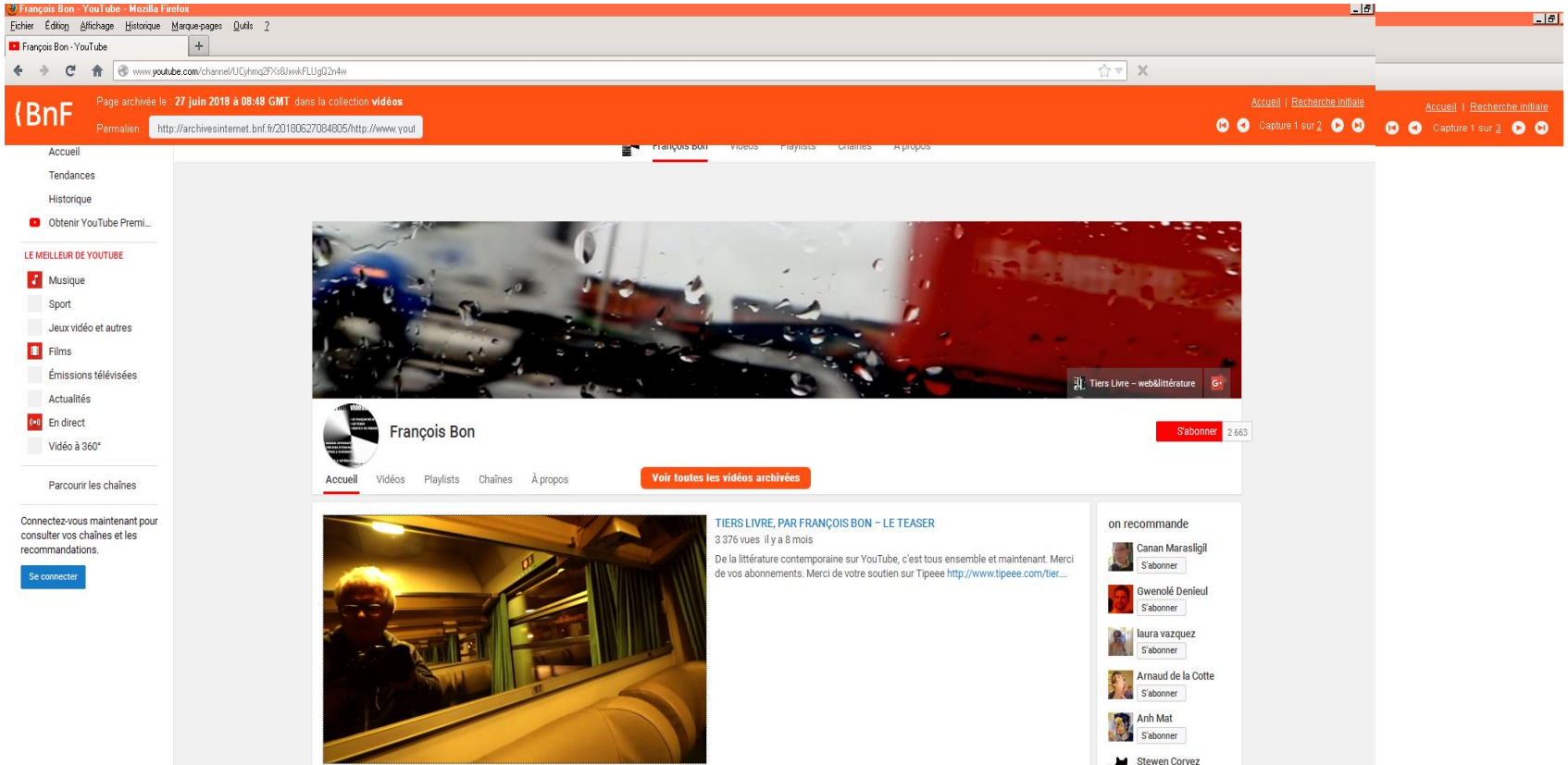
From videos to channels: archiving video content on the web

Géraldine Camile, BnF

Web archiving conference, IIPC, 6 juin 2019

Since the begining

- Our goal is to :
  - crawl and to give access the videos as they are published on the web
  - crawl channels content and not individual videos



