



Archivage du web à l'Ina



Les images qui vous parlent

PLAN

1. Présentation de l'Institut National de l'Audiovisuel
2. Le dépôt légal: définitions et applications
3. Le media web: quelques rappels
4. Chaîne d'archivage du web
5. Techniques de collecte du web
6. Aspects documentaires de l'archivage du web
7. Mise en consultation de l'archive du web

PARTIE 1

L'Institut National de l'audiovisuel

1974 : CRÉATION DE L'INA

Après l'éclatement de l'ORTF, l'Ina hérite

- de la recherche
- de l'archivage des chaînes publiques
- de la formation professionnelle

1992 : UNE INSTITUTION PATRIMONIALE

Extension du dépôt légal à la radio et la télévision

1999 : LE VIRAGE NUMÉRIQUE

- Plan de Sauvegarde Numérique jusqu'en 2015, 883 200 œuvres en danger à numériser;
- 2004 : InaMediaPro.com (1M d'heures avec droits de producteurs Ina);
- 2006 : Ina.fr, portail grand public
- 2010: Ina Global, revue en ligne sur les acteurs et les évolutions des industries créatives, des médias et du web.

- Etablissement public à caractère industriel et commercial
- Président: Laurent Vallet (mandat de 4 ans)
- Chiffre d'affaire : 37,9 millions d'euros (2017, +5%)
- Environ 1000 collaborateurs
- Répartition homme-femme : 478 femmes pour 473 hommes
- 4 sites en Île de France et 6 antennes en régions (Toulouse, Lyon, Strasbourg, Rennes, Marseille, Lille)
- Des métiers variés :
 - liés à l'audiovisuel: étalonneurs, acousticiens restaurateurs, monteurs, producteurs...
 - liés à l'archivage et à la documentation: archivistes, documentalistes multimédia...
 - liés à la formation : formateurs, experts...
 - liés aux fonctions supports: DSI, RH, juristes, acheteurs...

PARTIE 2

Le dépôt légal

Histoire et acteurs du dépôt légal français

Objet :

Constituer une collection de référence pour la mémoire collective du pays.

Historique :

28 décembre 1537: première ordonnance royale de François 1^{er}

20 juin 1992 : extension aux diffuseurs Radio/TV

1995: premiers dépôt à l'Ina d'émissions par les diffuseurs hertziens

2006: décret d'extension du périmètre aux sites web

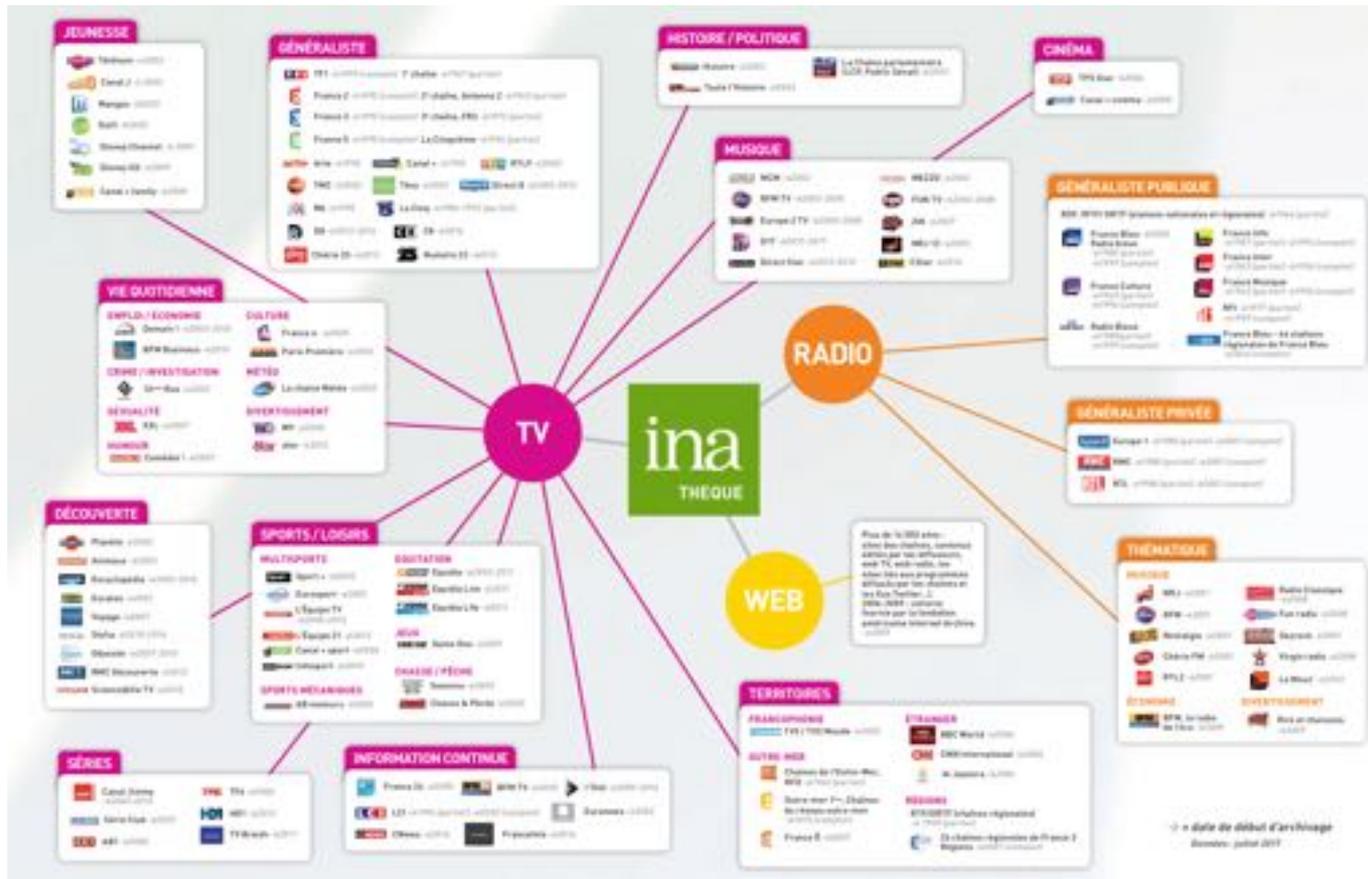
Acteurs :

- Bibliothèque nationale de France
- CNC
- Ministère de l'Intérieur
- Ina

Fonctionnement à l'Ina:

- captation numérique par liaisons satellites et fibres optiques
- 24h/24 7j/7
- 164 chaînes de télévision et de radio.
- 15 000 000 heures de programmes (1 000 000 heures / an)

Paysage du dépôt légal



La loi sur les Droit d'auteurs et droits voisins dans la société de l'information (DADVSI) du 1er août 2006, a apporté dans son titre IV quelques modifications au code du patrimoine.

*« Les logiciels et les bases de données sont soumis à l'obligation de dépôt légal dès lors qu'ils sont mis à disposition d'un public par la diffusion d'un support matériel, quelle que soit la nature de ce support. Sont également soumis au dépôt légal les signes, signaux, écrits, images, sons ou messages **de toute nature faisant l'objet d'une communication au public par voie électronique.** »*

Dépôt légal du Web : périmètre et acteurs

Le décret du 19 décembre 2011 modifie le code du patrimoine et « fixe les conditions de sélection des informations collectées sur Internet par la BnF et l'Ina au titre du dépôt légal »



Services de communication au public par voie électronique

- édités par les chaînes de radio et télévision;
- portant essentiellement sur les contenus de la radio et de la télévision;

{ BnF

Le domaine français à l'exception de ceux collectés par l'Ina

PARTIE 3

Le média Web

Quelques rappels

Le web: un média jeune

1989 : création du système hypertexte (Tim Berners Lee – CERN)

1991 : création officielle du projet WorldWideWeb

1992 : création du langage HTML

1993 : le projet WorldWideWeb passe dans le domaine public

1994 : création de Yahoo

1998 : création de Google

2001 : création de Facebook

2004 : création du Web 2.0

2006 : création de Twitter

2007 : création du HTML5

2008 : lancement du navigateur Chrome



Premier serveur Web



Data center Google

Le web: quelques principes

- Web (*application*) ≠ Internet (*structure*)
- Ensemble de **ressources autonomes** connectées entre elles
- Chaque ressource est localisée par une **URL** (*Uniform Resource Locator*) pouvant être rendue accessible par une **hyperlien**.
- Type de ressources: images, textes, vidéos, sons...
- Repose sur:
 - un protocole : HTTP
 - des langages: HTML, Java, Javascript, php, SQL, CSS
 - des standards ([W3C](#))



Le web: face visible et face cachée

The screenshot shows the ina.fr website with a network developer tool (likely Chrome DevTools) open on the right. The tool displays a list of 405 requests, including various images, CSS files, and scripts. The status bar at the bottom of the tool indicates: 405 requests | 9.2 MB transferred | Finish: 39.46 s | DOMContentLoaded: 6.18 s | Load: 7.33 s.

Name	Method	Status	Type	Initiator	Size	Time	Waterfall
fonda-jane_140x105.jpg	GET	200	jpeg	(index)	4.3 KB	73 ms	
aleveque-christophe_140x105.jpg	GET	200	jpeg	(index)	4.6 KB	74 ms	
seigner-mathilde_140x105.jpg	GET	200	jpeg	(index)	4.2 KB	74 ms	
boyer-laurent_140x105.jpg	GET	200	jpeg	(index)	3.3 KB	74 ms	
damieau-danielle_140x105.jpg	GET	200	jpeg	(index)	4.9 KB	75 ms	
higelin-jacques_140x105.jpg	GET	200	jpeg	(index)	4.1 KB	75 ms	
delair-suzy_140x105.jpg	GET	200	jpeg	(index)	4.6 KB	69 ms	
hossein-robert_140x105.jpg	GET	200	jpeg	(index)	4.0 KB	68 ms	
phrases-cultes_140x105.jpg	GET	200	jpeg	(index)	2.9 KB	67 ms	
les-actualites-francaises_140x105.png	GET	200	png	(index)	30.5...	66 ms	
theatre-de-l-etrange_140x105.jpg	GET	200	jpeg	(index)	5.1 KB	63 ms	
dim-dam-dom_140x105.jpg	GET	200	jpeg	(index)	3.7 KB	62 ms	
background_site+v2.1.png	GET	200	png	background...	502 B	62 ms	
strip-tease_140x105.jpg	GET	200	jpeg	(index)	3.8 KB	60 ms	
cinq-colonnes-a-la-une_140x105.jpg	GET	200	jpeg	(index)	3.9 KB	57 ms	
chorus_140x105.jpg	GET	200	jpeg	(index)	5.4 KB	56 ms	
cinema-cinemas_140x105.jpg	GET	200	jpeg	(index)	4.3 KB	53 ms	
age-tendre-et-tete-de-bois_140x105.jpg	GET	200	jpeg	(index)	4.4 KB	54 ms	
l-histoire-en-direct_140x105.jpg	GET	200	jpeg	(index)	5.1 KB	53 ms	
menu-premium_cndes-2.png	GET	200	png	(index)	3.6 KB	50 ms	
picto-menu-programme.png	GET	200	png	(index)	2.9 KB	49 ms	
picto-menu-abo.png	GET	200	png	(index)	3.7 KB	46 ms	
picto-menu-pub.png	GET	200	png	(index)	3.1 KB	35 ms	
picto-menu-exclu.png	GET	200	png	(index)	2.9 KB	36 ms	
picto-menu-premium-prix.png	GET	200	png	(index)	4.9 KB	35 ms	
documentaire_113x85.jpg	GET	200	jpeg	(index)	4.6 KB	33 ms	
emissions_113x85.jpg	GET	200	jpeg	(index)	3.1 KB	30 ms	
humour_113x85.jpg	GET	200	jpeg	(index)	4.3 KB	29 ms	
jeunesse_113x85.jpg	GET	200	jpeg	(index)	3.8 KB	42 ms	
musique_113x85.jpg	GET	200	jpeg	(index)	3.4 KB	41 ms	
serie-et-fiction_113x85.jpg	GET	200	jpeg	(index)	4.8 KB	39 ms	
spectacle_113x85.jpg	GET	200	jpeg	(index)	4.0 KB	38 ms	
sport_113x85.jpg	GET	200	jpeg	(index)	4.3 KB	37 ms	
ccadeau-menu-premium.png	GET	200	png	(index)	23.0...	37 ms	
js?lang=fr&site=ina.fr&za=homepage&id=7111778...	GET	200	script	(index)	5.8 KB	320...	
logo-ina.png	GET	200	png	(index)	5.5 KB	33 ms	
s704446_001-brigitte-bardot-proie-mediatique...	GET	200	jpeg	(index)	21.5...	34 ms	
i00018526-brigitte-bardot-a-propos-de-la-vivise...	GET	200	jpeg	(index)	7.5 KB	32 ms	
...	

