Archivage du web à l’Ina

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

Historique 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.

L’Ina: institut et entreprise

- → 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

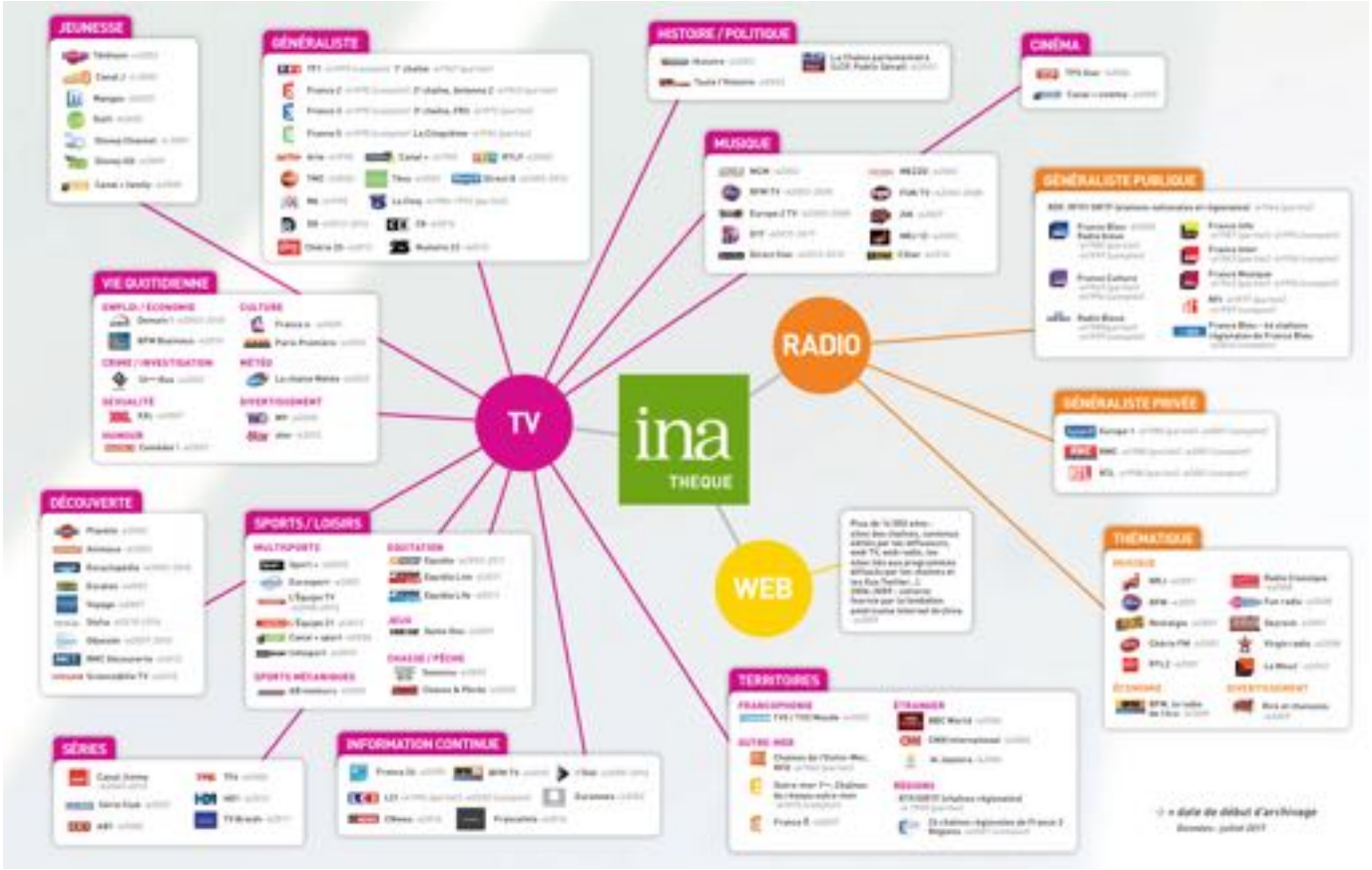
Objet :
Constituer une collection de référence pour la mémoire collective du pays.

Historique :
28 décembre 1537: première ordonnance royale de François 1er
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



Dépôt légal du Web: contexte 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 »

ina

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

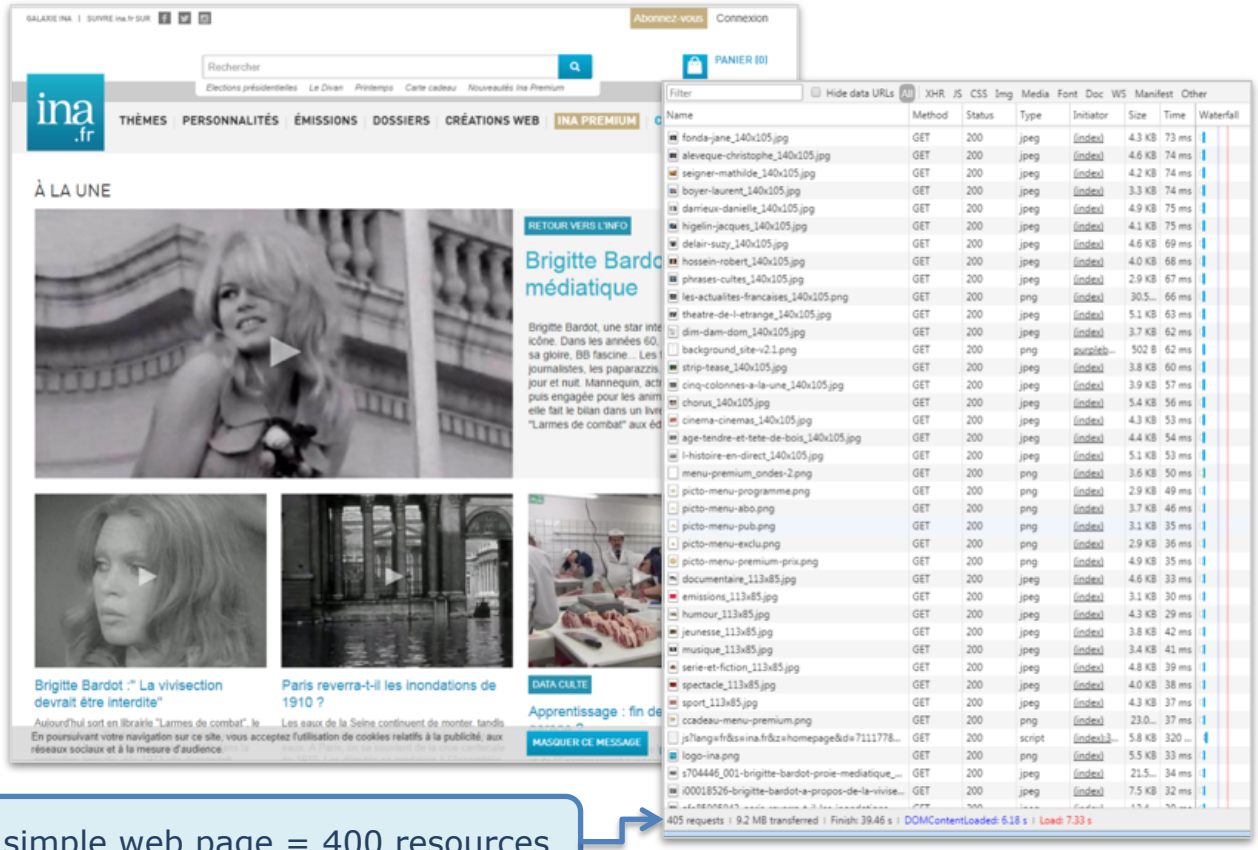
Premier serveur Web

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



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)

