Archivage du web

Bruno Bachimont, Université de technologie de Compiègne, France

Cadre : le dépôt légal

- Principe :
  - Patrimoine de ce qui a été diffusé à un public;
  - Mémoire de la nation, ce qui appartient à tous et à personne en particulier ;
  - Mémoire pour le futur et non pour l'exploitation utilitaire pour le présent.
- Une invention française:
  - Édit de Montpellier de 1537 (François 1er);
  - Présent dans de nombreux pays, mais pas dans tous.
- Une mise en œuvre en permanente évolution:
  - Chaque nouveau support, mode, procédé de publication pose la question d'un dépôt légal pour cette forme d'expression et de mise à disposition du public.

Étape du dépôt légal en France

- Une histoire riche :
  - 1537 Création du dépôt légal en France Imprimés (Livres)
  - 1648 Estampes dont cartes et plans
  - 1793 Partitions musicales
  - 1881 Périodiques (Loi sur la presse)
  - 1925 Photographies et « toute production d 'arts graphiques »
  - 1941 Affiches
  - 1963 Enregistrements sonores de toute nature
  - 1975 Image fixe et vidéo
    - « quel qu 'en soit le support ou le moyen technique de production »
  - 1992 Edition électronique sur support
    - dont « logiciels, bases de données et systèmes experts »
  - 1992 Dépôt légal audiovisuel, confié à l'INA.
  - 2005 Web
- Extension continue avec deux principes
  - Continuité des collections
  - Extension du champ par la prise en compte de tous les contenus véhiculés la nouvelle technique

La lettre patente de 1537

- On défend « à tous imprimeurs et libraires de mettre ni exposer en vente en notre Royaume soit en public ou en secret toutes les oeuvres nouvellement imprimées, sans qu'un exemplaire ne soit remis entre les mains de l'abbé Mellin de Saint-Gelais, ayant la charge et la garde de notre librairie étant en notre château de Blois ».

- Il est ordonné « de faire retirer, mettre et assembler en notre librairie toutes les oeuvres dignes d'être vues qui ont été et seront faites, compilées, amplifiées, corrigées et amendées de notre temps pour avoir recours aux dits livres, si, de fortune, ils étaient ci-après perdus de la mémoire des hommes ou aucunement immués ou variés de leur vraie et première publication ».

Pourquoi archiver le Web?

- Pas seulement un medium pour transmettre et diffuser des contenus, mais aussi un nouveau moyen de créer des contenus originaux.
- Deux enjeux :
  - Contenus classiques:
    - Le Web permet de récupérer des contenus classiqus mêmes si leur forme éditoriale vient d'autres traditions, imprimés ou diffusions audiovisuelles ;
  - Contenus propres au Web (web-born content) :
    - Le Web permet de trouver des contenus qui ne peuvent être trouvés ailleurs.

Une mémoire du Web

- Une véritable culture est en cours d'émergence avec le Web ;
- Chaque culture renvoie à des enjeux spécifiques de mémoire et relève d'une démarche patrimoniale spécifique.
- L'archivage du Web est désormais un enjeu sociétal et patrimonial.

Contexte français

- La loi sur la propriété intellectuelle ménage une nouvelle exception : le dépôt légal du Web ;
- La BNF et l'INA seront en charge de ce DL;
- La partie INA:
  - Continuer et enrichir les collections audiovisuelles actuelles : (e.g. les stations locales ou régionales)
  - Archiver le Web de la radio/télévision ainsi que les industries culturelles associées.