In the past…

Period	Crawls	Crawl tools	Access	Preservation
2008	- Dailymotion - 1rst BnF specific videos crawl	- Heritrix 1 - No specific technical parameter	In the Wayback, videos are visible in the original page	no
2009-2013	- Dailymotion - BnF videos crawl	- Heritrix 1 - Specific technical parameters to identify the videos URLs and bypass the limited duration of the URLs - NAS-Heritrix 1 since 2011	In 2013 - a FLV player replaces the dailymotion player. - the first format that the Wayback finds in its indexes - Specific indexation to ignore the key videos URLs	Only crawls made since 2011

- The main limits :
  - An increasing weigh which forced us to limit the number of videos per channel
  - Several formats for a same videos

Since 2017…

Period	Crawls	Crawl tools	Access	Preservation
2017	- YouTube *- 28 channels for presidential elections* - Subcontract with Internet Memory Research	- NAS- Heritrix 3 for the web pages - API services for metadata, images and videos	In 2018 - A data base in the wayback to make the link between the pages URLs and the videos URLs - A new flash player - A guided path to recompose the videos list	Ongoing development
2018	- YouTube *- 42 channels* - BnF videos crawl	- NAS- Heritrix 3 - YouTube-dl		yes

Today : 2 crawls per year

- A process included in our standard workflow
  - 10TB for 2019
  - Selection of a representative set of French channels on YouTube
    - on various topics (13)
    - according to department collection policy
    - with an important visibility on the web

- Have a comprehensive crawl of channels
  - first, crawl all videos available on the channel at crawl time
  - then, only crawl the new uploaded videos

First crawl of 2019

- 209 channels, 34 days, 3.5TB
- 37 403 new videos

(BnF

Configuration

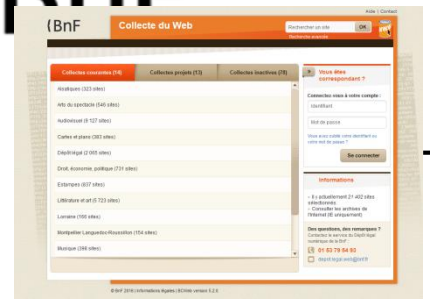
- A video profile
  - It rejects everything thing EXCEPT a defined YouTube channel list, video pages and video files
  - Filters on generic YouTube channels and other unwanted patterns

- The seeds lists
  - Seeds.txt = channels + navigation tabs
    - from Bcweb, our selection tool
  - Seeds-videos.txt = list of video pages to crawl
    - Extracts from youtube-dl at start of crawl
    - Deduplication of the video pages already crawled

Crawl specificities

- One job
  - First, crawl of the web pages
  - Then, crawl of the videos
    - MP4/best resolution
    - beanshell script used by Heritrix to execute youtube-dl during the crawl and get the video download URL for a specific page
    - In small batch to have time to crawl the videos before the URLs expiry time

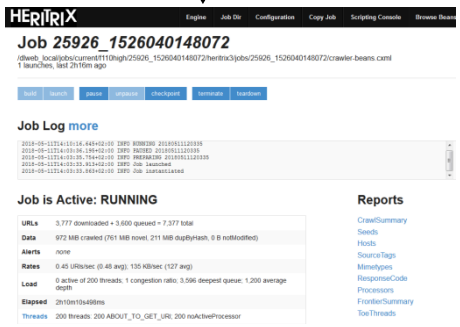
- Logs and reports
  - scripting_events.log
  - videos-detailed-list.txt
    - "[ChannelID]","[PageURL]","[UploadDate]","[Duration]","[Title]«
    - "UCFUeap7qpdvBetmpPvkhbFw","https://www.youtube.com/watch?v=0008_SlkRAg","20170921","183","Bande-annonce 2017 | Festival International du Film d'Aventure de La Rochelle"
  - videos-report.txt
    - "[PageURL]","[InitialVideoDownloadURL]","[FinalVideoDownloadURL]","[VideoFormat]","[ResponseCode]","[YoutubeDLFeedback]"



Selection of channels (navigation tabs)