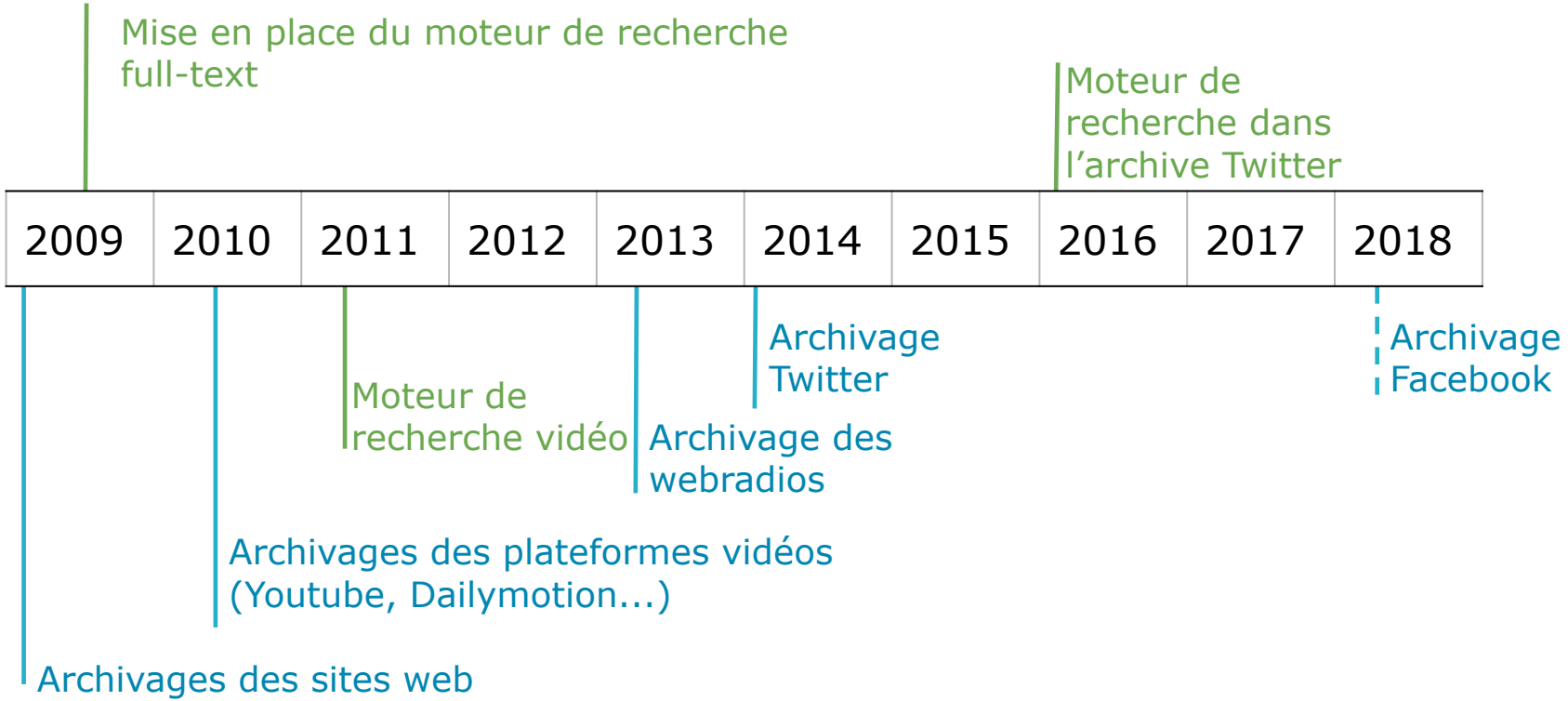
…et dans le temps

Source: visualcapitalist.com

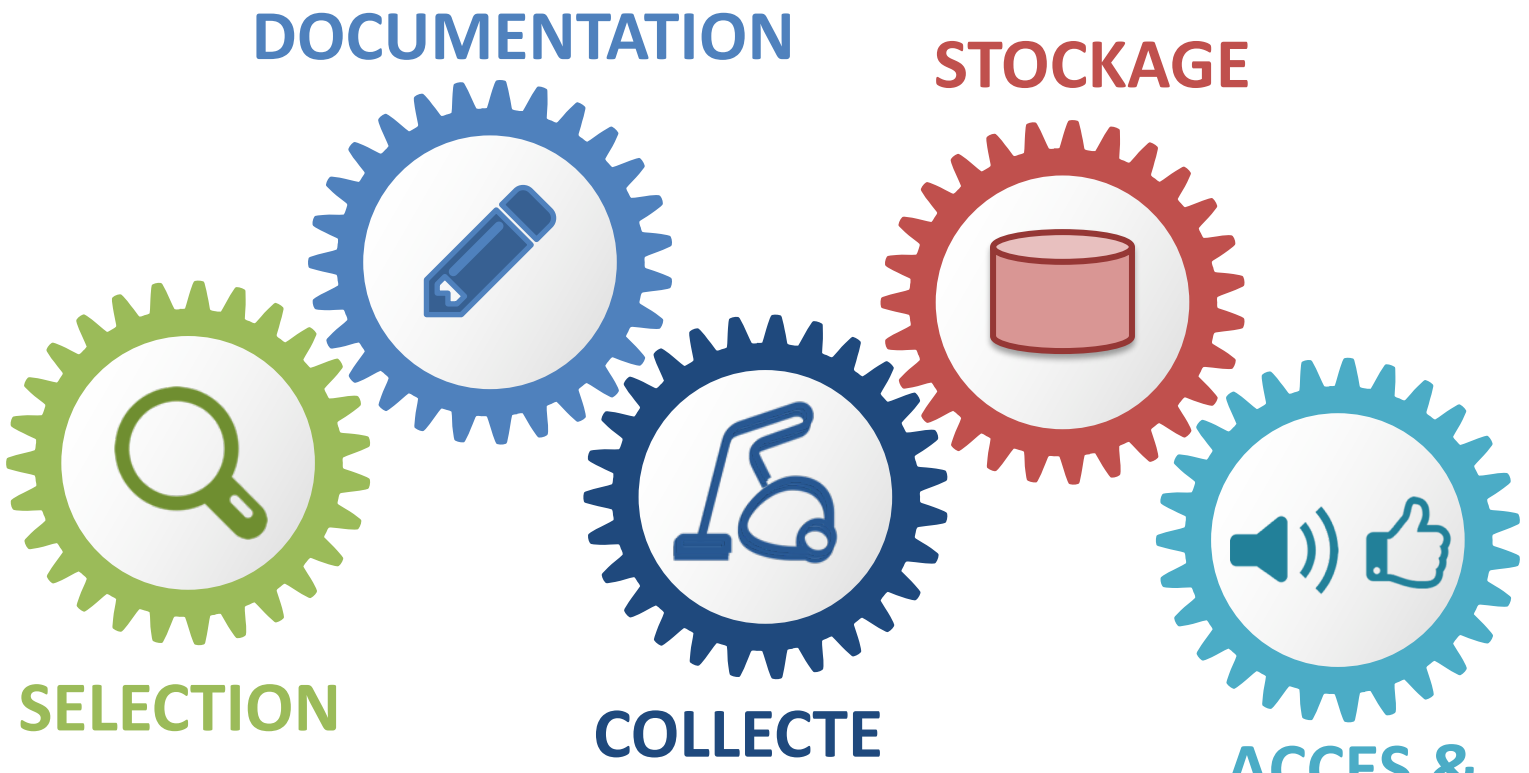
PARTIE 4

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 média (www.plusbellelavie.org)
- Comptes de réseaux sociaux liés aux médias (@CashInvestigations, @TPMP)
- UGC: Chaînes Youtube, Dailymotion, Viméo 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, Viméo
- 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), LinkedIn, GitHub…

Philosophie de la chaîne d'archivage

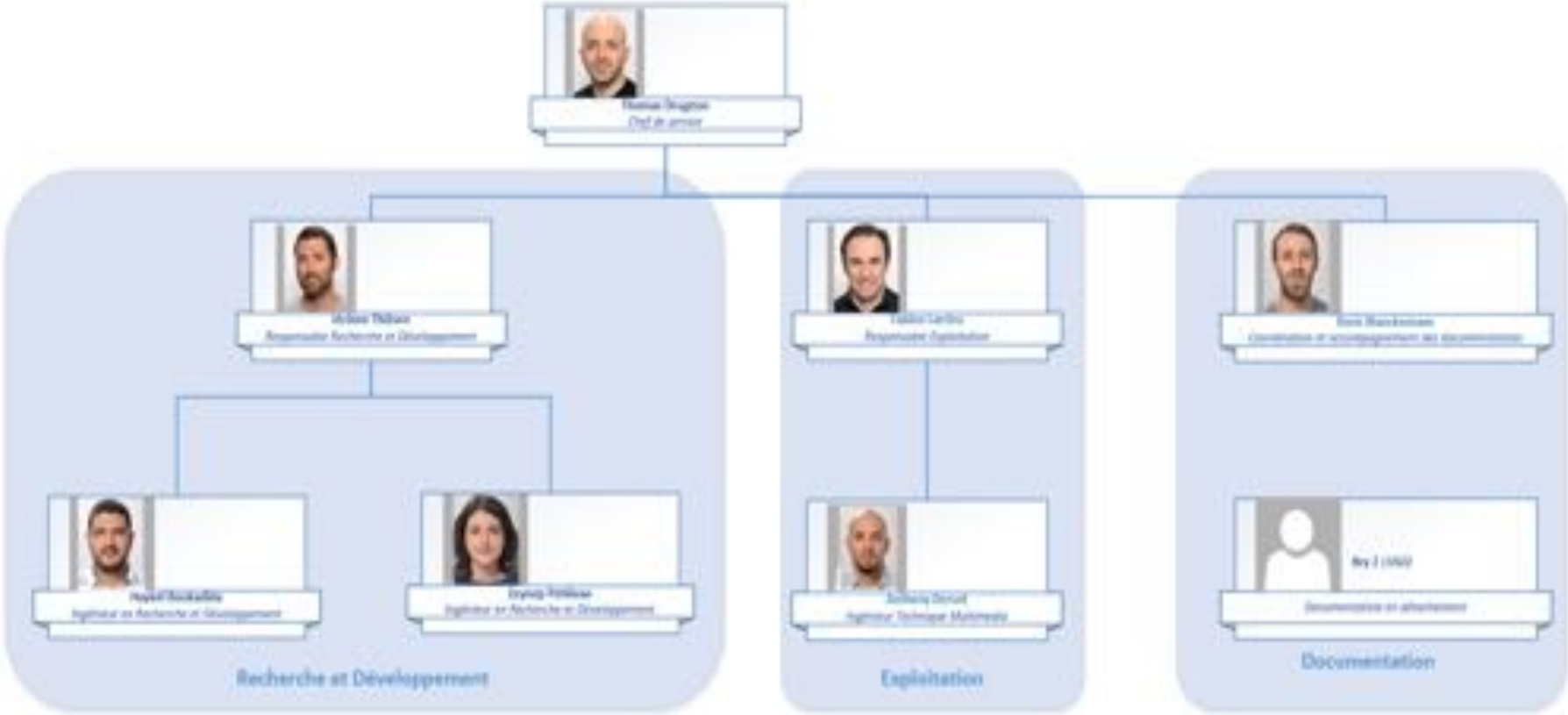
→ Prise en charge complète et autonome d'un maximum des aspects de l'archivage du web: stockage, collecte, infrastructure…

→ 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

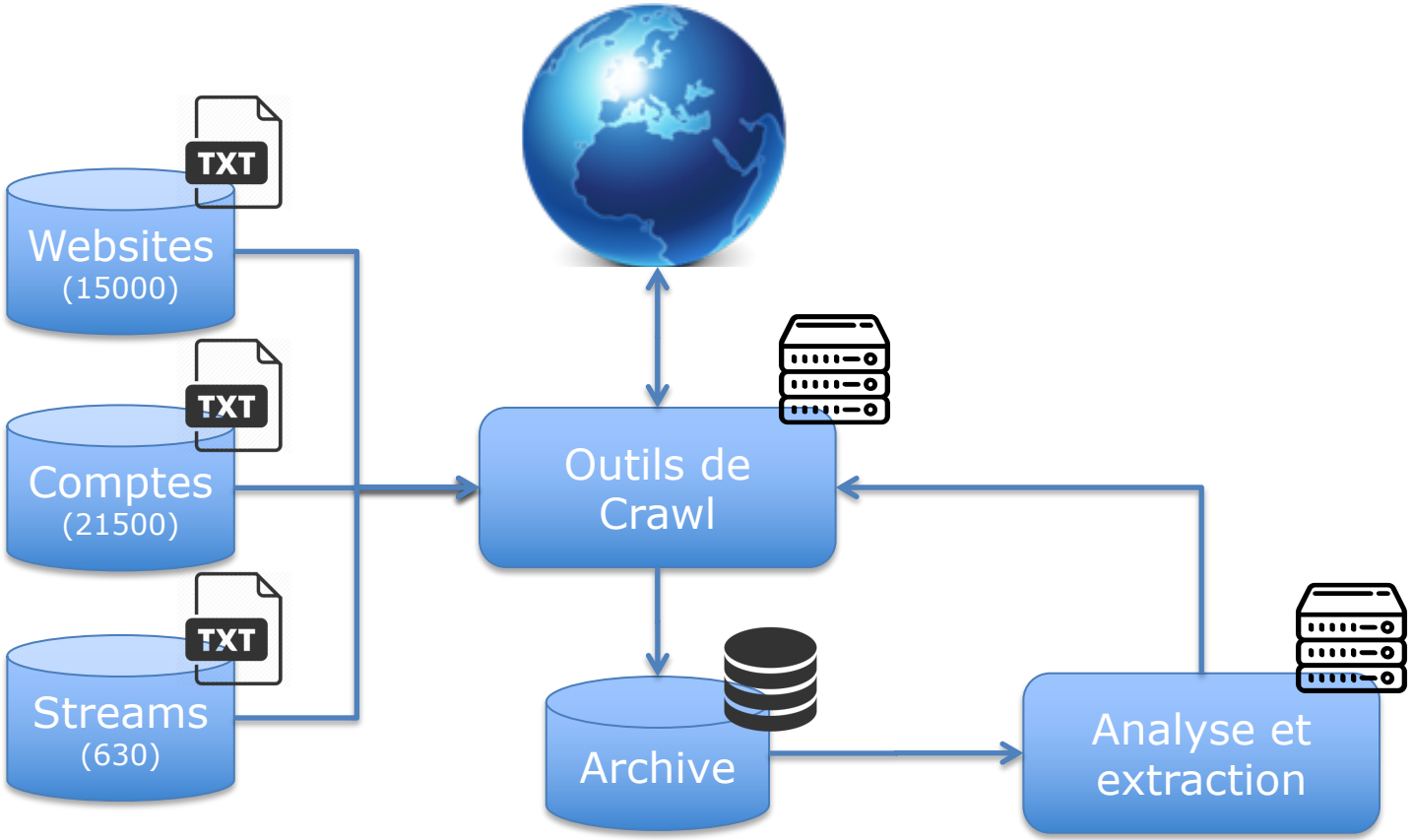
Équipe en charge de la collecte



PARTIE 5

Techniques de collecte du web

Principe général de la collecte



Robot de collecte - définition

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 :