Le contexte légal

Définir un cadre : l'exemple de la France

Motivations

- Une nouvelle rédaction du décret du 31 décembre 1993 pour :
  - Actualiser la mise en œuvre du dépôt légal
    - de la télévision (extension aux chaînes du câble, du satellite, de la TNT)
    - de la radio (extension aux radios privées généralistes et aux réseaux nationaux thématiques)
  - Mettre en œuvre le DL du web
  - Clarifier les domaines de compétence de l'Ina et de la BNF

La loi de 2006

- Le titre IV de la loi DADVSI du 1er août 2006
  - Article 39 :
    - « [...] 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 [...] »
  - Article 45 :
    - « [...]l'institut est seul responsable de la collecte, au titre du dépôt légal, des documents sonores et audiovisuels radiodiffusés ou télédiffusés ; il participe avec la Bibliothèque nationale de France à la collecte, au titre du dépôt légal, des signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication publique en ligne [...] »
  - Ce texte précise qu'un décret en Conseil d'Etat fixera les conditions de sélection et de consultation après avis de la CNIL (Art 41 – 2)

Dépôt légal du web : domaine INA

- Le domaine Ina :
  - un domaine estimé à environ 35 000 sites

- 5 grandes catégories :
  1. Les sites de radio et de télévision issus d'un média AV préexistant :
     - de 1500 sites à 2000 sites (ex : les sites France TV, de TF1, de BFM…)
  2. Les sites liés aux programmes diffusés sur une chaîne (de 2000 à 3000 sites) :
     - Sites d'émissions ou de séries (ex : Plus belle la vie, Ushuaïa-terre, A la recherche de la nouvelle star…)
     - Sites de personnalités des médias, artistes ou animateurs (ex : Arthur on line).
     - Sites événementiels et blogs liés à l'actualité (ex : site festival de Cannes et environ 2000 blogs liés aux médias)
  3. Les webradios et webTV, environ 5000 sites (ex : Clap TV consacré au cinéma et à la musique, Mizik, la TV des Caraïbes, Art total sur l'art vidéo et infographique…)
  4. les sites en relation directe ou indirecte avec l'activité radio et télévision : sites institutionnels (ex : CSA, sites des sociétés d'auteurs…), de sociétés (ex : Vivendi pour Canal+) ou de prestataires. Ils sont estimés à environ 150 sites auxquels s'ajoute une centaine de sites annuaires.
  5. les sites de partage vidéo, les UGC (DailyMotion et YouTube par exemple) et les blogs diffusant des extraits vidéo (environs 20 000 blogs)

Mettre en œuvre un projet opérationnel

Des initiatives nombreuses

- Dépôt légal :
  - Danemark
  - France
  - Suède,
  - Australie, etc.
- Initiatives internationales:
  - Internet Archive
  - Nedlib
  - Nordic Web Archive
  - Etc.
- Initiatives nationales :
  - UKWAC: UK web archive consortium

Plusieurs approches

- **Périmètre de l'archive**
  - Tous le Web
  - Une partie déterminée, selon différents critères :
    - Linguistique
      - Le suédois, le danois, le français…
    - Territorial
      - Sites .fr,
    - Thématique :
      - Sites médicaux
    - Événementiel :
      - Jeux olympiques, élections (présidentielles)…
- **Stratégie de collecte**
  - Exhaustive
    - Tous les sites du périmètre
  - Sélective
    - Stratégie de filtrage : e.g. algorithme « page ranking » par exemple ;
  - Échantillonnage :
    - Des sites représentatifs du périmètre
- **Procédure de collecte**
  - Captation automatique
  - Dépôt manuel.

De nombreuses difficultés

- Masse importante de données
- Complexité éditoriale:
  - Interactivité;
  - Connectivité.
- Perplexité documentaire :
  - Qu'est-ce qu'un site?
  - Que doit-on indexer ?
    - Site, page, unités graphiques, blocs textuels ?
    - Pas de critères reconnus et consensuels qu'ils soient techniques ou sémiotiques.
  - Comment les indexer ?
    - Quel format, quel standard ?
    - Prendre en compte les versions et le temps.