Harvest Def. Videos

Initial H3 video template video
Initial seed list (seeds.txt)
Complementary videos seed list (seeds-videos.txt)
A single job



rg3 / youtube-dl **jq**

1) Metadata extraction (a JSON file for each channel)
2) Preparing the comprehensive list of video web pages to crawl :
videos-detailed-list.txt
seeds-videos.txt

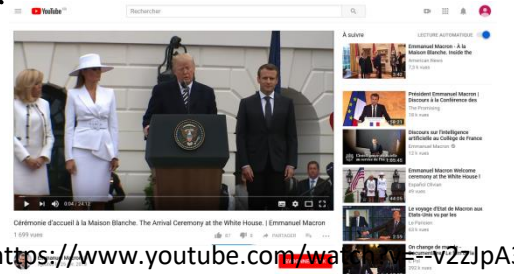
manage-youtube-videos

PAUSE
Phase 1 : crawl of navigation tabs for all channels
Adding positive filters matching seeds.txt + scripting_events.log
PAUSE
Phase 2 : iterative crawl of a batch of 50 video web pages + video files Changing positive filters to match the batch of video web pages + scripting_events.log
FINAL PAUSE
Consolidates videos-report.txt

mk_videos-report

ExtractorVideo.bsh
HERITRIX rg3 / youtube-dl

Execute youtube-dl to get the video download URL (MP4/best resolution)
Build videos-report.txt + scripting_events.log



https://www.youtube.com/watch?v=V2zJpA3PU

Access in OpenWayback

← Retour à la liste des parcours

Des vidéos à la chaîne, un état des lieux

Date de publication : 01/09/2018

Présentation
Thème
art
arts du spectacle
cinéma
culture
éducation
information
jeux vidéo
littérature

Thème : cinéma

Filmsactu

https://www.youtube.com/channel/UC_i8X3p8oZNaik8X513Zn1Q

Créée en octobre 2009 par Mixicom (jeuxactu.com), la chaîne Filmsactu compte en juillet 2018 plus de 2 300 000 abonnés. Elle propose des bandes annonces, des critiques de films et de séries télévisées, ainsi que des avis sur les éditions DVD et Bluray.

» Archive

Le fossoyeur de films

https://www.youtube.com/channel/UCwbV8cTR4yBgFdfa_BXV2OA

À partir de septembre 2012, François Theurel propose des analyses de l'histoire du cinéma, notamment des films de genre. Avec près de 700 000 abonnés en 2018, il est le youtubeur français le plus suivi sur le cinéma.

» Archive

TV Festival de Cannes

https://www.youtube.com/channel/UC-PTqxIdpslwlcaj_0dpKQA

La webtélévision officielle du Festival du film de Cannes, avec photos calls, conférences de presse, montées des marches, interviews... Sur YouTube depuis avril 2008, elle compte plus de 51 000 abonnés en 2018.

» Archive

BnF Archives de l'internet
A propos | Aide
https://www.youtube.com/channel/UCwbV8cTR4yI
Rechercher l'URL
» Recherche avancée
Parcours guidés
» Presse et actualité
Recherche par URL
Titre : Le fossoyeur de films
URL : https://www.youtube.com/channel/UCwbV8cTR4yBgFdfa_BXV2OA
Date : du 1 juin 2018 au 15 fév. 2019
Collection : vidéos
1 résultat
» Modifier la recherche
Tout ouvrir | Tout fermer
2019
0 résultat
2018
1 résultat
jan. fév. mar. avr. mai juin jui. août sep. oct. nov. déc.
27

BnF
Page archivée le : 27 juin 2018 à 08:36 GMT dans la collection vidéos
Permalien : http://archivesinternet.bnf.fr/20180627083619/http://www.youl
Accueil | Recherche initiale
Capture 1 sur 1

Le Fossoyeur de Films

S'abonner 693 477

Vidéos Playlists Communauté Chaînes À propos

Voir toutes les vidéos archivées

OPÉRATION 5 ANS !

292 614 vues il y a 9 mois

Aujourd'hui, la chaîne a 5 ans.
CÉLÉBRONS !