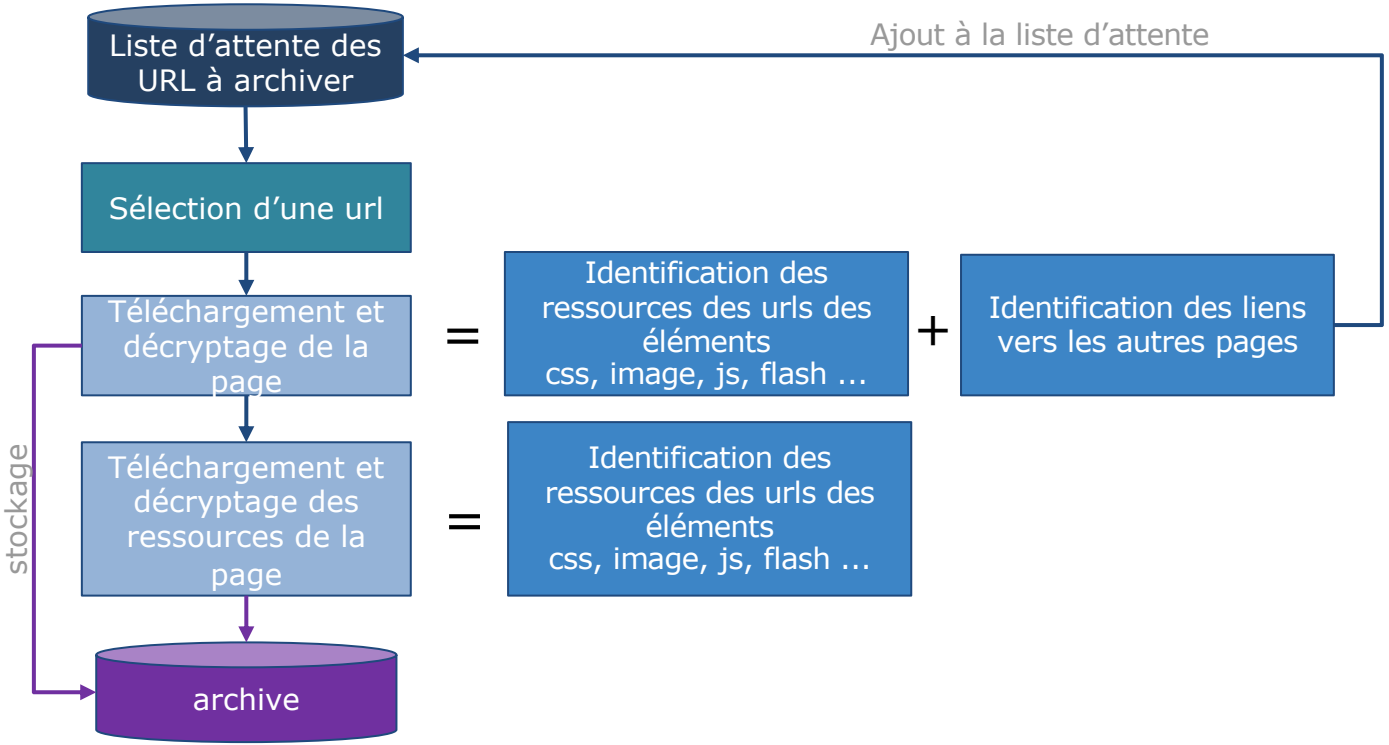
→ Contenu manquant :

  - 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

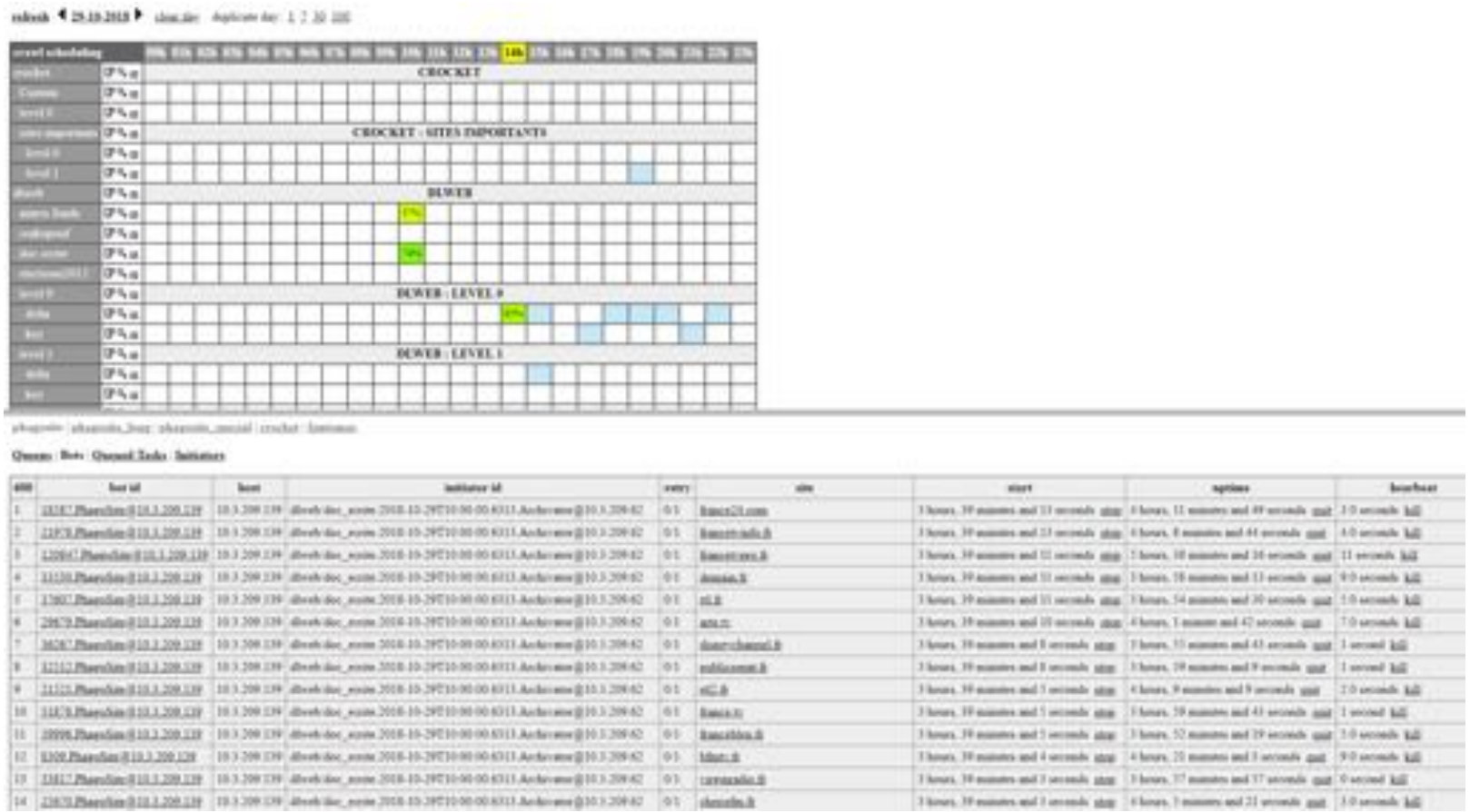
→ Crocket: robot plus spécialisé

- 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…)

Ordonnanceur des robots de collecte



Format d’archive de la collecte

→ **DAFF** : Digital Archiving File Format

→ Format containeur

→ 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
12345678901234567890 1234567890	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*)

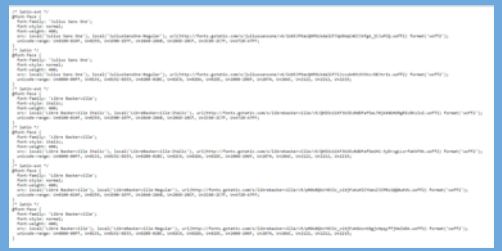
→ DAFF: sha256

Le format DAFF en pratique

daff data records

sha_256: e4ba78b2c0034f

sha_256: babc4e3130004def6



daff metadata records

```
url:      http://prog-tv.fr/images/mazinger.jpg
date:     2018-01-02 08:51:00Z
sha_256:  e4ba78b2c0034f
```

```
url:      http://prog-tv.fr/images/mazinger.jpg
date:     2018-01-13 10:02:00Z
sha_256:  e4ba78b2c0034f
```

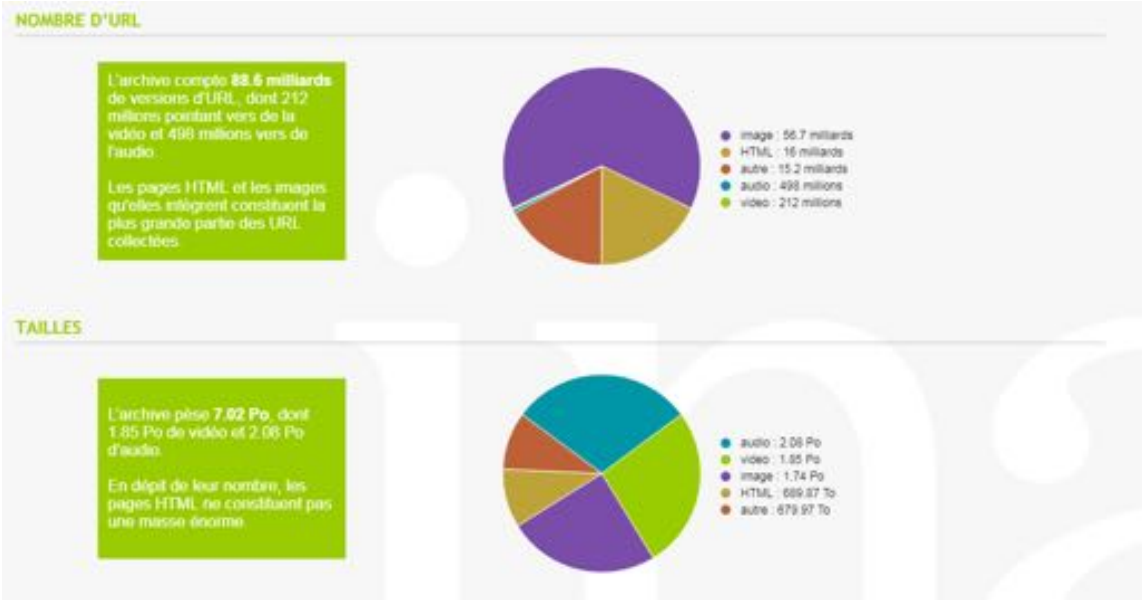
```
url:      http://anime-fan.fr/img/mazingerZ.jpg
date:     2018-02-02 18:33:00Z
sha_256:  e4ba78b2c0034f
```

```
url:      http://prog-tv.fr/style/main.css
date:     2018-01-02 08:50:00Z
sha_256:  babc4e3130004def6
```

```
url:      http://prog-tv.fr/style/main.css
date:     2018-01-13 10:01:00Z
sha_256:  babc4e3130004def6
```