Pour un résultat étrange

- Difficultés
  - Le temps de captation d'un site est souvent plus long que sa mise en jour
  - Il n'est pas toujours possible de mettre à jour l'intégralité d'un site lors d'une nouvelle captation :
    - fréquence variable selon la profondeur des pages
  - Le contenu n'est accessible qu'en exécutant du code (javascript, Flash):
    - page Web = application
  - Rejouer les contenus anciens dans un environnement actuel
- Conséquences
  - Le site archivé et consulté n'a jamais existé comme tel ;
  - Le site consulté d'une journée donnée ne correspond pas forcément à la version de ce jour dans sa profondeur
  - Des parties / segments ne sont pas/plus consultables :
    - Formats techniques obsolètes
    - Des contenus générés par du code qu'il faut reconnaître et exécuter (javascript)

Et trois niveaux d'objectivité à confronter

- Les réalités socio-culturelles qu'on veut étudier
- Le Web d'un moment temporel qui en était une manifestation
- L'archive qu'on a constituée.

- Naïvement, l'enjeu est d'exploiter l'archive réalisée pour l'interroger et trouver des réponses sur des réalités étudiées :
- Possible en principe : l'archive n'est après tout que la trace laissée par ces réalités, elle en est l' « archive »

Différences avec les archives

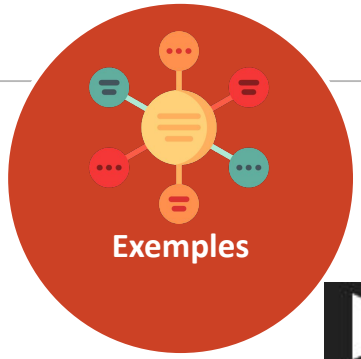
- L'archive du Web n'est pas organique, il relève d'un choix éditorial :
  - S'inscrit dans le périmètre du DL (pour la France par exemple) ;
  - Relève du droit d'auteur ;
- L'archive du Web n'a pas de rôle probatoire comme tel d'un événement survenu ;
- Pas un fonds (fons = source) mais une sélection.
- Stricto sensu : l'archivage du Web relève de la bibliothéconomie.

L'approche INA

Apud Thomas Drugeon, Jérôme Thièvre, INA ©

Périmètre : Sites Web

20 milliards de pages à partir de 15 000 sites web :
- **éditeurs TV/Radio**
- **sites professionnels**
- **sites & blogs de fans**