1 simple web page = 400 resources

Le web: un média éphémère...

...soumis à la disparition:

- Espérance de vie d'une page Web : 44 jours
- Permanence d'un lien 56h en moyenne pour un titre de presse (étude menée pour "Chicago Tribunal" - 2011)
- Durée de vie d'un tweet: 4h04
- Durée de vie d'un post Facebook: 14h42
- Durée de vie d'une photo Instagram: 21h36

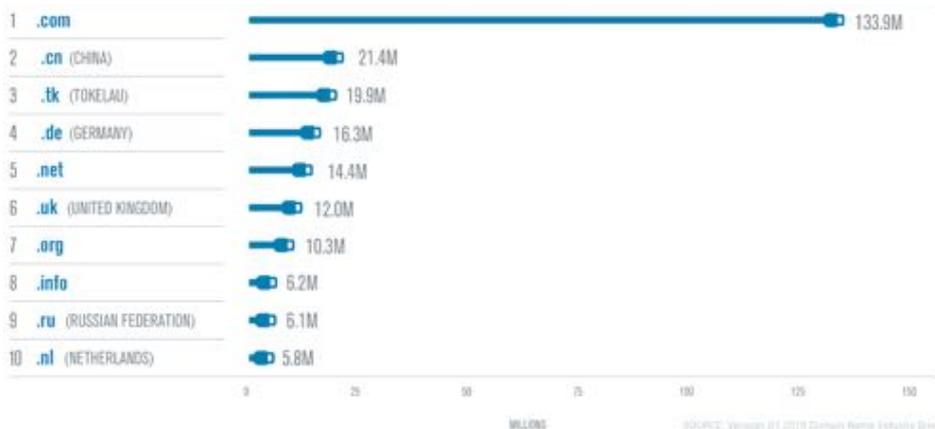
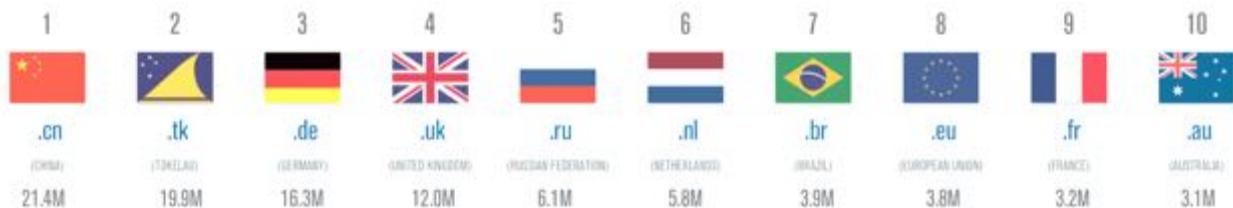
...et à la modification:

- mises à jour des sites régulières:
 - du contenu : textes, medias...
 - du contenant: éditorialisation, mise en page...
- changement des URLs
- modification des noms de domaines

...inscrit dans l'espace...

334 millions de domaines :

- .com: +3 % par an ([source](#))
- .fr: + 6% par an ([source AFNIC](#))



Source: VeriSign Q1 2019 Domain Name Industry Brief

...et dans le temps

2017 *This Is What Happens In An Internet Minute*



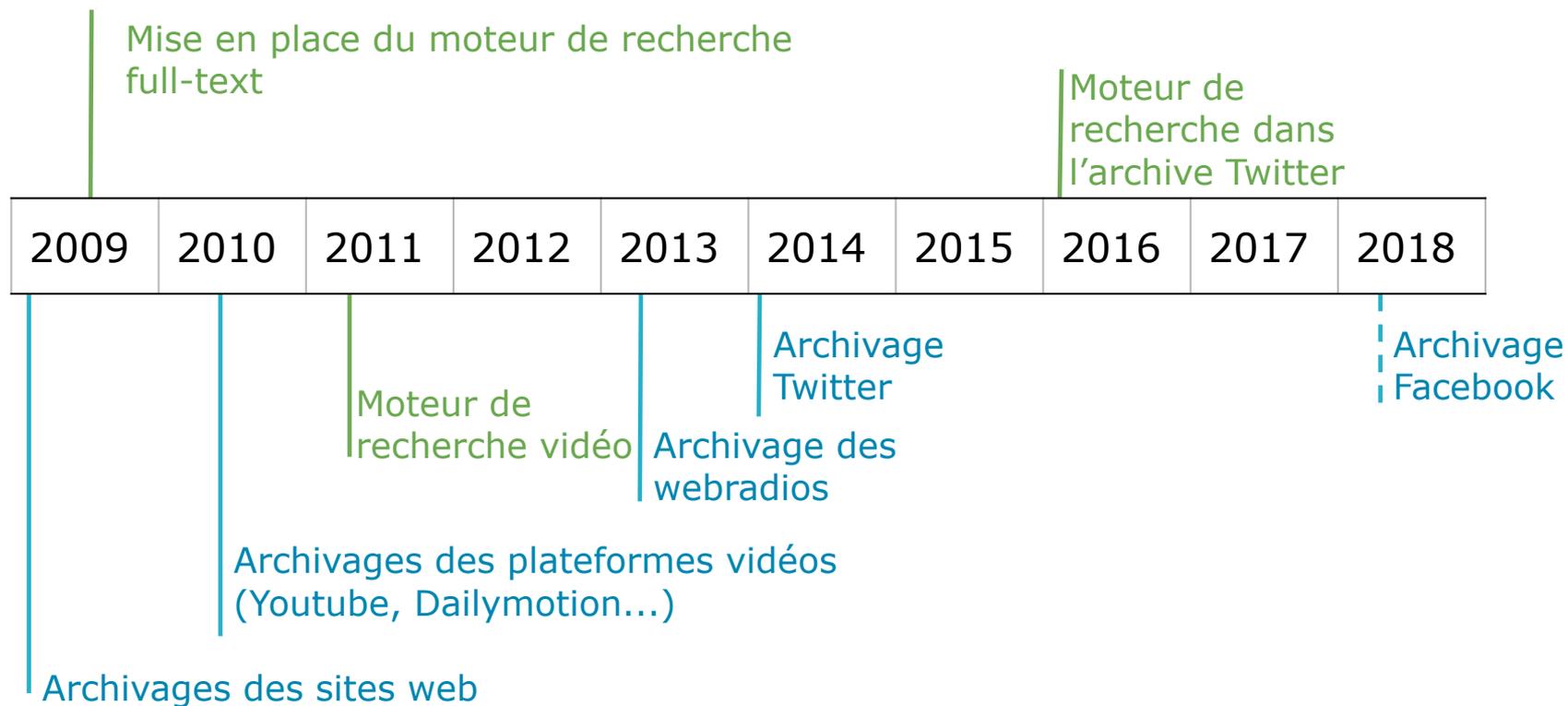
2018 *This Is What Happens In An Internet Minute*



PARTIE 4

Chaîne d'archivage du web

Historique de l'archive du web à l'Ina



Rouages de la chaîne d'archivage du web



Chaîne d'archivage: sélection

Quoi :

→ Tous les objets web en liens avec l'audiovisuel francophone :

- Sites web des groupes media (www.tf1.fr ...)
- Sites personnels liés aux médias (www.plusbellelavie.org)
- Comptes de réseaux sociaux liés aux médias (@CashInvestigations, @TPMP)
- UGC: Chaînes Youtube, Dailymotion, Vimeo francophones (Norman fait des vidéos)
- Hashtags liés aux médias (#LMP) et aux sujets d'actualités (#MeToo)

Qui :

→ L'équipe documentation

Comment :

→ Outils dédiés maisons (Excel)

→ Veilles sur le web: Hootsuite, Twitter, Talkwalker...

Chaîne d'archivage: documentation

Quoi :

→ Une partie des objets web archivés :

- Sites web
- Comptes réseaux sociaux
- Chaînes Youtube, Dailymotion, Vimeo
- Hashtags

→ Ne sont pas documentés:

- Pages web
- Tweets
- Vidéos

Qui :

→ L'équipe documentation

Comment :

→ Outils dédiés maisons

Chaîne d'archivage: collecte

Quoi :

→ Développer :

- les outils et les méthodes de collecte
- les formats et les outils de stockage
- les méthodes d'accès : fouille et indexation

Qui :

→ L'équipe Recherche & Développements

Comment :

→ Outils dédiés maisons

→ Solutions open-source: Elastic Search, Hadoop, Kibana...

Chaîne d'archivage: stockage

Quoi :

- Mettre en œuvre les outils R&D
- Assurer le monitoring et l'analyse des serveurs de stockage
- Etablir la stratégies de préservation et consultation à long terme
- Gérer les aspects technique de la consultation

Qui :

- L'équipe Exploitation

Comment :

- Solutions open-source

Chaîne d'archivage: valorisation

Quoi :

- Publier les résultats des travaux de R&D
- Proposer des corpus thématiques aux usagers
- Présenter les différents aspects de l'archivage du web

Qui :

- L'équipe Recherche et développement
- L'équipe Documentation
- L'équipe Exploitation
- L'Inathèque

Comment :