Musiques extraites de :
- Cannibal Ferox (Umberto Lenzi, 1981)
- Ninja Terminator (Godfrey Ho, 1986)
- Virus Cannibale (Bruno Mattei, 1980)
- L'homme Puma (Alberto de Martino, 1980)

...

Lire la suite

Activités récentes

Hola amigos.

NEXUS VI S'abonner

StudioUha S'abonner

Axolot S'abonner

La Brigade du Livre S'abonner

ALT 236 S'abonner

Zone 42 S'abonner

Hokku S'abonner

Vidéos archivées

- Titre : Le fossoyeur de films
- URL : https://www.youtube.com/channel/UCwbV8cTR4yBgFdfa_BXV2OA
- URL : https://www.youtube.com/user/deadwattsofficiel
- Collection : vidéos

214 résultats

Résultats 1 à 60 Mot(s) du titre OK Date de publication ↓ 1 2 3 4 > >>

LES PLANS WTF 10min11s
27-05-2018

L'homme qui tua Don Quichotte : ENFIN ! 11min54s
25-05-2018

FDF #35 - Making of / bêtisier ! 2min23s
14-05-2018

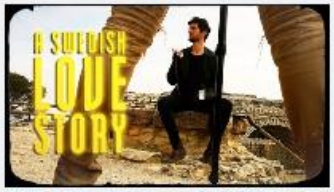
LE FOSSOYEUR DE FILMS #35 - Prince des ténèbr... 17min21s
07-05-2018

AVENGERS 3 - Bonjour spoilers (duh) 23min30s
02-05-2018

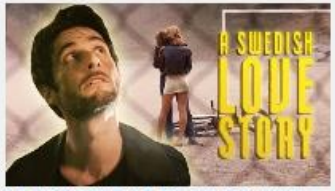
PANDORA'S BOX #2 - Ready Player One 1h15min35s
12-04-2018

Le cinéma de genre est enfin respecté ? 11min18s
04-04-2018

ANNIHILATION - La lecture cachée 13min42s
25-03-2018



FDF #34 - Making of / bêtisier ! 2min29s
14-03-2018



LE FOSSOYEUR DE FILMS #34 - Une histoire d'am... 19min08s
08-03-2018

BnF

BnF Page archivée le : 28 juin 2018 à 05:22 GMT dans la collection vidéos
Permalien : http://archivesinternet.bnf.fr/20180628052206/http://www.youl
Accueil | Recherche initiale
Capture 1 sur 1

L'homme qui tua Don Quichotte : ENFIN !

Le Fossoyeur de Films
S'abonner 693 k

168 868 vues

Ajouter à Partager Plus 7 026 59

Ajoutée le 25 mai 2018
Quand un film qu'on a rêvé pendant plus de 20 ans apparait enfin, la réalité est-elle à la hauteur du fantasme ? VERDICT.

PLUS

À suivre Lecture automatique

LE FOSSOYEUR DE FILMS #30 - Le Nom de la Rose
Le Fossoyeur de Films
589 509 vues

5 preuves que "Gotham" est débile - FERMEZ LA
MJ - FERMEZ LA
240 372 vues

LES PLANS WTF
Le Fossoyeur de Films
264 456 vues

PVR #39 - QUI A TUÉ RARE ? LE STUDIO MYTHIQUE DES FRÈRES STAMPER
La chaîne de P.A.U.L
84 080 vues

La trilogie du Hobbit de Peter Jackson : l'analyse de M. Bobine
Le ciné-club de M. Bobine
24 854 vues

Bernard Minet - Pourquoi ça existe !? - Ermite Moderne
L'Ermite Moderne
194 637 vues

Access specificities

- Specific developments to:
  - Browse and display thumbs channel pages (URL rewriting)
  - Insert « View all archived videos » button
  - « All archived videos » page (search, sort, page)
  - Fix video page
  - Replay video with a customised archived Flash player (flvplayer.net)
  - Specific database to make the link between the pages and the URLsvideos, to add a channel information

Perspectives

- Automate the maximum of manual tasks
- Update our « guided path » with 2009-2013 Dailymotion collection (early web videos)
- Try to look beyond:
  - harvest embeded videos
  - harvest other plateforms (Vimeo, Dailymotion)