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 archivées

Collecte des UGC: contexte et sources

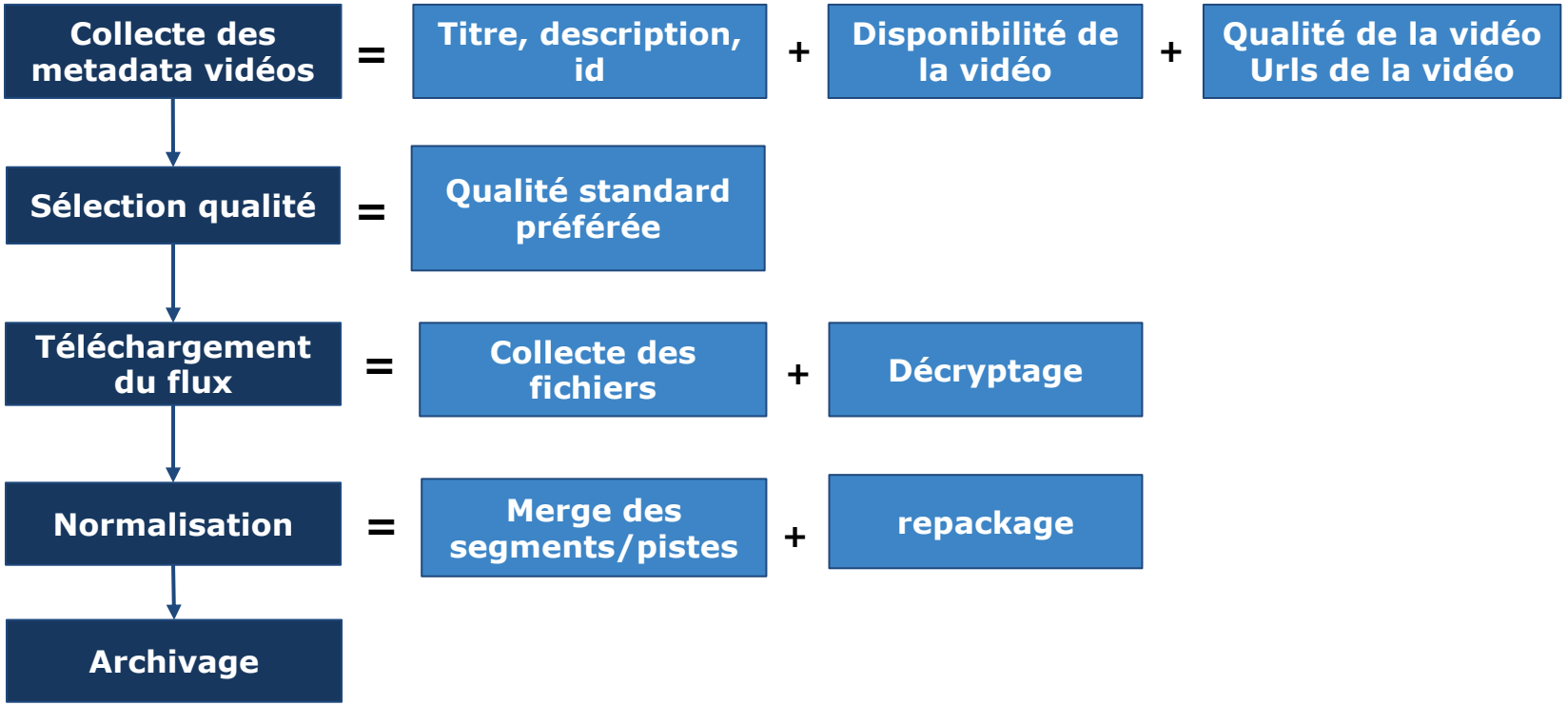
Spécificité des vidéos sur le web :

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

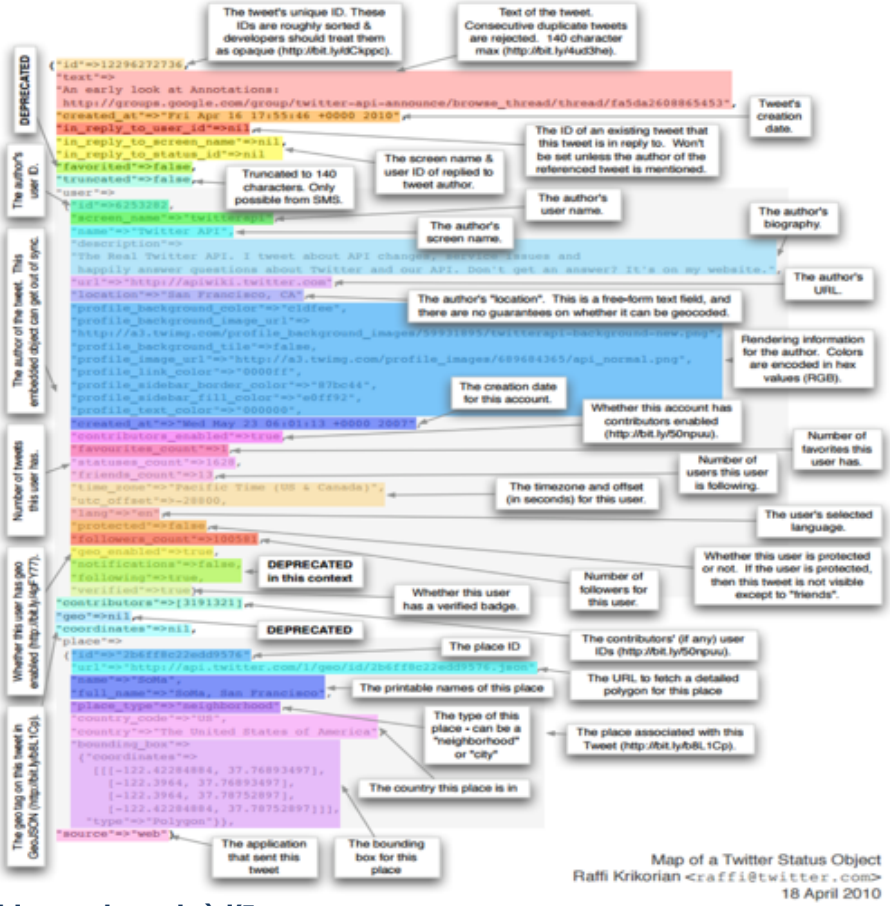
Sources de collecte :

- 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

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
- Evénements nationaux importants (élections présidentielles, jeux olympiques…)

→ ≈ 270 000 tweets archivés par jour (hashtags)

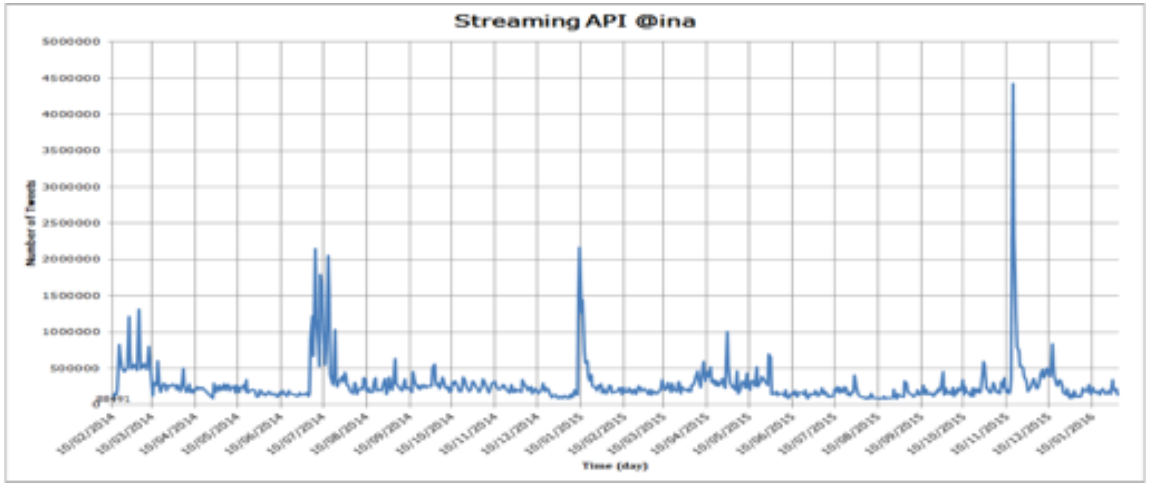
→ ≈ 48 000 tweets archivés par jour (timeline)

→ 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

Critère et méthode de sélection des sources

→ Quoi ?

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

→ Comment ?

- Veille en mode push
- Veille en mode pull

→ Fréquence ?