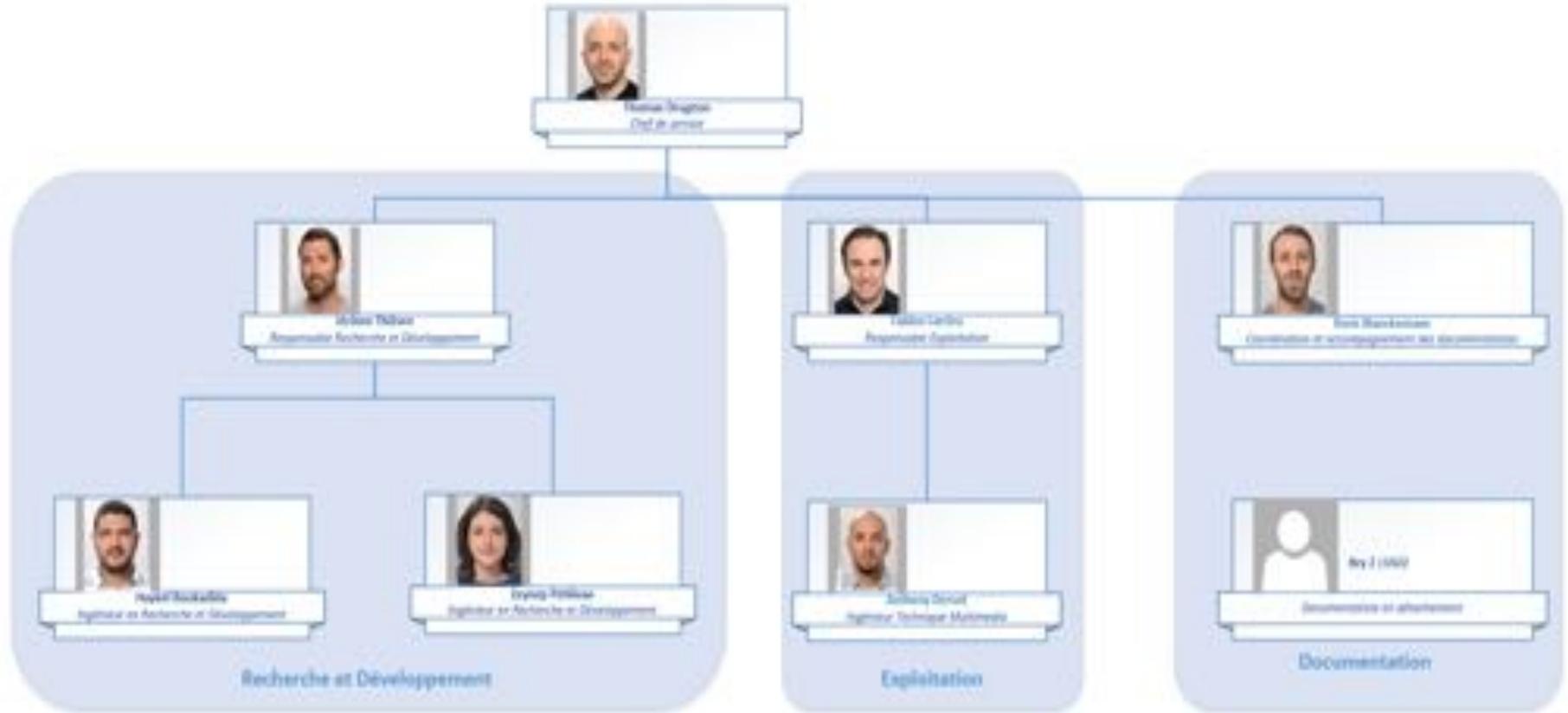
- Conférences, séminaires, formations, ateliers...
- Communications sur Twitter ([@inadlweb](https://twitter.com/inadlweb)), LinkedIn, GitHub...

- Prise en charge complète et autonome d'un maximum des aspects de l'archivage du web: stockage, collecte, infrastructure...

- Développements maison:
 - agilité
 - réactivité
 - autonomie logicielle (pas de dépendances vis-à-vis des éditeurs tiers)
 - maîtrise des coûts

- Créer, développer et maintenir des passerelles avec les autres métiers/services de l'Ina

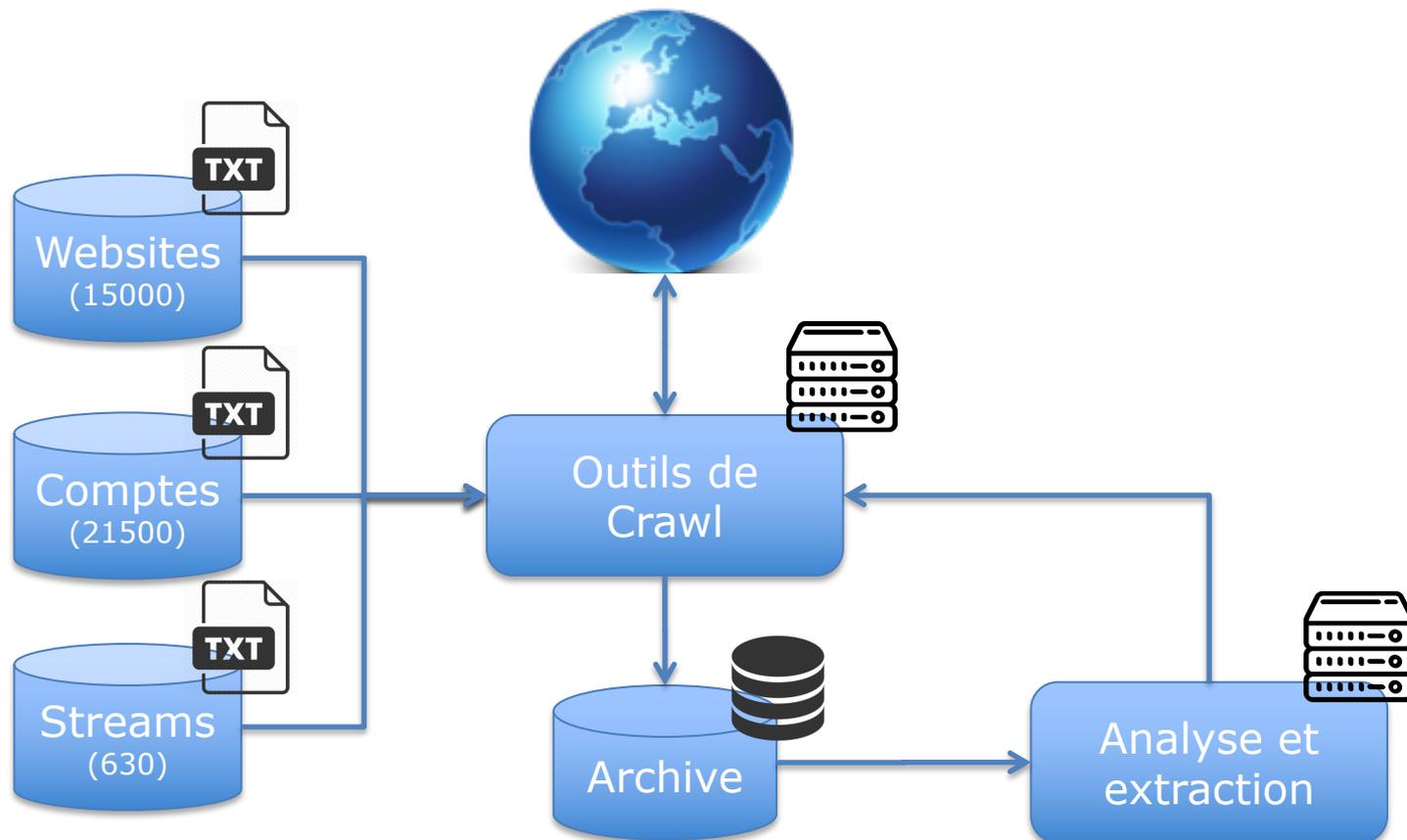
Équipe en charge de la collecte



PARTIE 5

Techniques de collecte du web

Principe général de la collecte



Définition :

- Crawler: programme qui parcourt et enregistre:
 - l'ensemble des liens HTML d'un site et des ressources vers lesquels ils mènent : médias, CSS...
 - une partie des interactions entre le client (poste de l'utilisateur) et serveur (machine hébergeant le site web): javascript, html5

Principes :

- Principe de sélection
- Principe de re-visite
- Principe de politesse
- Principe de parallélisation/coordination

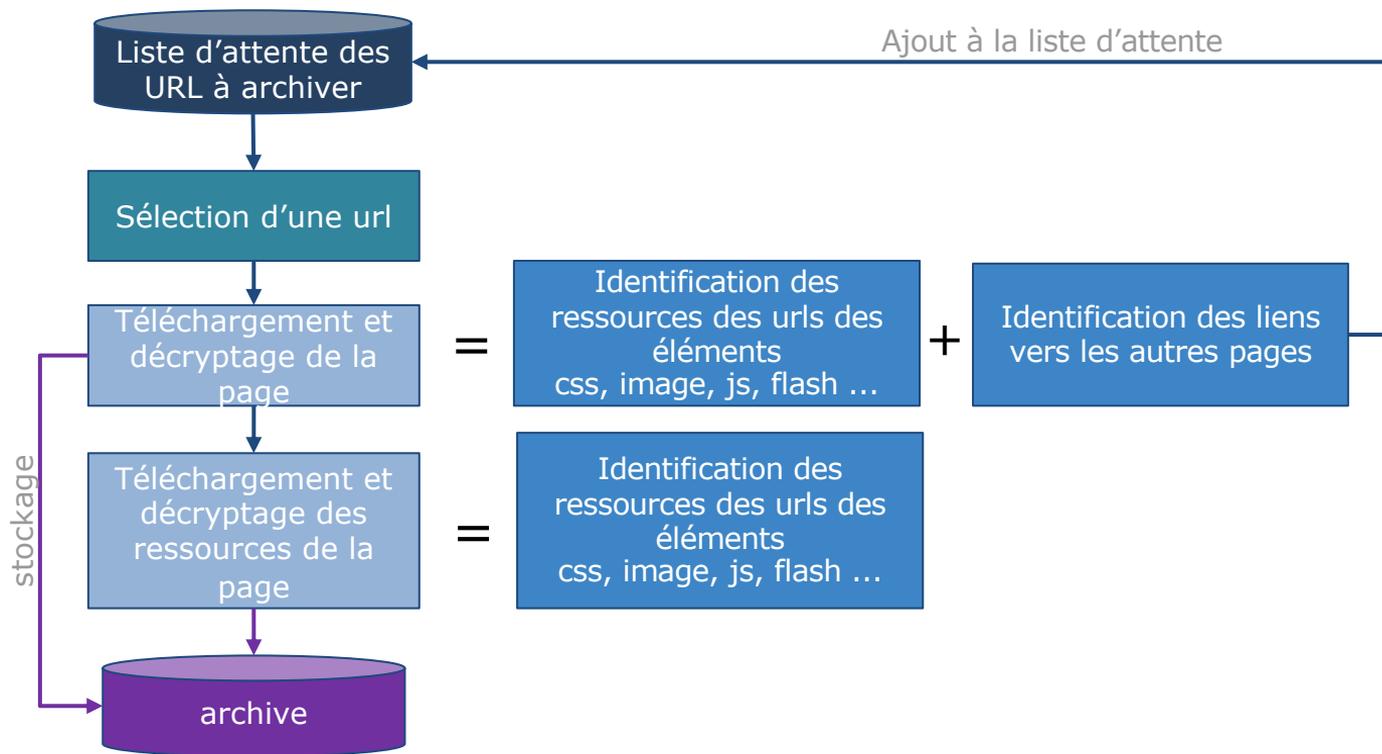
Exemples :

- Heritrix (Internet Archive)
- Qwantify (Qwant)
- Googlebot (Google)

Limites :

- Contenu manquant :
 - Certaines interactions (moteur de recherche...)
 - Certains paramètres de navigation (calendrier perpétuel...)
- Inconsistance temporelle :
 - mises à jour de pages manquantes
 - lien menant à une page à une autre date

Robot de collecte: fonctionnement



Organisation des robots de collecte

→ Un ordonnanceur gérant plusieurs robots

→ Chaque robot archive un site à la fois

→ Plusieurs fréquences d'archivage :

- Pages d'accueil crawlées toutes les heures
- Pages de niveau 1 crawlées plusieurs par jour
- Pages profondes crawlées plusieurs par jour

→ Plusieurs types de robots:

- Robots génériques, légers et simples à mettre en œuvre
- Robots lourds pour les pages plus complexes
- Robots dédiés à un site ou un type de site

→ Robots développés en mode agile.

Spécialisation des robots de collecte

→ Phagosite: robot générique, le plus utilisé

→ Crocket: robot plus spécialisé

- Exécute le javascript
- Détecte les évènements utilisateurs (clic menu, défilement de page)

→ LAP: robot manuel commandé par un humain

→ Utilisé pour les sites à l'interactivité très forte (web documentaires...)