MY TFI

france•tv

6play

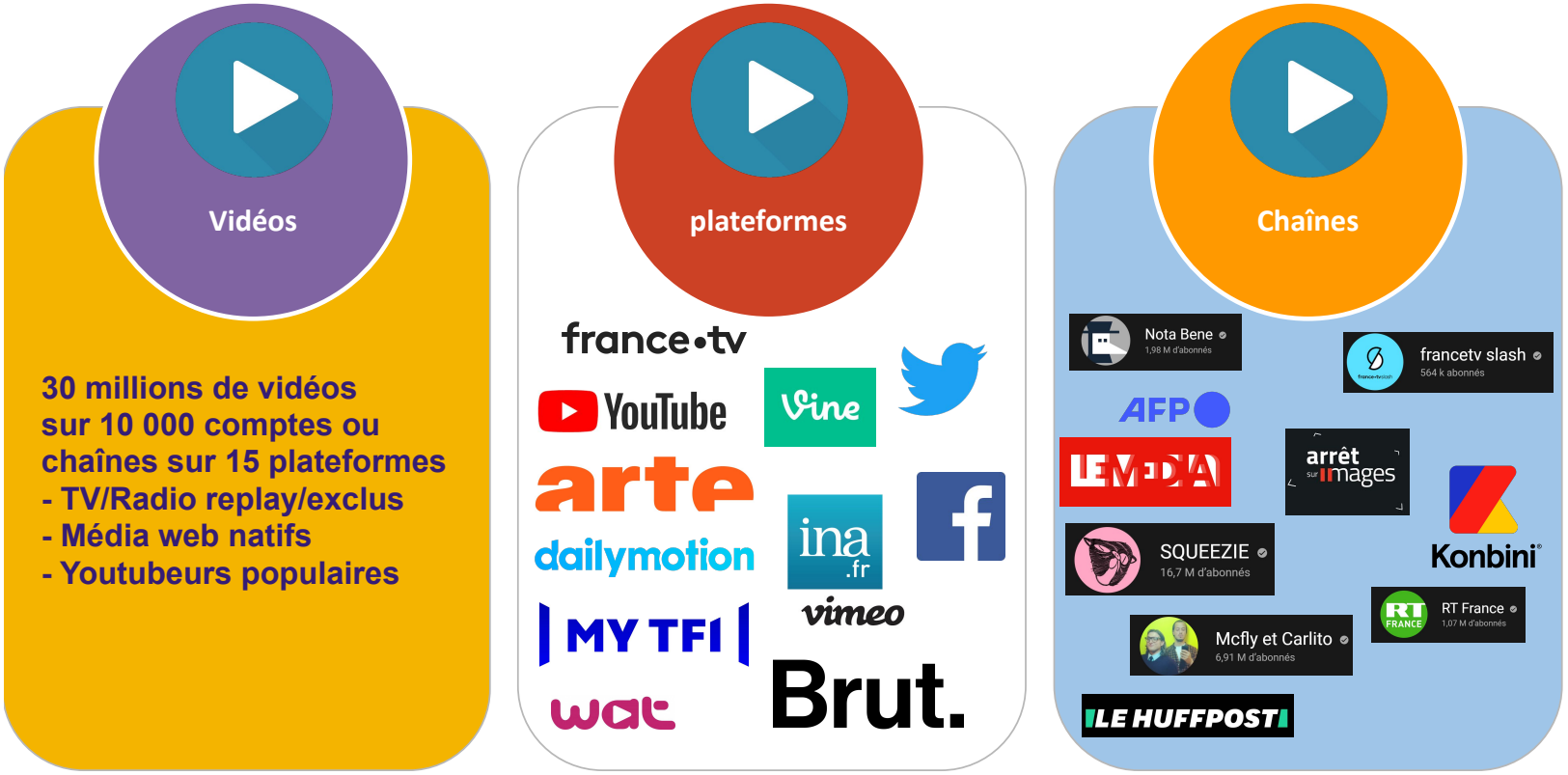
Brut.

