002 - C comme

zotero.org/groups/5010222/bnf-test/collections/BT2RTTVS/items/N4EZD977/collec...

zotero

Search: Title, Creator, Year

Title	Date	Creator	Extra
Robert Hue 2002 - L'invente-ère	2002-01-31		
Robert Hue 2002 - B comme Bioéthique	2002-01-31		
Robert Hue 2002 - L'invente-ère	2002-01-31		
Robert Hue 2002 - C comme Culture	2002-01-31		
Robert Hue 2002 - Actualité	2002-01-31		
Salle de presse	2002-01-31		
Chevènement 2002	2002-01-31		
Chevènement 2002	2002-01-31		
Chevènement 2002	2002-01-31		

Info Notes Tags Attachments Related Show Empty Fields

Item Type: Web Page

Title: Robert Hue 2002 - C comme Culture

Website Title: roberthue2002.net

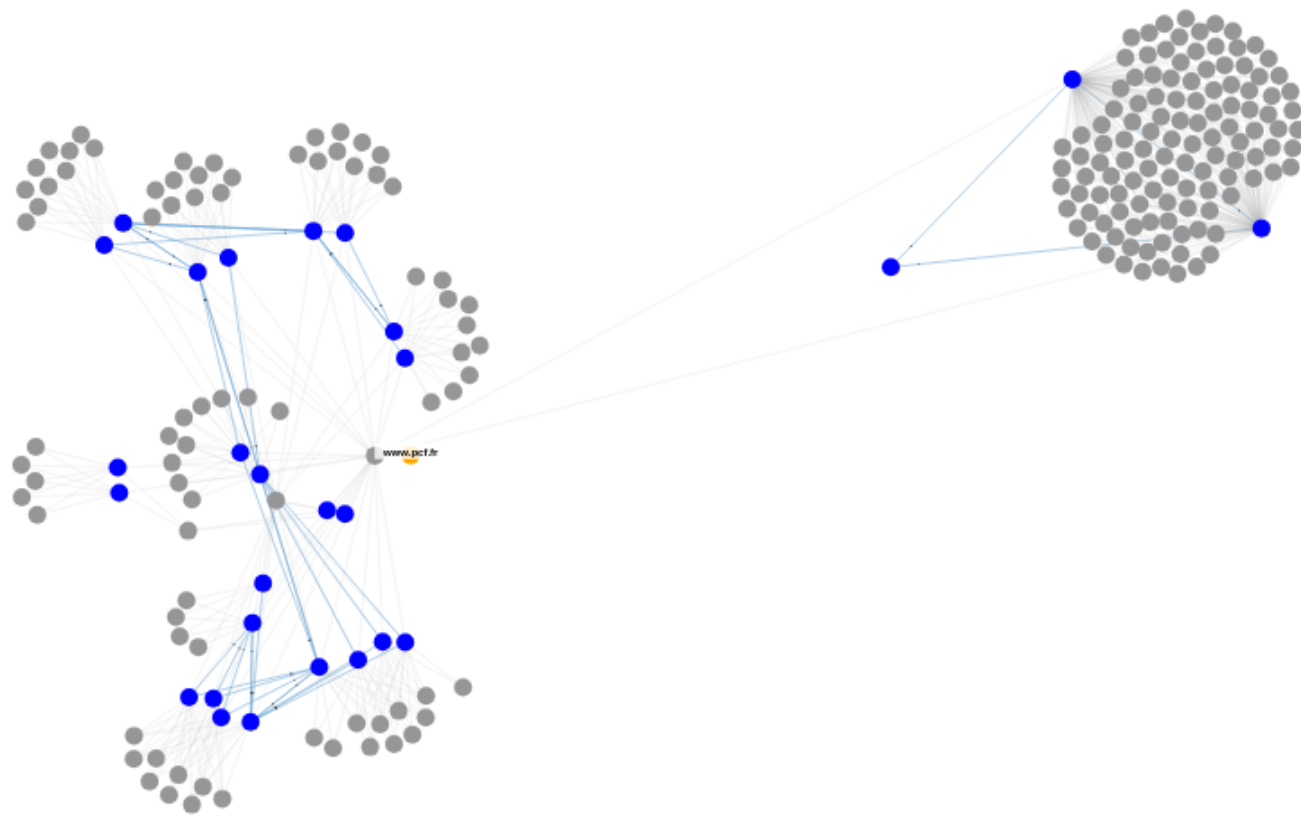
Website Type: html

Date: 2002-01-31T07:23:15Z

Short Title: {"id":"20020131072315/ISysYmageZaFNARVm5gcOw==","collections":["élections"],"links":["http://roberthue2002.net/index-3.html","http://roberthue2002.net/mot.php3?"]}



<> TYPE – Explorer des corpus avec PANDORÆ



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

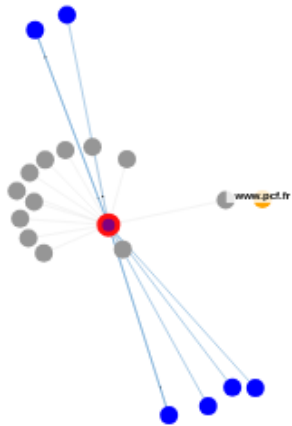
Hide solo ghosts

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

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

ynHgTRZoU7jllgtH8oZl0xz5lvjSz+CZv+a90JU4S

ssdeep_hash_bs_1536

ynyRZoUt2/8oZmvSz+ib90JnS

version

1779471020176965600

content

Collège exécutif Particomunisme français exco
dossiers:Collège exécutifDéclar
(notes,etc.)DiscoursCommuniquésCommunistes
journalContenu du dossier Collège exécutif.préc
exécutif du 06 juin 2001Collège exécutif du 29 m

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.



Dorothee
Benhamou-Suesser



Guillaume Levrier



Closed data, open software
Building new ways into the
French web archives