Format d'archive de la collecte

- **DAFF** : Digital Archiving File Format
- Format conteneur
- Format simple et extensible
- Auto-décrit: contient les données et les métadonnées
- Agnostique aux protocoles
- Contrôle d'intégrité implémenté : utilisation d'une fonction de hashage
- Non ISO

Hachage

→ Permet d'identifier de manière unique un contenu numérique

Exemples:

Données en entrée	Signature
123	202cb962ac59075b964b07152d234b70
124	c8ffe9a587b126f152ed3d89a146b445
1234	81dc9bdb52d04dc20036dbd8313ed055
123456789012345678901234567890	a46857f0ecc21f0a06ea434b94d9cf1d
abcde	ab56b4d92b40713acc5af89985d4b786
abcdef	e80b5017098950fc58aad83c8c14978e

→ Basé sur des calculs mathématiques

→ Ne fonctionne que dans un seul sens (impossible de retrouver le contenu à partir de sa signature)

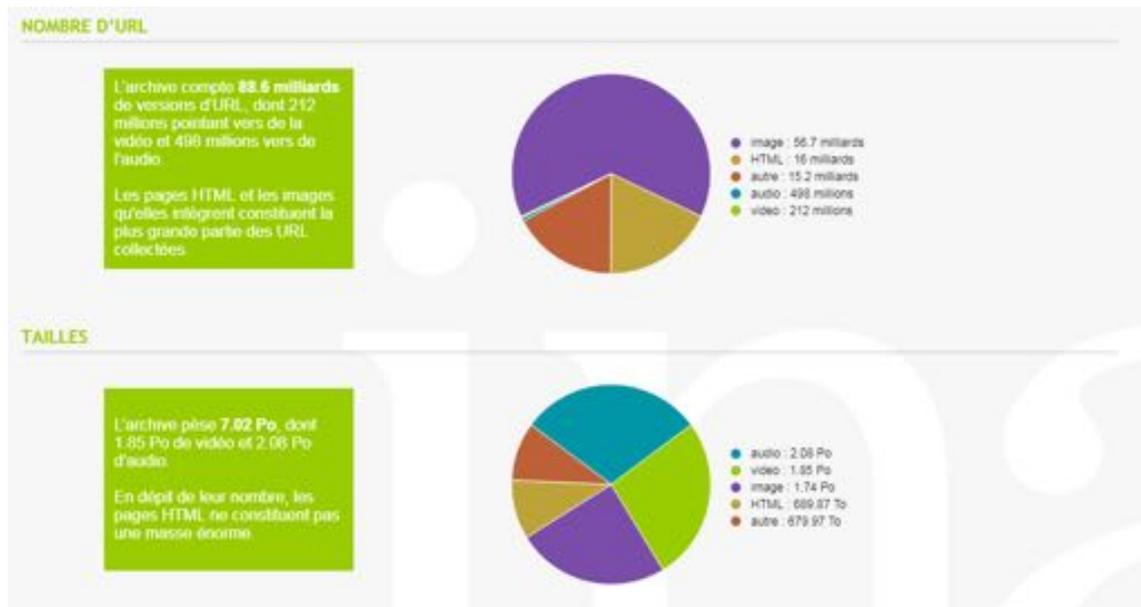
→ Plusieurs fonctions de hachage: md5, sha (*Secure Hash Algorithm*)

→ DAFF: sha256

Le format DAFF en chiffres

➤ 88,6 milliards d'enregistrements

➤ Taille de l'archive: 7,02 Po



Collecte des UGC: états des lieux

Depuis 2010: 24 millions de contenus archivés



vimeo



dailymotion

CULTUREBOX



**francetv
pluzz**

**arte
CONCERT**

wat

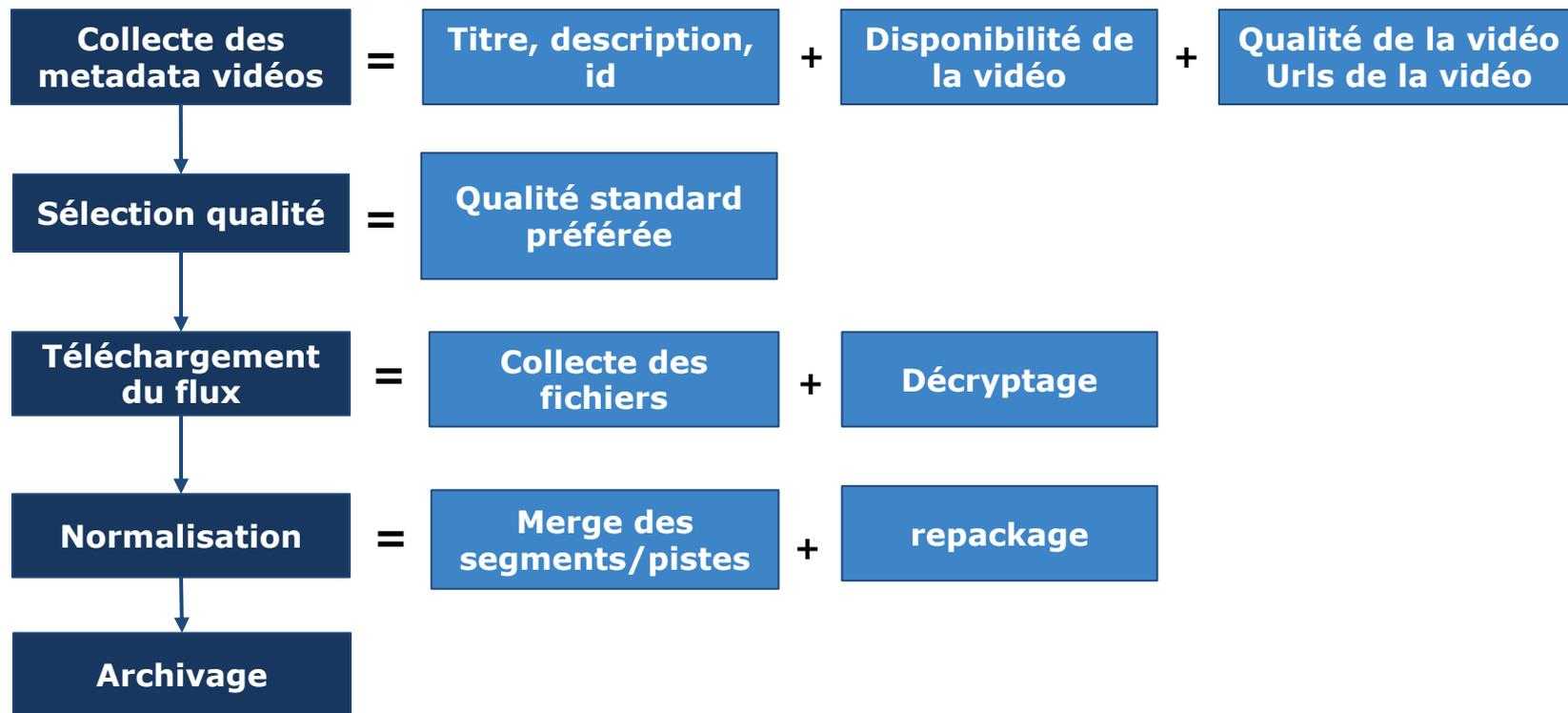
Spécificité des vidéos sur le web :

- plusieurs *providers*
- plusieurs formats
- volumineuses
- pas de mises à jour
- difficile à détecter

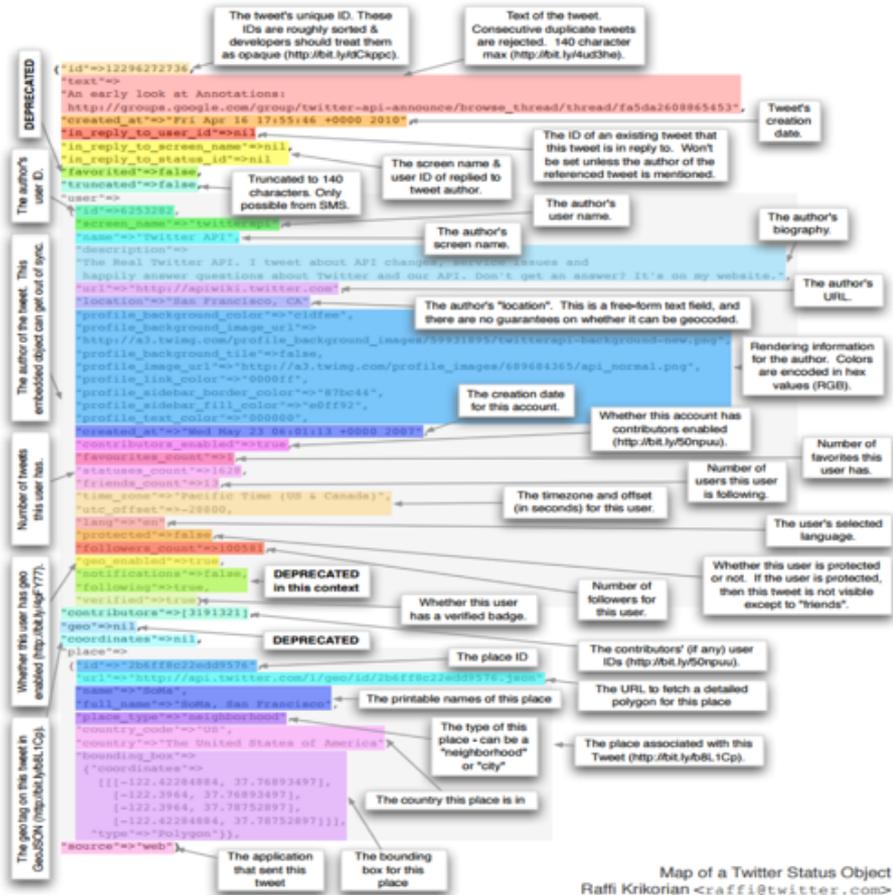
Sources de collecte :

- à partir des pages web
- à partir des tweets
- à partir des API (Dailymotion, Youtube)

Collecte des UGC: principe



Collecte des tweets : définition



- 1 tweet = 30 metadatas
- Texte du tweet : 5% des données du tweet

Map of a Twitter Status Object
Raffi Krikorian <raffi@twitter.com>
18 April 2010

Collecte des tweets : états des lieux

Depuis 2014:

- 1 milliard de tweets archivés

- 15 000 comptes et 900 hashtags liés à:
 - aux média radio/tv
 - Événements nationaux importants (élections présidentielles, jeux olympiques...)

- ≈ 270 000 tweets archivés par jour (hashtags)

- ≈ 48 000 tweets archivés par jour (timeline)

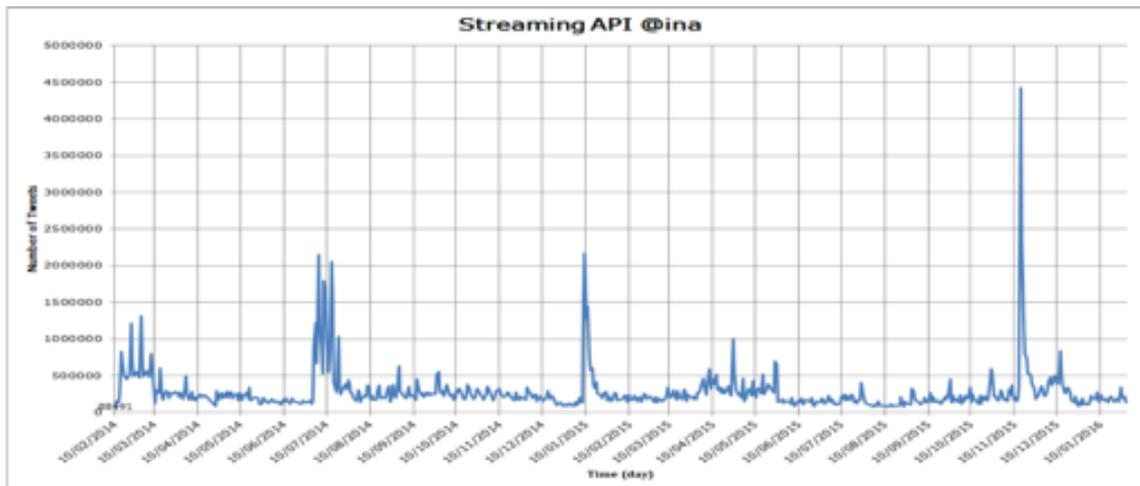
- Croissance: ≈ 100 comptes et 10 hashtags par mois

Collecte des tweets : principe de collecte

→ Utilisation de l'API publique twitter

Limitations:

- 400 hashtags et 5000 utilisateurs maximum
- 1% des tweets publiés à l'instant
- 3200 derniers tweets
- Uniquement les tweets de moins de 7 jours



PARTIE 6

Aspects documentaires de l'archivage du web

→ Quoi ?