Quotidienne

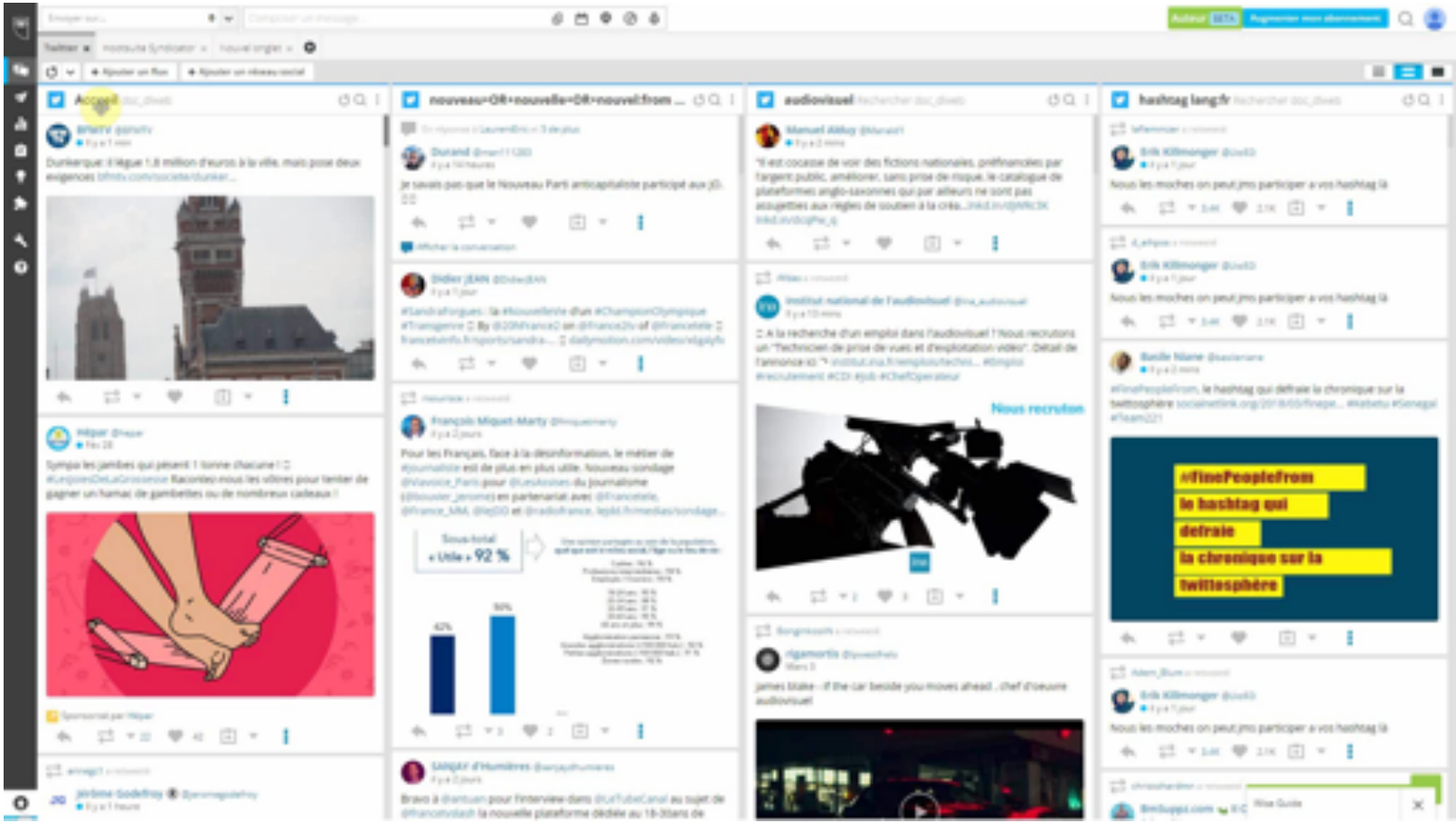
→ Outils ?

Outils maison: tableur excel + macro

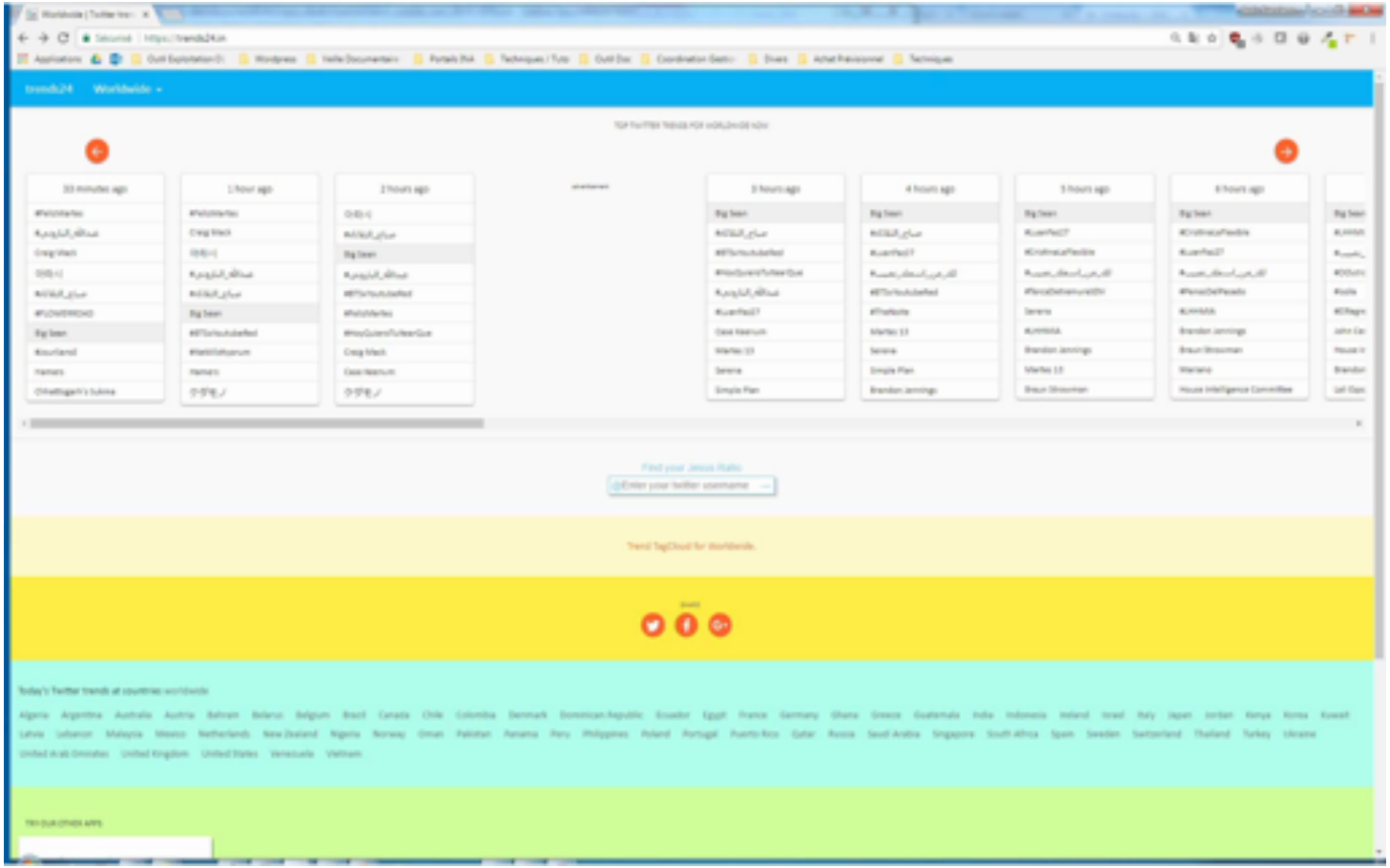
Veille: push vs. pull

Méthodes	Principes et outils	Avantages	Inconvénients
Push	▪ Sur les réseaux sociaux, abonnement aux comptes "clés" : • comptes de chaines 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)	• Automatisation forte (gain de temps) • Outils existants (gratuits / freemium): Hootsuite, Talkwalker Alert • Repose peu sur l'humain	• Bruit / résultats non pertinents • Paramétrage des outils • Sérendipité faible • Faible maîtrise du processus de recherche
Pull	Recherche dans un moteur	• Sérendipité forte • Maîtrise du processus de recherche	• Chronophage • Repose fortement sur l'humain

Veille : hootsuite



Veille : tendances



Veille: alerte mails

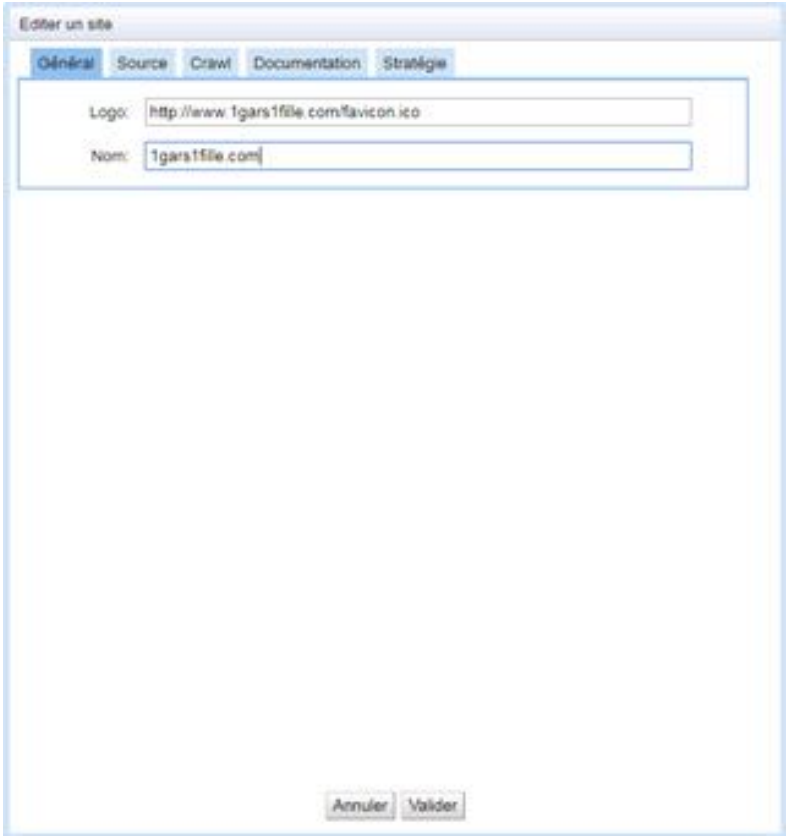
Outil d’inscription des sources

Notice des sites

Double usage

- Pour le catalogage
- Pour la collecte

Notice des sites – onglet *général*



- Logo du site
- Nom de l’entrée

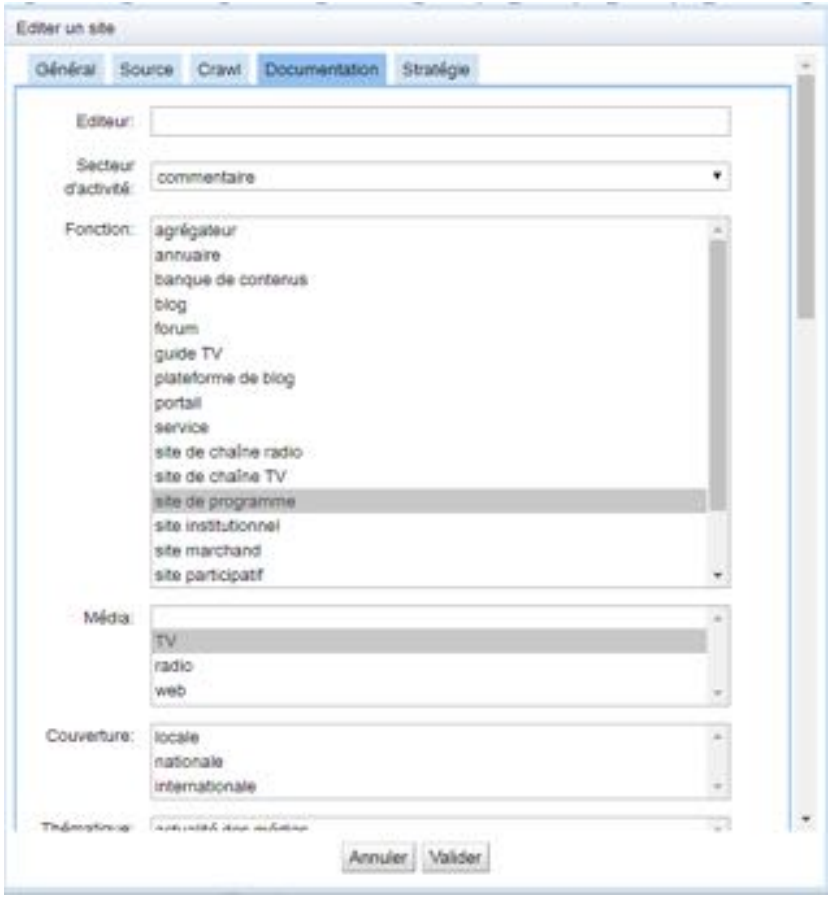
Notice des sites – onglet *source*

- Id du site

Notice des sites – onglet *crawl*

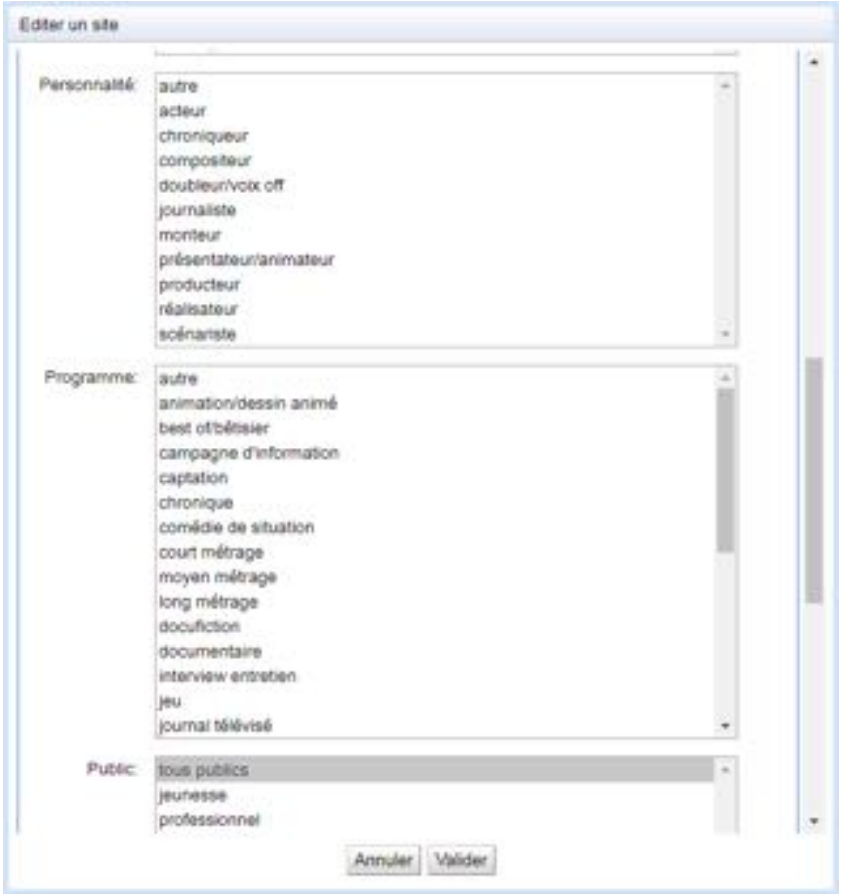
- URL(s) du site à archiver
- Hosts des pages du site à archiver
- Hosts des pages du sites à ne pas archiver (permet d'externaliser la partie d'un site pour la considérer comme un site à part entière)

Notice des sites – onglet *documentation*



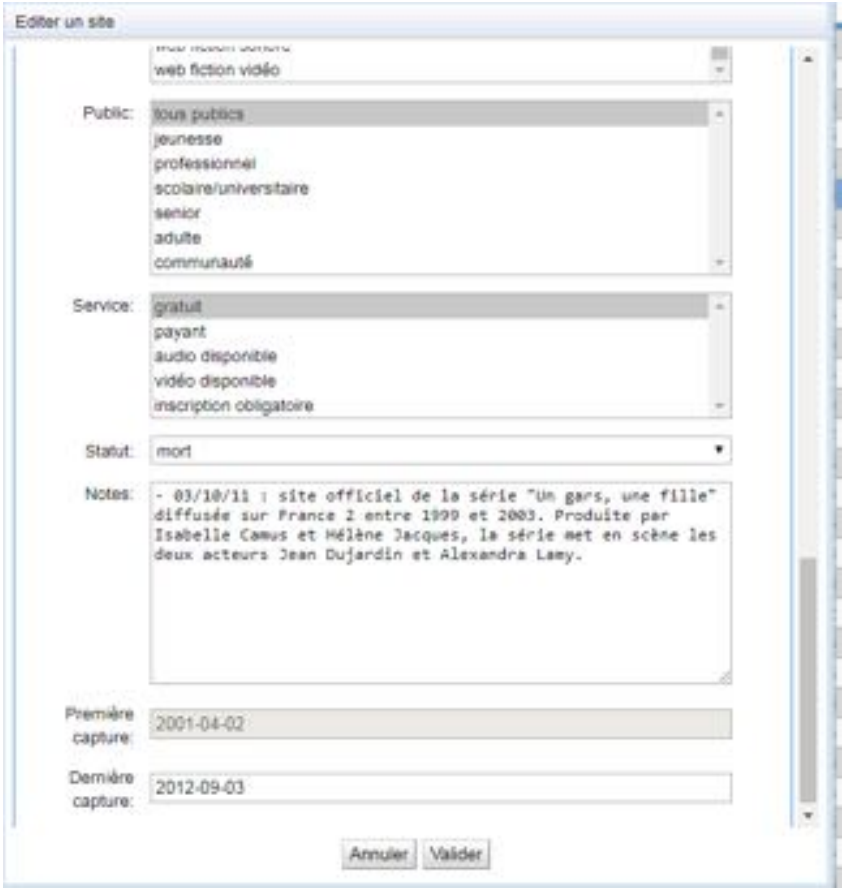
- Nom de l'éditeur
- Secteur d'activité
- Fonction du site
- Type de média représenté
- Couverture du média

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



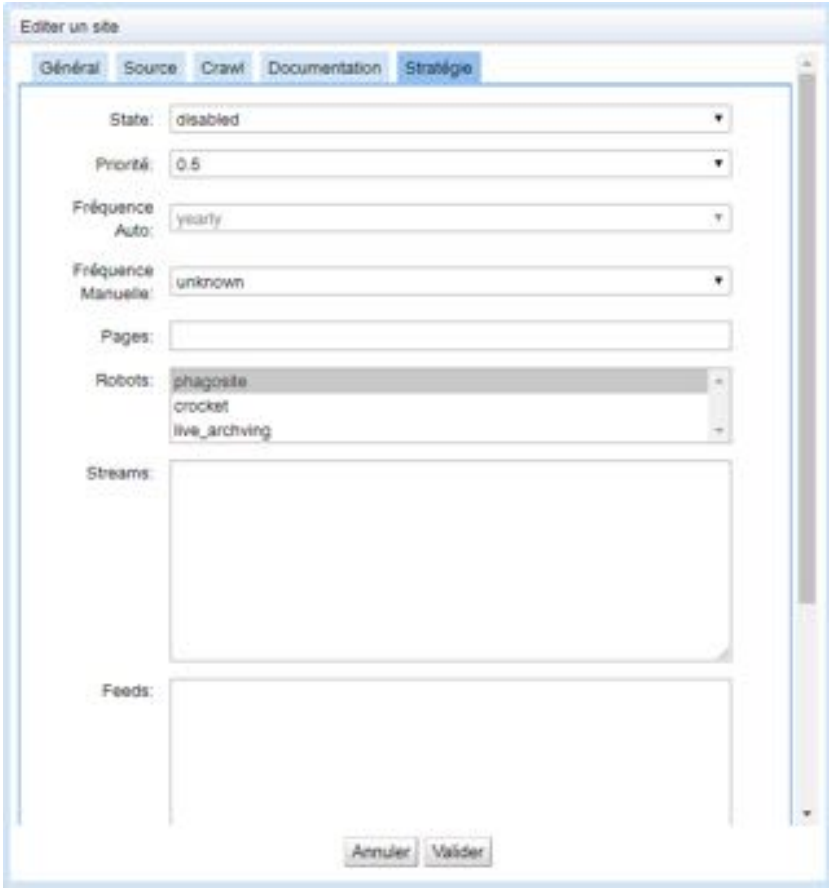
- Personnalité
- Type de programmes
- Public cible du site

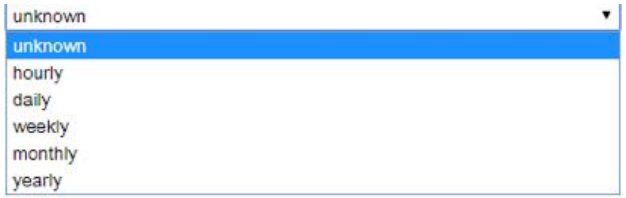
Notice des sites – onglet *documentation* (2/2)



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

Notice des sites – onglet *stratégie*





Exemple de notice en consultation

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

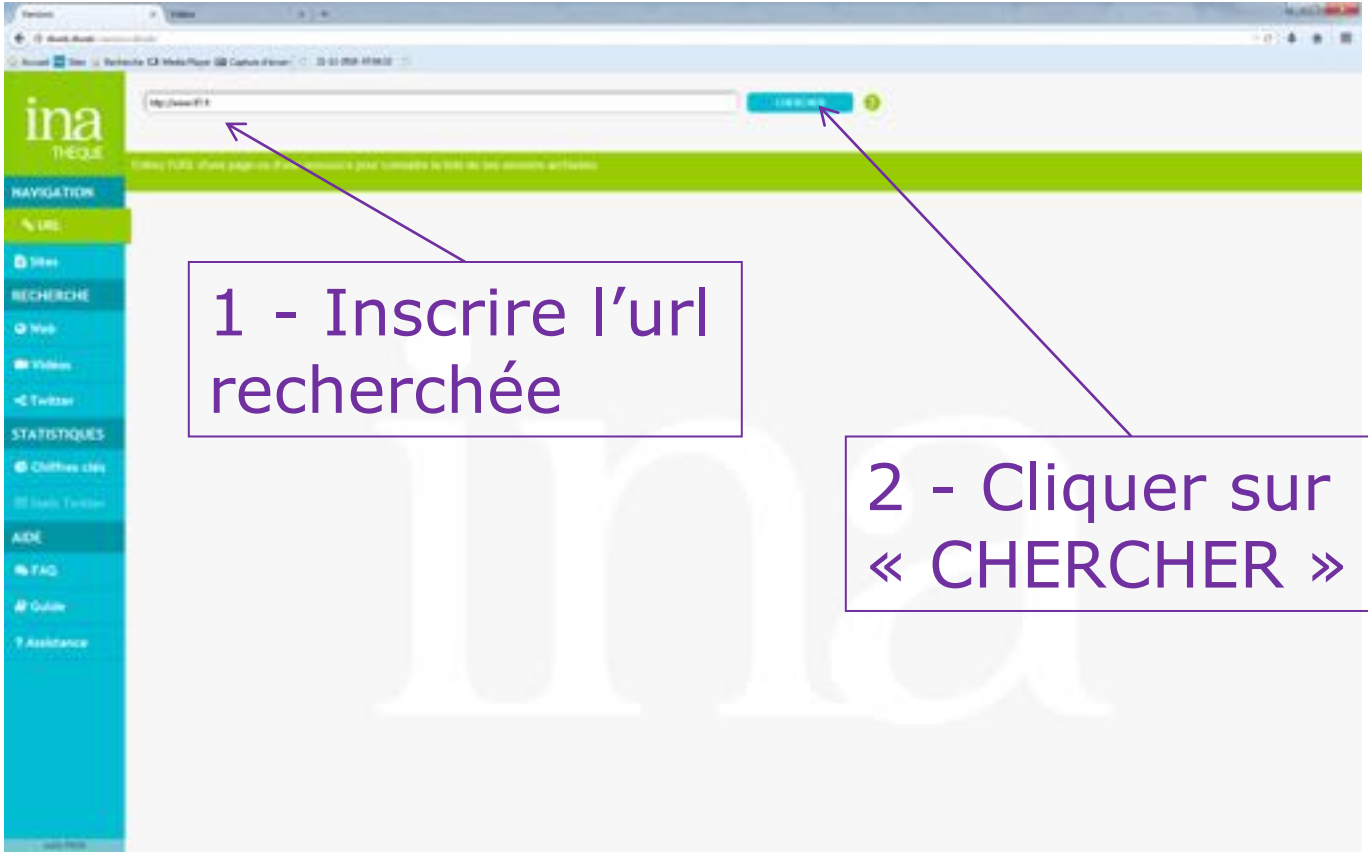
Consultation autonome

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)