Tout contenu web en rapport avec l'audio-visuel français

→ Comment ?

- Veille en mode push
- Veille en mode pull

→ Fréquence ?

Quotidienne

→ Outils ?

Outils maison: tableur excel + macro

Veille: push vs. pull

Méthodes	Principes et outils	Avantages	Inconvénients
Push	<ul style="list-style-type: none">▪ Sur les réseaux sociaux, abonnement aux comptes "clés" :<ul style="list-style-type: none">• comptes de chaînes TV/Radio• comptes de personnalités TV/Radio• comptes d'émissions• comptes de fans▪ Surveillance des tendances▪ Sur le web: création d'alertes mails à partir de mot(s) clef(s)	<ul style="list-style-type: none">• Automatisation forte (gain de temps)• Outils existants (gratuits / freemium): Hootsuite, Talkwalker Alert• Repose peu sur l'humain	<ul style="list-style-type: none">• Bruit / résultats non pertinents• Paramétrage des outils• Sérendipité faible• Faible maîtrise du processus de recherche
Pull	Recherche dans un moteur	<ul style="list-style-type: none">• Sérendipité forte• Maîtrise du processus de recherche	<ul style="list-style-type: none">• Chronophage• Repose fortement sur l'humain

Veille : hootsuite

The screenshot displays a Hootsuite dashboard with a grid of social media feeds. The top navigation bar includes a search bar, a 'Twitter' tab, and a 'nouveau' (new) notification. The main content area is divided into several columns and rows of tweets and posts.

- Column 1 (Twitter):** A tweet from 'Dunkerque' with a photo of a building and a tweet from 'Miguel' with a graphic of hands holding pencils.
- Column 2 (Twitter):** A tweet from 'Duranid' with a bar chart showing '92%' and a tweet from 'Françoise Marty' with a bar chart showing '50%'.
- Column 3 (Twitter):** A tweet from 'Manuel Akbay' with text about national fiction, a tweet from 'Institut national de l'audiovisuel' with a 'Nous recrutons' graphic, and a tweet from 'James Blake' with a photo of a stage.
- Column 4 (Twitter):** A tweet from 'Erlis Kilbinger' with text about a hashtag, a tweet from 'Belle Nième' with a '#InaPeopleFrom' graphic, and a tweet from 'Adem Dum'.

Veille : tendances

WordCloud24 | Twitter trends

Top Twitter trends for worldwide sites

33 minutes ago	1 hour ago	2 hours ago	3 hours ago	4 hours ago	5 hours ago	6 hours ago
#Policierite	#Policierite	Q&A	Big Beer	Big Beer	Big Beer	Big Beer
سوالگاہ کیلئے	سوالگاہ کیلئے	سوالگاہ کیلئے	سوالگاہ کیلئے	سوالگاہ کیلئے	سوالگاہ کیلئے	سوالگاہ کیلئے
Crug Mack	Crug Mack	Big Beer	#Policierite	سوالگاہ کیلئے	سوالگاہ کیلئے	سوالگاہ کیلئے
Crug Mack	Crug Mack	سوالگاہ کیلئے	#Policierite	سوالگاہ کیلئے	سوالگاہ کیلئے	سوالگاہ کیلئے
Q&A	سوالگاہ کیلئے					
سوالگاہ کیلئے	سوالگاہ کیلئے	#Policierite	سوالگاہ کیلئے	سوالگاہ کیلئے	سوالگاہ کیلئے	سوالگاہ کیلئے
#Policierite	Big Beer	سوالگاہ کیلئے				
Big Beer	#Policierite	سوالگاہ کیلئے				
سوالگاہ کیلئے	سوالگاہ کیلئے	Crug Mack	سوالگاہ کیلئے	سوالگاہ کیلئے	سوالگاہ کیلئے	سوالگاہ کیلئے
names	names	Crug Mack	سوالگاہ کیلئے	سوالگاہ کیلئے	سوالگاہ کیلئے	سوالگاہ کیلئے
Crug Mack's Subira	Crug Mack	Crug Mack	سوالگاہ کیلئے	سوالگاہ کیلئے	سوالگاہ کیلئے	سوالگاہ کیلئے

Find your Jesus Ratio

Enter your twitter username

Trend TagCloud for WordCloud

Twitter Facebook Google+

Today's Twitter trends at countries worldwide

Algeria Argentina Australia Austria Bahrain Belarus Belgium Brazil Canada Chile Colombia Denmark Dominican-Republic Ecuador Egypt France Germany Ghana Greece Guatemala India Indonesia Ireland Israel Italy Japan Jordan Kenya Korea Kuwait Latvia Lebanon Malaysia Mexico Netherlands New Zealand Nigeria Norway Oman Pakistan Panama Peru Philippines Poland Portugal Puerto-Rico Qatar Russia Saudi-Arabia Singapore South-Africa Spain Sweden Switzerland Thailand Turkey Ukraine United-Arab-Emirates United-Kingdom United-States Venezuela Vietnam

View our other APIs

Veille: alerte mails

The screenshot shows the Talkwalker Alerts website interface. At the top, there is a navigation bar with the logo 'talkwalkeralerts' and three menu items: 'CREATE', 'MANAGE', and 'IMPORT'. A user email address 'documentaliste.dewind@gmail.com' is visible in the top right corner. The main content area is divided into two sections. On the left, under the heading 'Create Alert', there is a form with the following fields: 'SEARCH QUERY' (empty), 'RESULT TYPE' (set to 'Everything'), 'LANGUAGE' (set to 'All languages'), 'HOW OFTEN' (set to 'Once a day'), 'HOW MANY' (set to 'Only the best results'), and 'YOUR EMAIL' (set to 'documentaliste.dewind@gmail.com'). A red 'CREATE Alert' button is located at the bottom right of the form. On the right, there is a promotional text block titled 'Talkwalker Alerts, the best free and easy alternative to Google Alerts'. Below the text is a cartoon character with a red light on its head. At the bottom of the page, there is a footer with the text 'Talkwalker.com | Search Syntax | Terms of Service'.

Outil d'inscription des sources

Notice des sites

Double usage

- Pour le catalogage
- Pour la collecte

Corpus

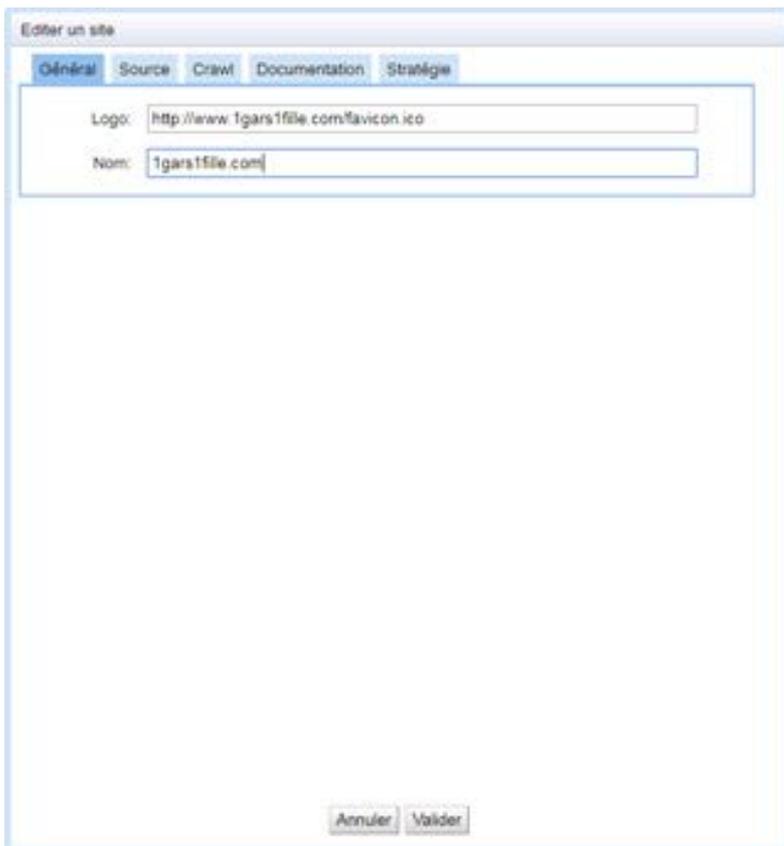
Masquer: Source Crawl Documentation Stratégie

Exporter:

Général		Source		Crawl		Documentation				Stratégie																					
Logo	Nom	Id	Brouillon	Url	Accept	Refuse	Paramètres	Editeur	Secteur d	Fonction	Média	Couverture	Thématique	Personnal	Programme	Public	Service	Statut	Notes	Première	Dernière	c State	Priorité	Fréquence	Fréquence	Robots	Pages	Streams	Feeds	Comment	

Aucune Donnée
0 sites

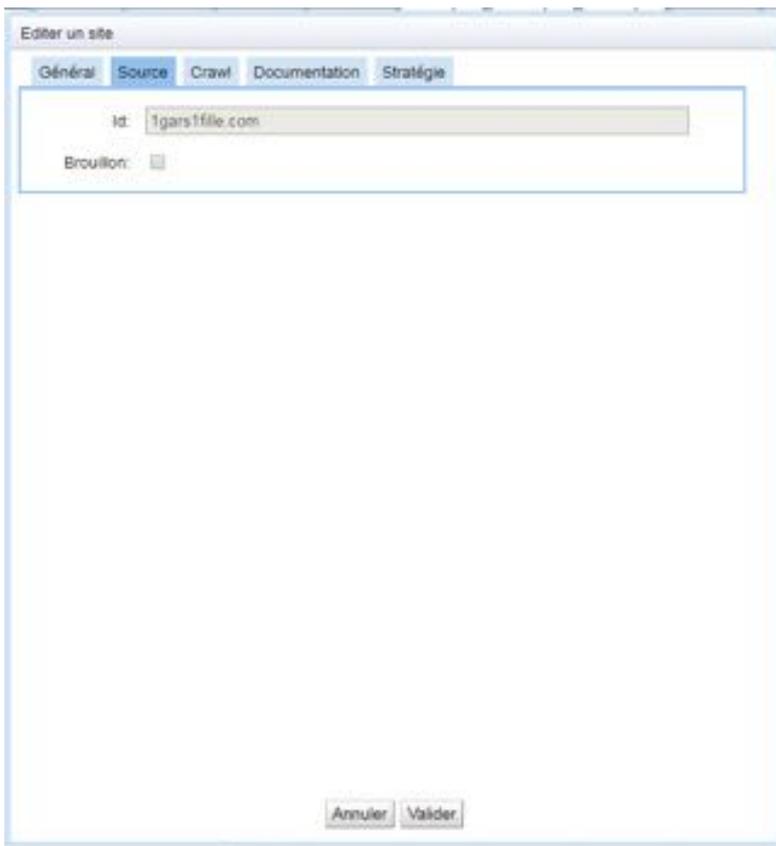
Notice des sites – onglet *général*



The screenshot shows a web application window titled "Editer un site". At the top, there are five tabs: "Général", "Source", "Crawl", "Documentation", and "Stratégie". The "Général" tab is selected. Below the tabs, there are two input fields. The first is labeled "Logo:" and contains the text "http://www.1gars1file.com/favicon.ico". The second is labeled "Nom:" and contains the text "1gars1file.com". At the bottom of the window, there are two buttons: "Annuler" and "Valider".