Recherche par URL (2/4)

Recherche par URL (3/4)

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:

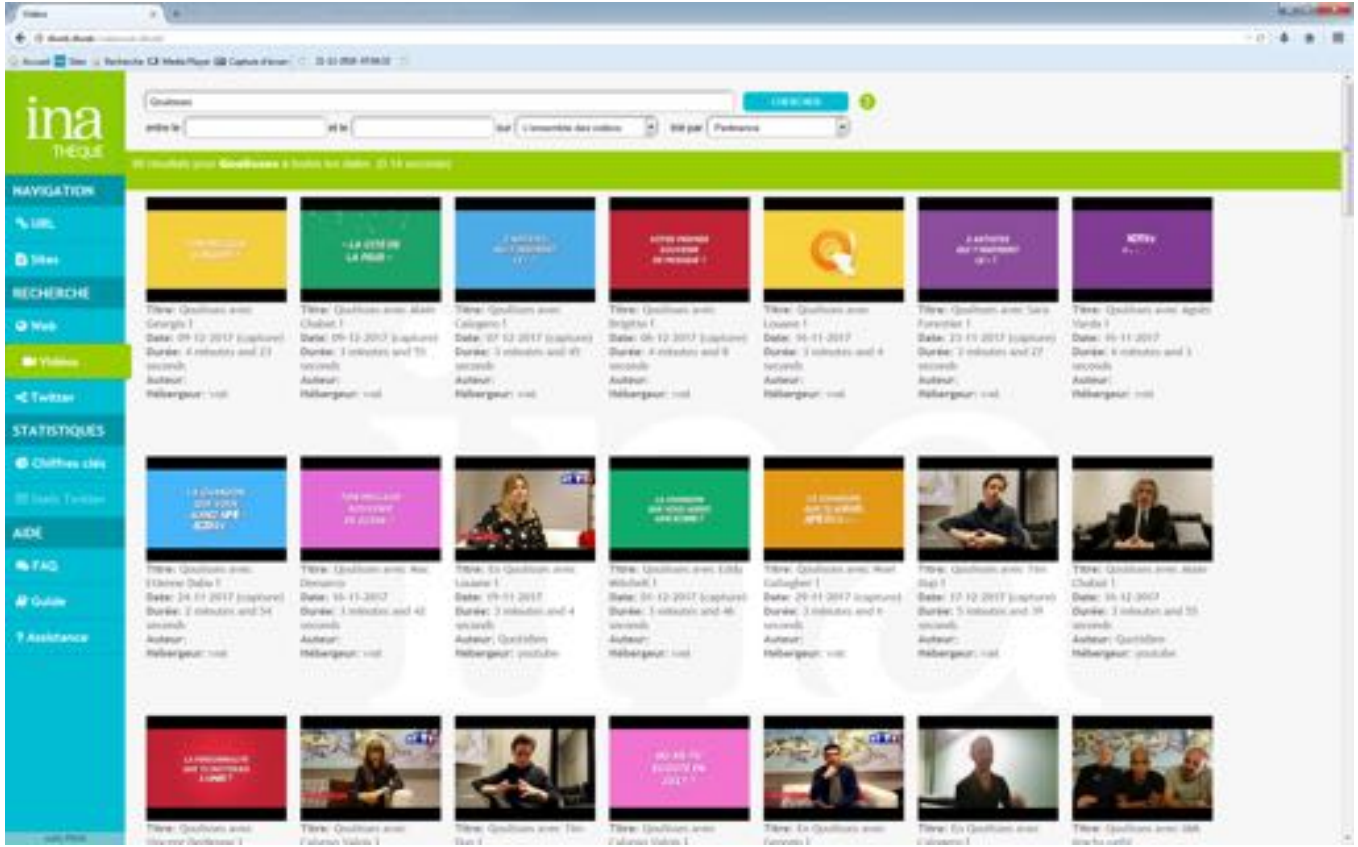
- À partir de l'url de la racine d'un site
- A partir d'un mot contenu dans l'url
- A partir d'un ou plusieurs critères (fonctions, thématique, etc)

Recherche full-text

Permet de chercher toutes les pages contenant un ou plusieurs mots clés

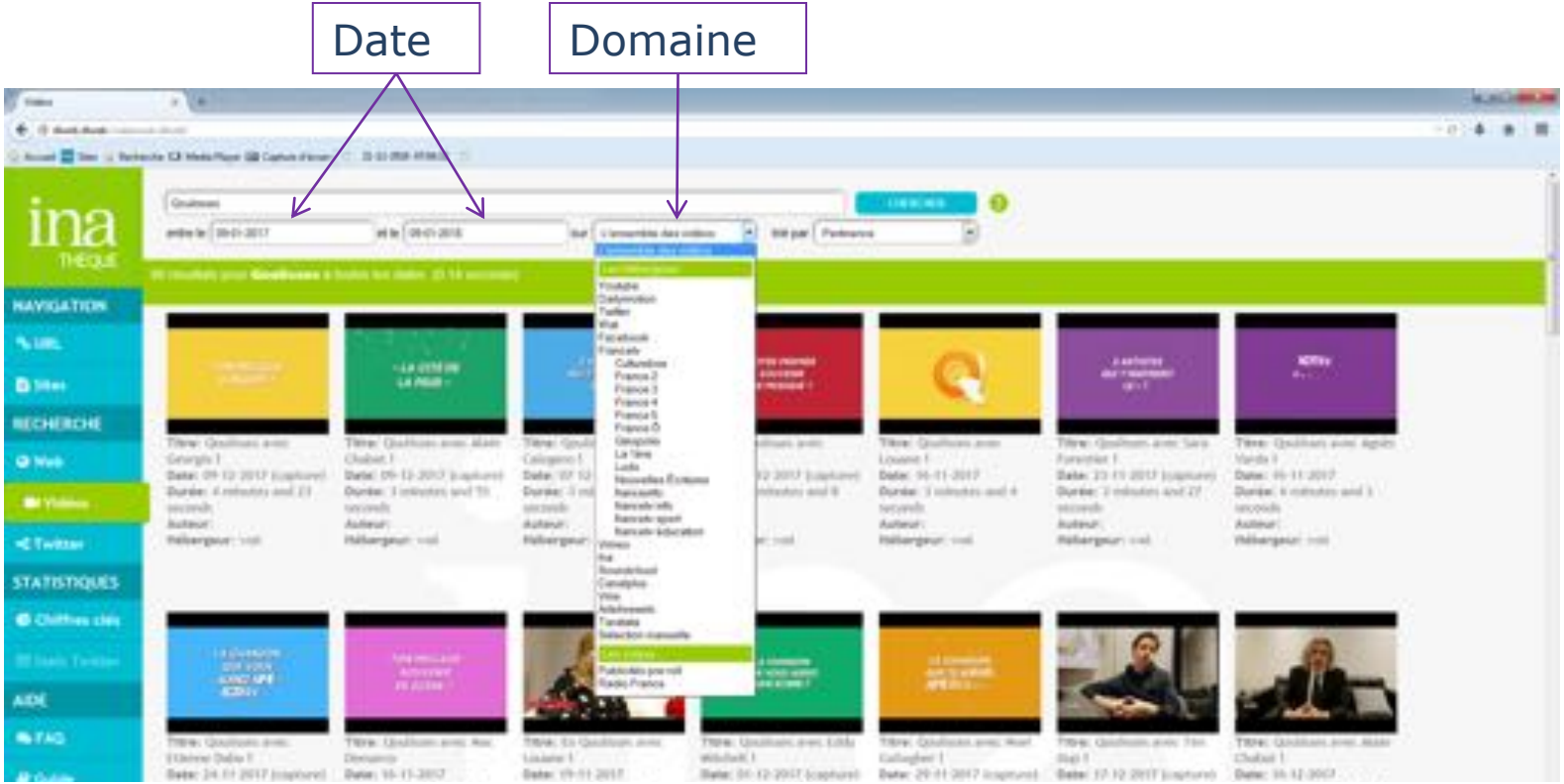
Outil de consultation archive vidéos (1/2)

Permet de consulter les vidéos issues du web et archivées:



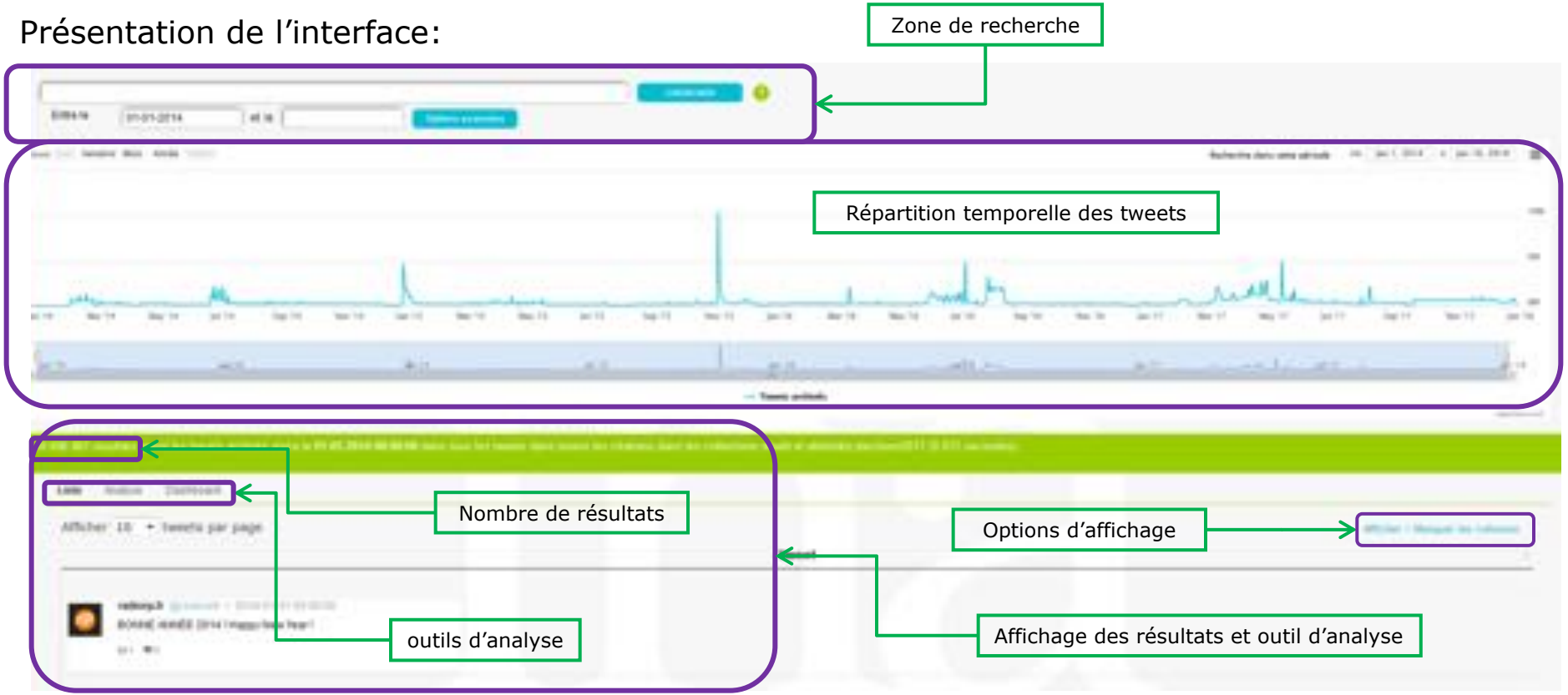
Outil de consultation archive vidéos (2/2)

Possibilité de restreindre la recherche par :



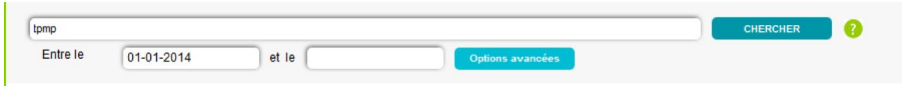
Outil de consultation Twitter (1/2)

Présentation de l'interface:

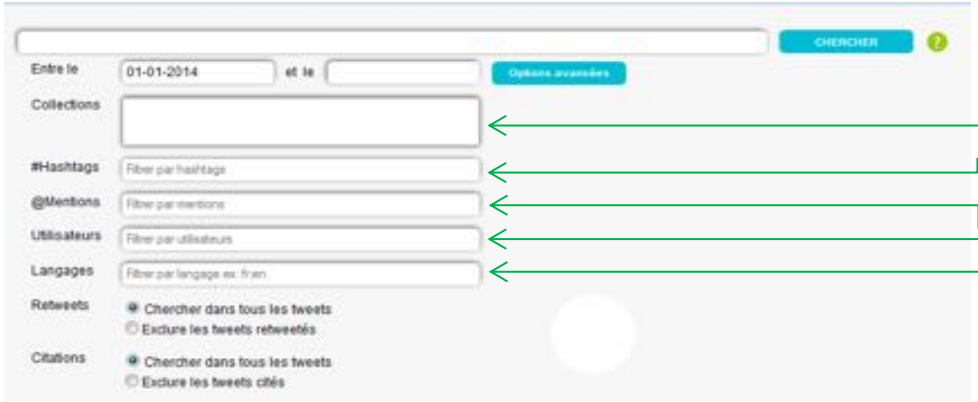


Outil de consultation Twitter (2/2)

→ 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 correspondants 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:

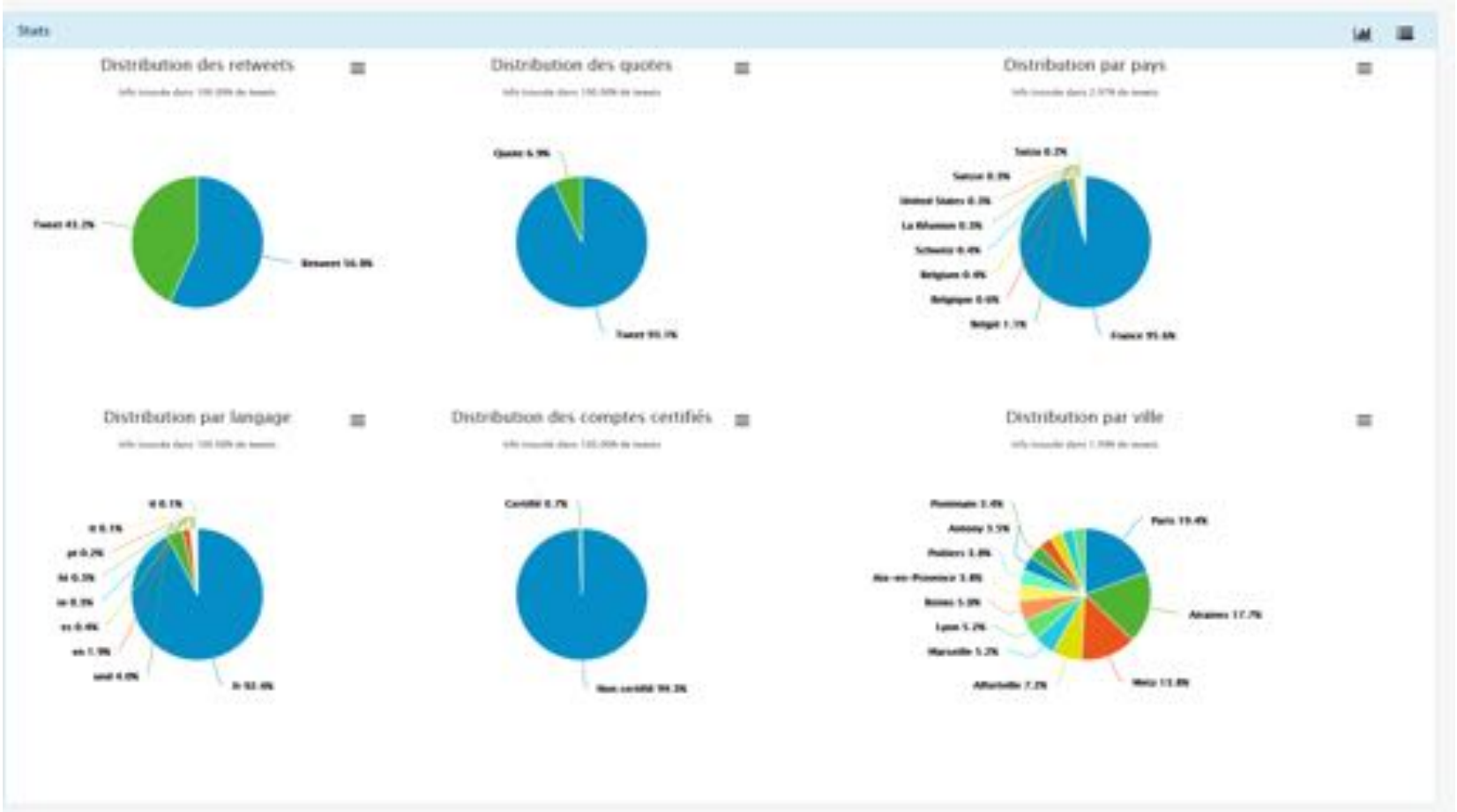
→ Outil mettant à disposition des outils complémentaires pour l’analyse des résultats

- **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)

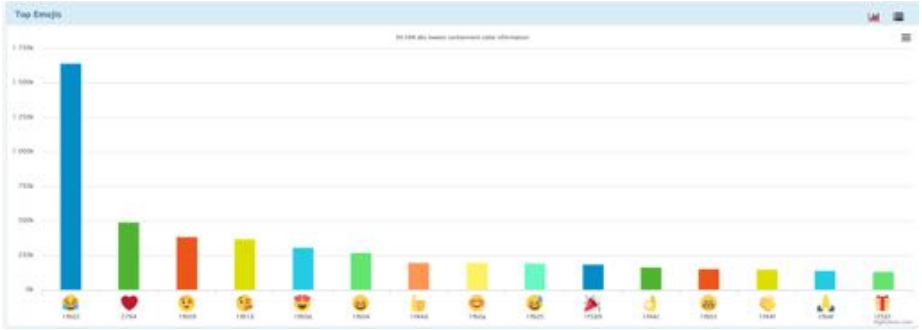
→ Utilisations :

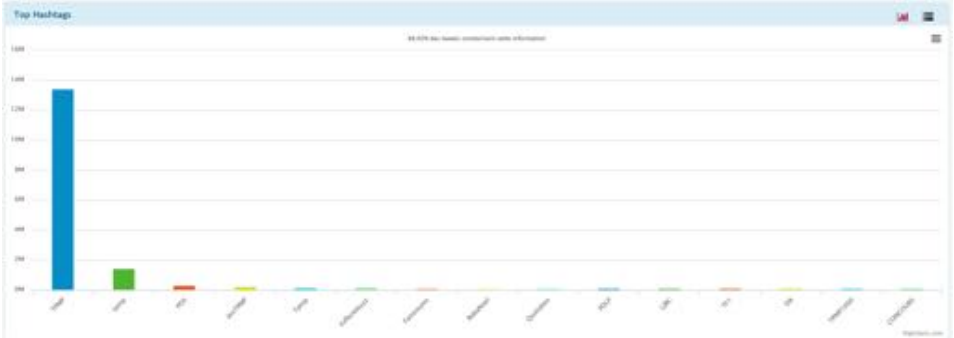
- Analyse géographique, comparative, de sentiments, sérendipité….

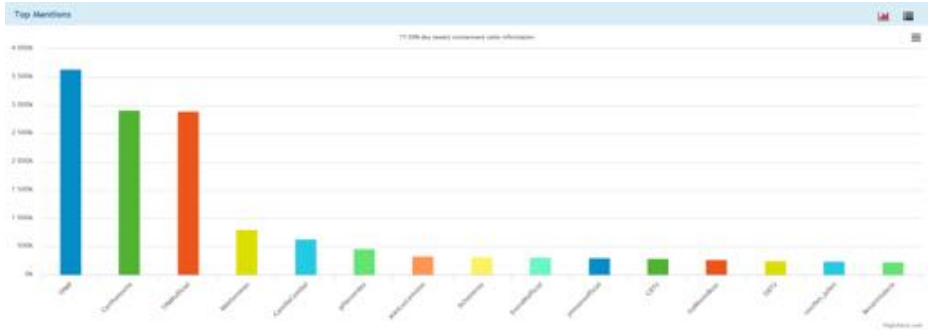
Outils d’analyse Twitter (3/4)



Outils d’analyse Twitter (4/4)







Merci pour votre attention

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

ina
AUGMENTED MEMORY