- Logo du site
- Nom de l'entrée

Notice des sites – onglet *source*



The screenshot shows a dialog box titled "Editer un site" with a tabbed interface. The "Source" tab is selected. It contains a text input field for "Id:" with the value "1gars1file.com" and a "Brouillon:" checkbox which is unchecked. At the bottom, there are "Annuler" and "Valider" buttons.

- Id du site

Notice des sites – onglet *crawl*

The screenshot shows a web browser window titled 'Editer un site' with a tabbed interface. The 'Crawl' tab is active. It contains four text input fields: 'Url' with two lines of text, 'Accept' with 'hosts[lgarsifille.com/]', 'Refuse' which is empty, and 'Paramètres' with '--- {}'. At the bottom are 'Annuler' and 'Valider' buttons.

- URL(s) du site à archiver
- Hosts des pages du site à archiver
- Hosts des pages du sites à ne pas archiver (permet d'externaliser la partie d'un site pour la considérer comme un site à part entière)

Notice des sites – onglet *documentation*

Editer un site

Général Source Crawl **Documentation** Stratégie

Editeur:

Secteur d'activité: commentaire

Fonction: agrégateur
annuaire
banque de contenus
blog
forum
guide TV
plateforme de blog
portail
service
site de chaîne radio
site de chaîne TV
site de programme
site institutionnel
site marchand
site participatif

Média: TV
radio
web

Couverture: locale
nationale
internationale

Thématique: Actualités des médias

Annuler Valider

- Nom de l'éditeur
- Secteur d'activité
- Fonction du site
- Type de média représenté
- Couverture du média

Notice des sites – onglet *documentation* (1/2) ina

Editer un site

Personnalité:

- autre
- acteur
- chroniqueur
- compositeur
- doubleur/voix off
- journaliste
- monteur
- présentateur/animateur
- producteur
- réalisateur
- scénariste

Programme:

- autre
- animation/dessin animé
- best of/bibliothèque
- campagne d'information
- captation
- chronique
- comédie de situation
- court métrage
- moyen métrage
- long métrage
- fiction
- documentaire
- interview entretien
- jeu
- journal télévisé

Public:

- tous publics
- jeunesse
- professionnel

Annuler Valider

- Personnalité
- Type de programmes
- Public cible du site

Notice des sites – onglet *documentation* (2/2) ina

The screenshot shows a web form titled "Modifier un site". At the top, there is a dropdown menu with "web fiction vidéo" selected. Below this are several fields:

- Public:** A dropdown menu with "tous publics" selected. Other options include "jeunesse", "professionnel", "scolaire/universitaire", "senior", "adulte", and "communauté".
- Service:** A dropdown menu with "gratuit" selected. Other options include "payant", "audio disponible", "vidéo disponible", and "inscription obligatoire".
- Statut:** A dropdown menu with "mort" selected.
- Notes:** A text area containing the text: "- 03/10/11 : site officiel de la série 'Un gars, une fille' diffusée sur France 2 entre 1999 et 2003. Produite par Isabelle Camus et Hélène Jacques, la série met en scène les deux acteurs Jean Dujardin et Alexandra Lamy."
- Première capture:** A text field containing "2001-04-02".
- Dernière capture:** A text field containing "2012-09-03".

At the bottom of the form, there are two buttons: "Annuler" and "Valider".

- Nature du service
- Statut du site
- Notes : champs libre
- Date de première et de dernière capture

Notice des sites – onglet *stratégie*

Editer un site

Général Source Crawl Documentation **Stratégie**

State: disabled

Priorité: 0.5

Fréquence Auto: yearly

Fréquence Manuelle: unknown

Pages:

Robots: phagosite
crocket
live_archiving

Streams:

Feeds:

Annuler Valider

- unknown
- unknown**
- hourly
- daily
- weekly
- monthly
- yearly

Exemple de notice en consultation

Logo	Nom	Editeur	Secteur d'activité	Fonction	Média	Couverture	Thématique	Personnalité	Programme	Public	Service	Statut	Notes	Première capture	Dernière capture
	RTS														
	RTS	RTS	site institutionnel	site institutionnel	TV		éducation et enseign			professionnel	gratuit	actif		2008-03-18	
	RTS	RTS	groupe média	site institutionnel	TV							actif		10/10/11 12:00:00	2011-04-13
	RTS	RTS	diffuseur	site de programme	web						actif, vidéo disponible	actif		2008-12-18 09:00:00	2012-04-11
	RTS	RTS	diffuseur	site de programme	web						actif, vidéo disponible	actif		2008-12-18 09:00:00	2012-01-09
	RTS	RTS	diffuseur	site de chaîne TV	TV						actif, vidéo disponible	actif		12/01/07 12:00:00	12-12-12

CONSULTEZ UN SITE

Général

Logo:

Nom: <http://www.rts.ch/fr/actualites/index.html#RESULT>

Site: RTS

Documentation

Editeur: RTS

Secteur d'activité: diffuseur

Fonction: site de chaîne TV

Média: TV

Couverture: nationale

Thématique:

Personnalité:

Programme:

Public: tous publics

Service: gratuit vidéo disponible

Statut: actif

Notes: - 12/01/07
Site de la chaîne de télévision RTS.
Créé le 01/01/07 puis rebaptisé le 15/04/08, elle est maintenant disponible en multilingue sur le RTS, le site, par le biais de l'interface, sur le site de la RTS, sur le site de la RTS, sur le site de la RTS.
- 20/04/11
Création du site internet en décembre 1995 avec une nouvelle version en mai 1999.
En juin 2000 création de RTS, plateforme de vidéo-online. En novembre 2000 création du portail internet RTS.ch. En 2013, RTS.ch propose la fonction Connect.
RTS.ch est un portail d'accès aux programmes de la chaîne direct, replis ou connect, à des sites thématiques, vidéo à la demande, jeux, service de relations presse, etc.
- 30/04/12
Relance du site le 30/04/12 qui double son offre en regroupant les contenus issus de ses 4 chaînes (RTS, RSI, RSI, RSI), à donner toujours accès aux contenus parents avec un lien direct à l'information avec un lien direct (le site est intégré à l'écran sur lequel il est consulté, il propose un "menu pop-up" qui permet à l'utilisateur de donner une vidéo tout en poursuivant sa navigation sur le site, il donne accès au direct des émissions, au replis des émissions classées par thème et par genre. Les films sont également. En plus des contenus TV, annonce d'une plateforme RTS.ch qui rassemblera des programmes diffusés uniquement sur internet. Les sites de site sont successivement RTS, puis RTS.ch et RTS.
Les sites de groupe sont hébergés sur la plateforme RTS.ch et sont accessibles pour la consultation, les par leurs liens uniquement.

Première capture: 1996-12-13
Dernière capture:

PARTIE 7

Mise en consultation de l'archive du web

Profils des utilisateurs



- Chercheurs (majoritairement en Sciences Humaines)
- Étudiants

Accès

→ Uniquement dans les instances Ina via les postes de consultations multimédia (PCM)

→ 43 sites



en rouge médiathèques et bibliothèques municipales équipées
en jaune cinémathèques équipées
en vert les projets d'implantation
en bleu les délégations régionales de l'Ina

Type de consultations

Consultation experte

Centres de consultation Ina
La consultation dans les centres Ina permet de profiter de l'accompagnement des équipes, d'outils d'aide à l'analyse et d'un environnement de travail numérique personnel.



7 sites

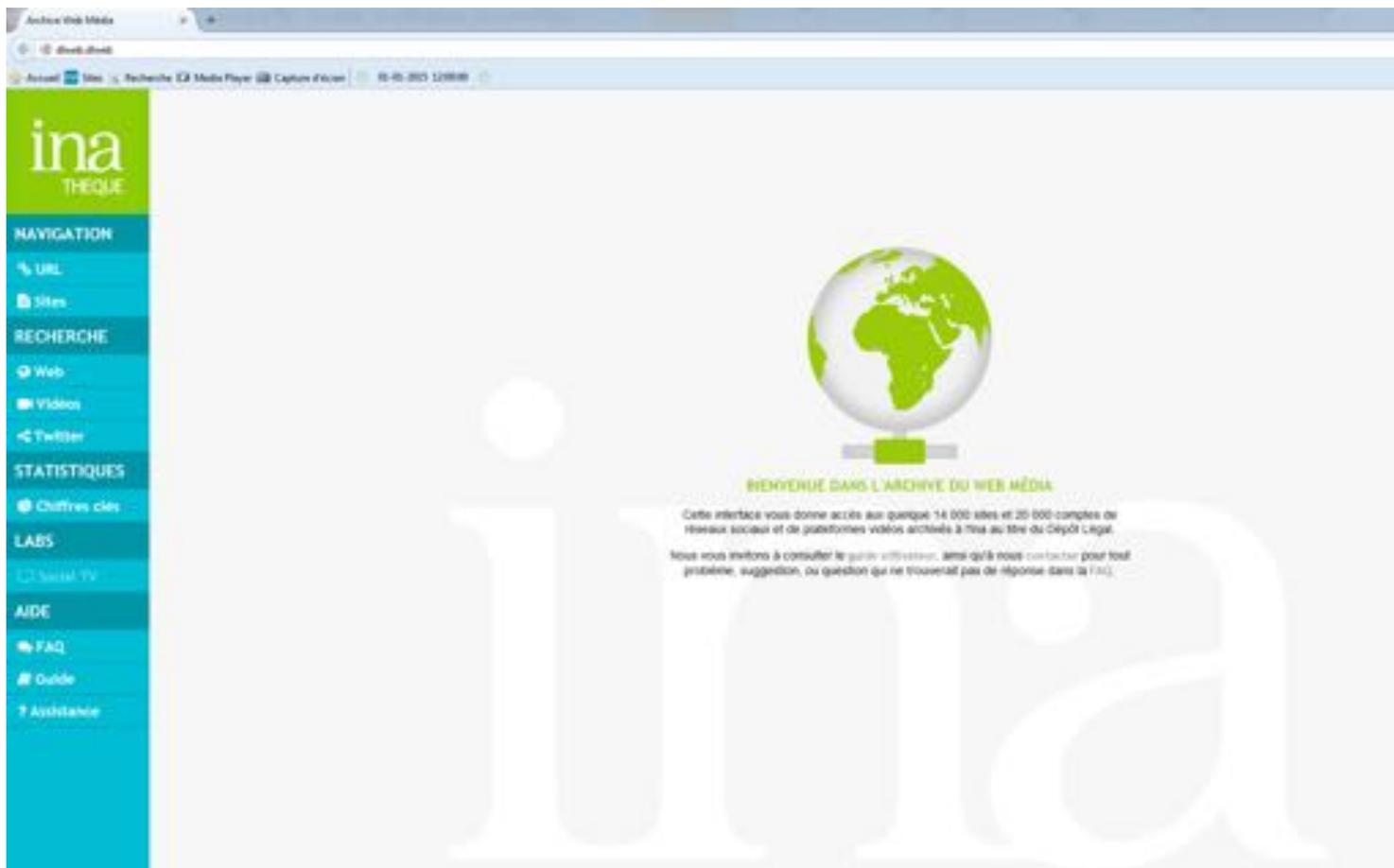
Consultation autonome

Consultation hors sites Ina
Des postes de consultation autonome sont installés dans des bibliothèques et médiathèques municipales à vocation régionale. Ils permettent de consulter, près de chez vous, l'intégralité de nos fonds audiovisuels.

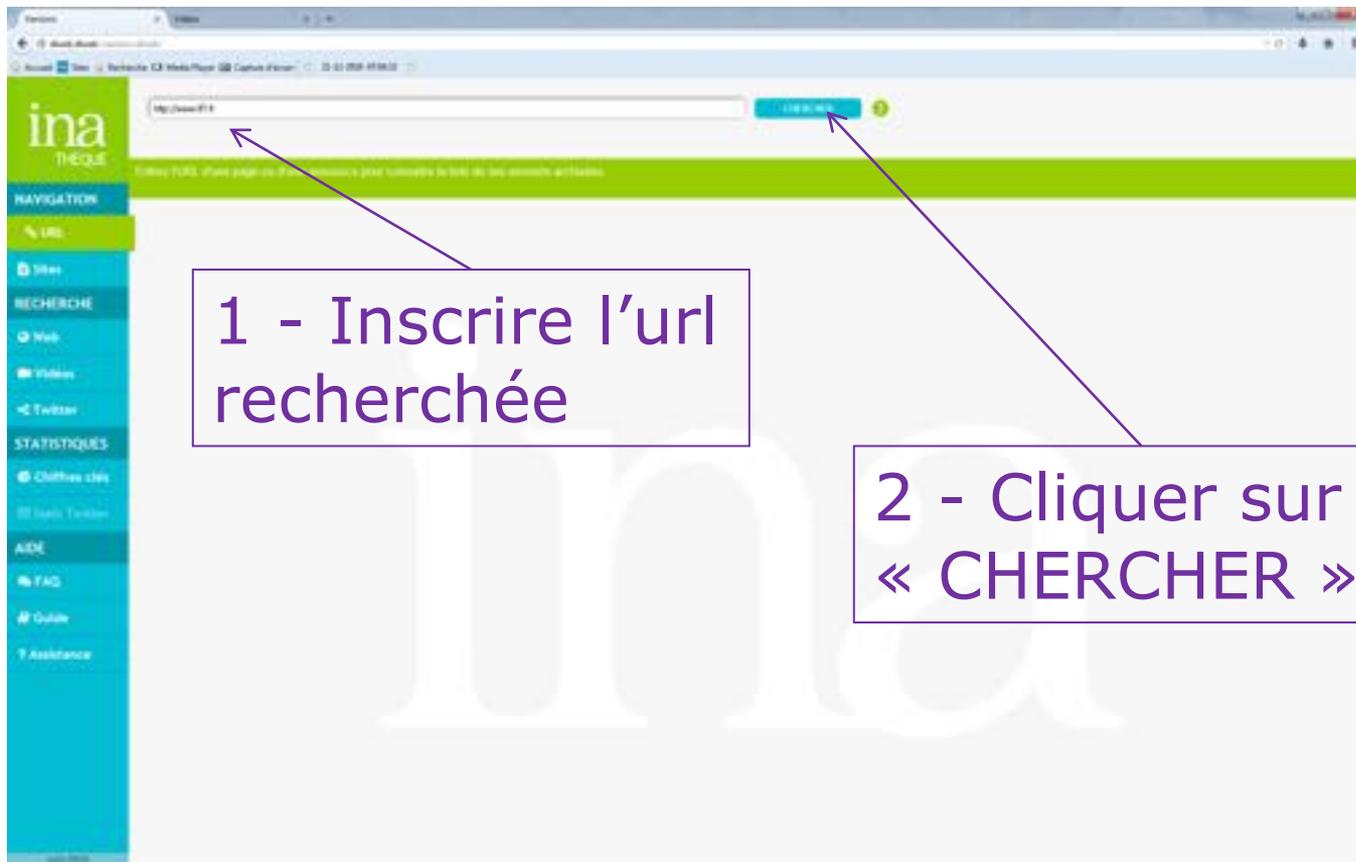


36 sites

Interface de consultation



Recherche par URL (1/4)



1 - Inscrire l'url recherchée

2 - Cliquer sur « CHERCHER »

Recherche par URL (2/4)



Répartitions des versions par année

Nombre de versions disponibles

Recherche par URL (3/4)

ina THEQUE

2016

2017

2018

NAVIGATION

URL

Site

RECHERCHE

Web

Vidéo

Twitter

STATISTIQUES

Cliffhies clés

Insights Twitter

AIDE

FAQ

Guide

Assistance

04:55 07:38 07:45
05:16 07:04 07:25
05:14 07:20 07:35

Choix d'une version en cliquant sur l'heure correspondante

Recherche par URL (4/4)



- 👉 A partir de là, la navigation se fait intégralement dans l'archive.
- 👉 A chaque clic, l'outil de navigation recherche la version de l'url en date la plus proche de celle demandée

Recherche Catalogue

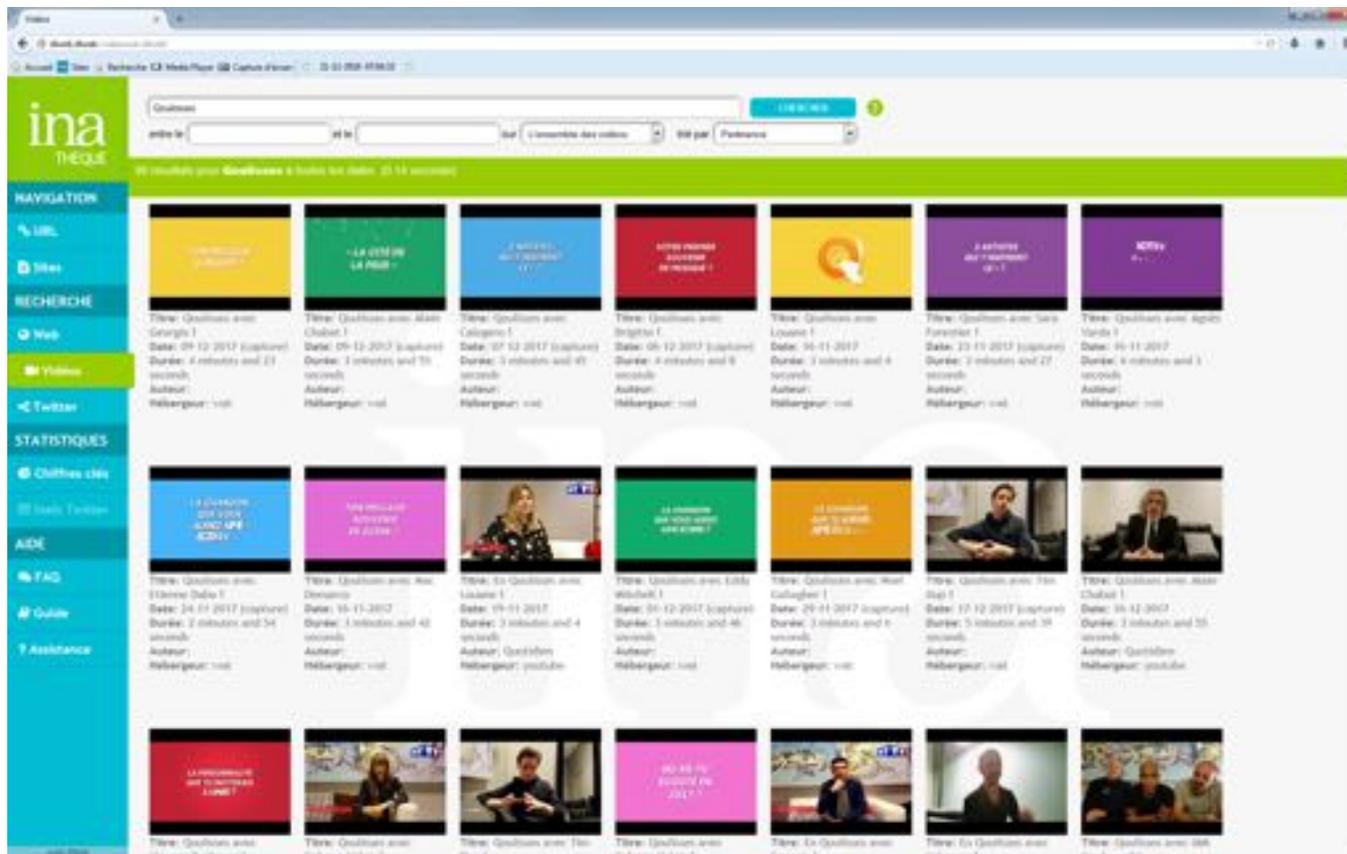
Recherche directement dans la liste des sites du corpus:

- À partir de l'url de la racine d'un site
- A partir d'un mot contenu dans l'url
- A partir d'un ou plusieurs critères (fonctions, thématique, etc)

The screenshot shows the 'ina TheCat' search interface. On the left, there is a navigation sidebar with options like 'URL', 'Site', 'Web', 'Vidéo', 'Twitter', 'Statistiques', 'Classements', 'Aide', 'FAQ', and 'Guide'. The main area displays a table of search results with columns for 'Site', 'Fonction', 'Thématique', 'Périodicité', 'Langue', 'Statut', 'Date de mise à jour', and 'Dernière capture'. The table lists various media sites such as 'lepoint.com', 'lefigaro.fr', 'leparisien.fr', 'lemonde.fr', etc., with their respective categories and update dates.

Outil de consultation archive vidéos (1/2)

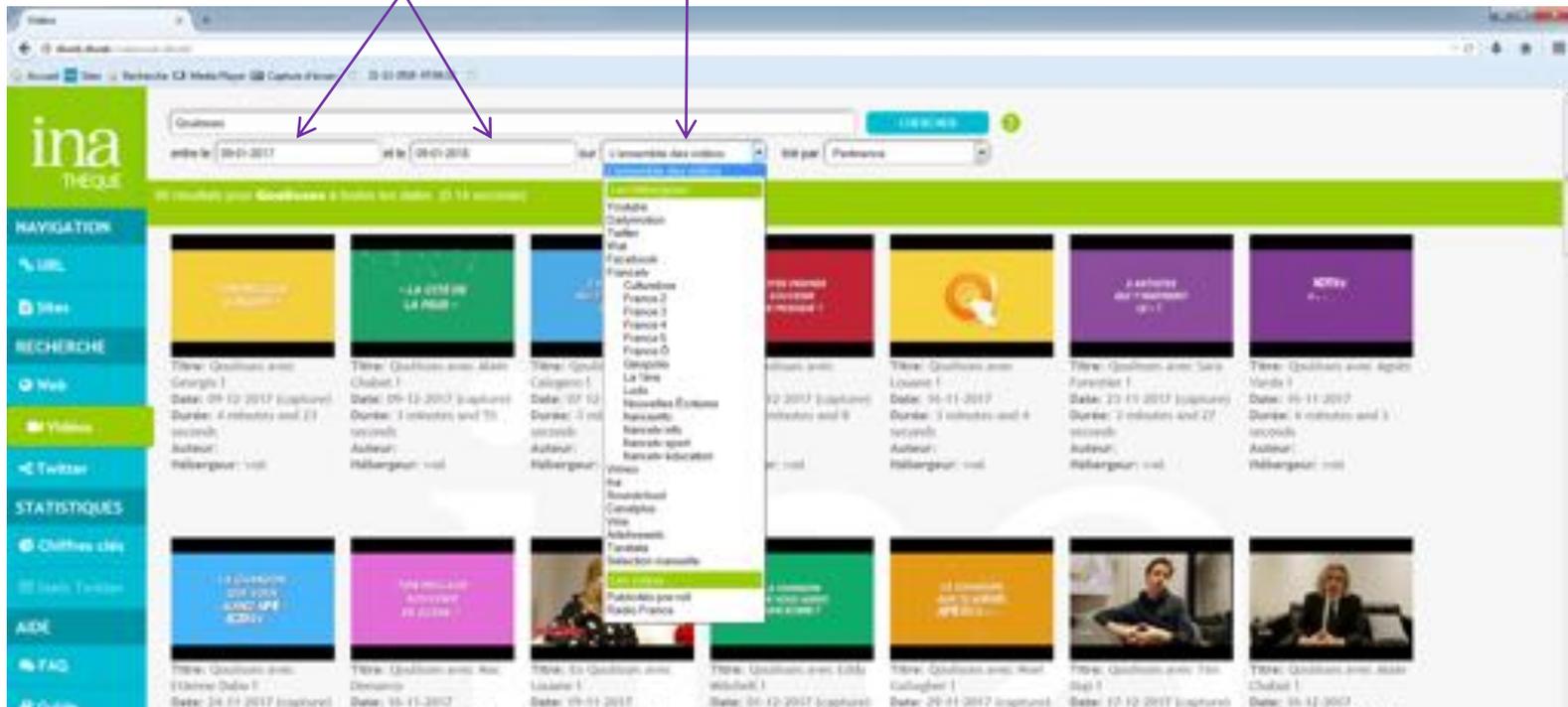
Permet de consulter les vidéos issues du web et archivées:



Outil de consultation archive vidéos (2/2)

Possibilité de restreindre la recherche par :

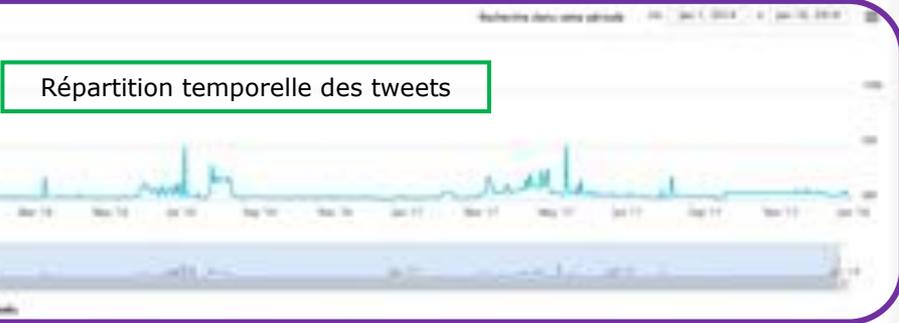
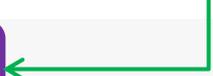
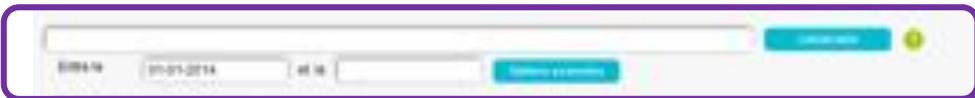
Date Domaine



Outil de consultation Twitter (1/2)

Présentation de l'interface:

Zone de recherche



Répartition temporelle des tweets

Nombre de résultats

outils d'analyse

Options d'affichage



Affichage des résultats et outil d'analyse

Outil de consultation Twitter (2/2)

→ Recherche simple : recherche d'un mot dans tous les tweets de l'archive



→ Recherche avancée: accessible en cliquant sur Options avancées



Recherche dans un des sous-corpus

Recherche par hashtag

Recherche par mention

Recherche par nom d'utilisateur

Recherche par langue

→ 4 collections:

- Médiasphère : les tweets liés à l'audiovisuel français
- Attentats : les tweets liés aux attentats (janvier 2015, novembre)
- Elections 2017 : les tweets liés aux elections 2017
- Trends : les tweets tendance

Outils d'analyse Twitter (1/4)

Liste des résultats:

- Présente de façon brute les résultats
- Possibilité d'afficher/masquer les colonnes correspondant au méta-données des tweets (date, nombre de retweet, langue, localisation) en cliquant sur

Exemple:

Ici on affiche les métadonnées *Tweet*, *Texte brut du tweet* et *location*



Outils d'analyse Twitter (2/4)

Dashboard:

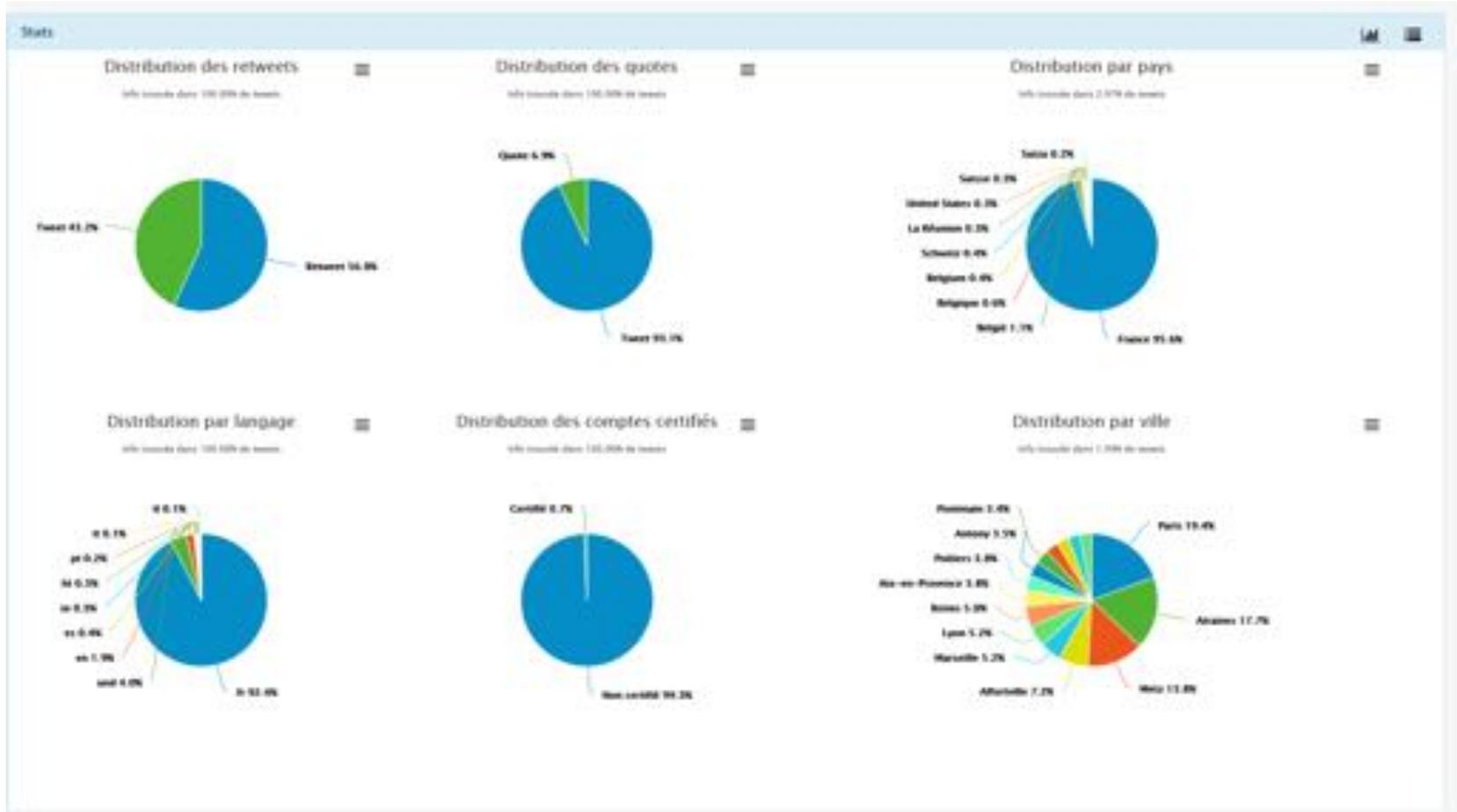
→ Outil mettant à disposition des outils complémentaires pour l'analyse des résultats

- **Stats** : statistique de distribution par pays, ville, langue dans les tweets archivés et contenant le terme recherché
- **Top Hashtags**: hashtags les plus présents dans les tweets archivés et contenant le(s) terme(s) recherché(s)
- **Top Mentions**: mentions les plus présentes dans les tweets archivés et contenant le(s) terme(s) recherché(s)
- **Top liens**: liens les plus partagés dans les tweets archivés et contenant le(s) terme(s) recherché(s)
- **Top Utilisateurs** : utilisateurs les plus représentés dans les tweets archivés et contenant le(s) terme(s) recherché(s)
- **Top Media**: liens les plus partagés dans les tweets archivés et contenant le(s) terme(s) recherché(s)
- **Top Emoji**: émoji les plus utilisés dans les tweets archivés et contenant le(s) terme(s) recherché(s)

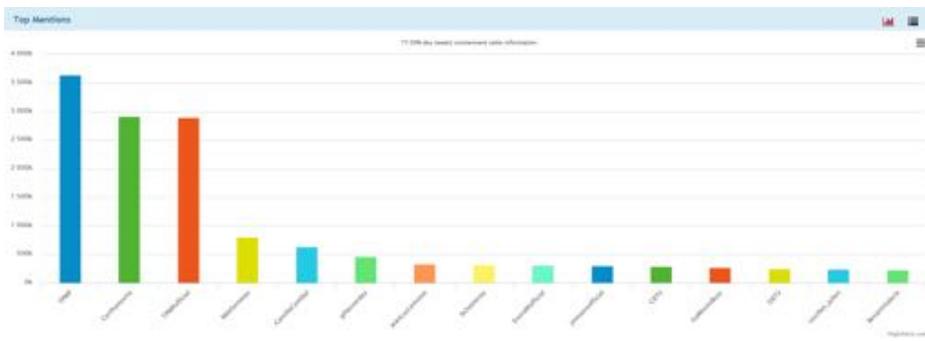
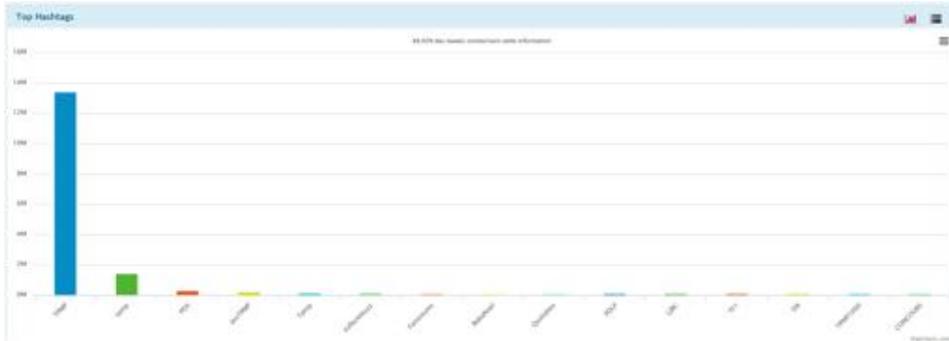
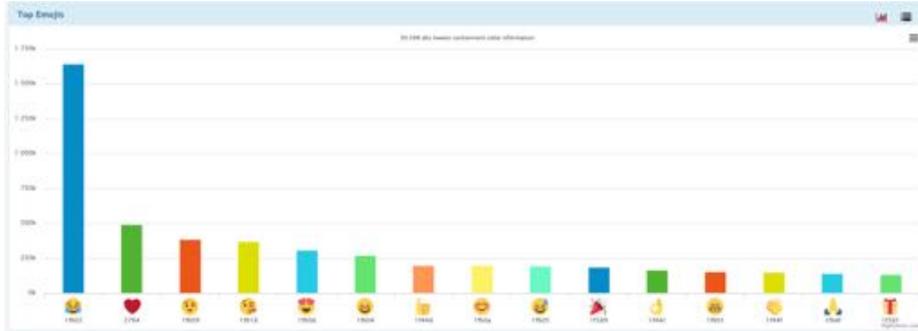
→ Utilisations :

- Analyse géographique, comparative, de sentiments, sérendipité....

Outils d'analyse Twitter (3/4)



Outils d'analyse Twitter (4/4)



Merci pour votre attention

Questions: bblanckemane@ina.fr
dlweb@ina.fr