Chérie FM

TLM

GRENOUILLE 88.8fm MARSEILLE grenouille888.org

Périmètre : Vidéo



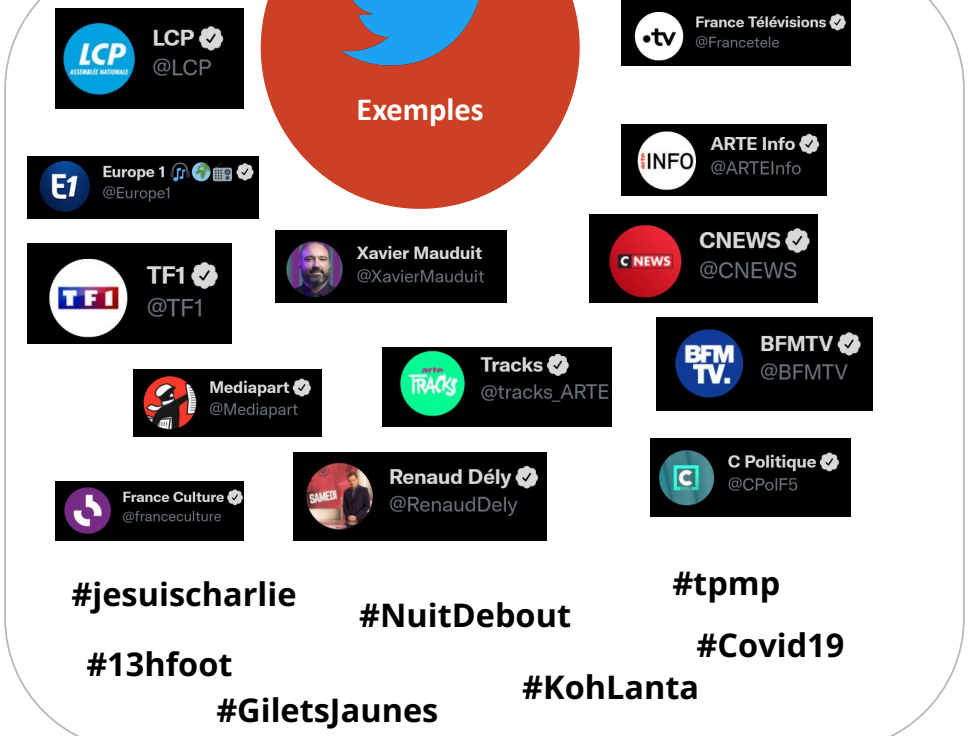
Périmètre : audio et podcasts

Périmètre : Twitter

Twitter

2.5 milliards de tweets à partir de 13 500 comptes & 2 700 hashtags :
- chaînes, émissions et personnalités TV/Radio
- actualités des médias
- grands évènements
- comptes @elysee, PR

Exemples



1. Principe

1. Définir, catégoriser et faire évoluer une liste de sites pertinents

2. Archiver ces sites à des intervalles de temps adaptés

3. Proposer des enrichissements pour l'analyse de cette archive

1. Mettre en place une consultation de cette archive

1. Principe

1. Définir, catégoriser et faire évoluer une liste de sites pertinents

2. Archiver ces sites à des intervalles de temps adaptés

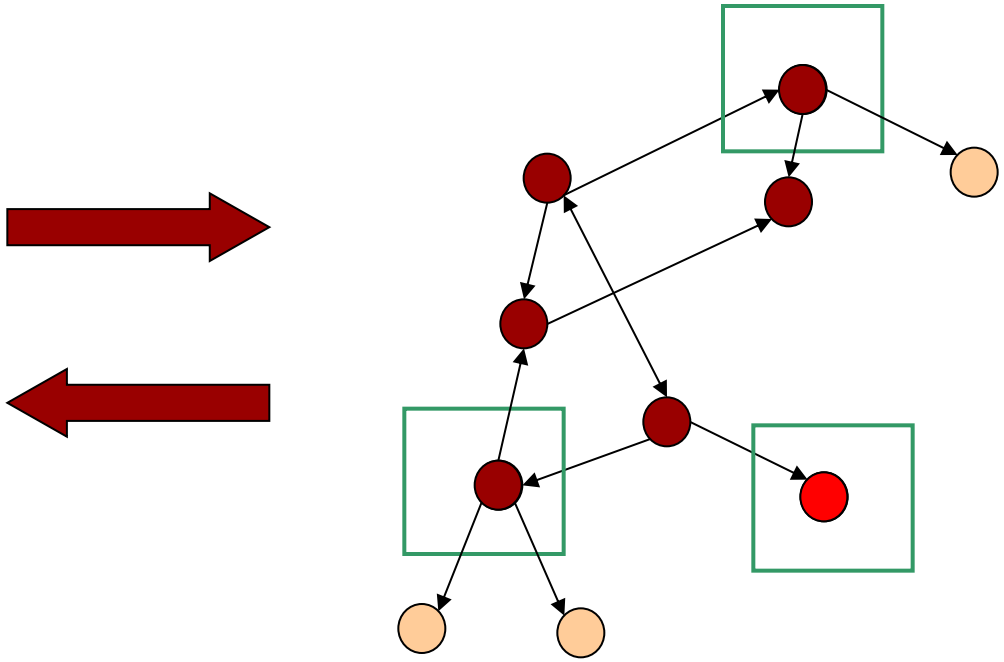
1. Proposer des enrichissements pour l'analyse de cette archive

2. Mettre en place une consultation de cette archive

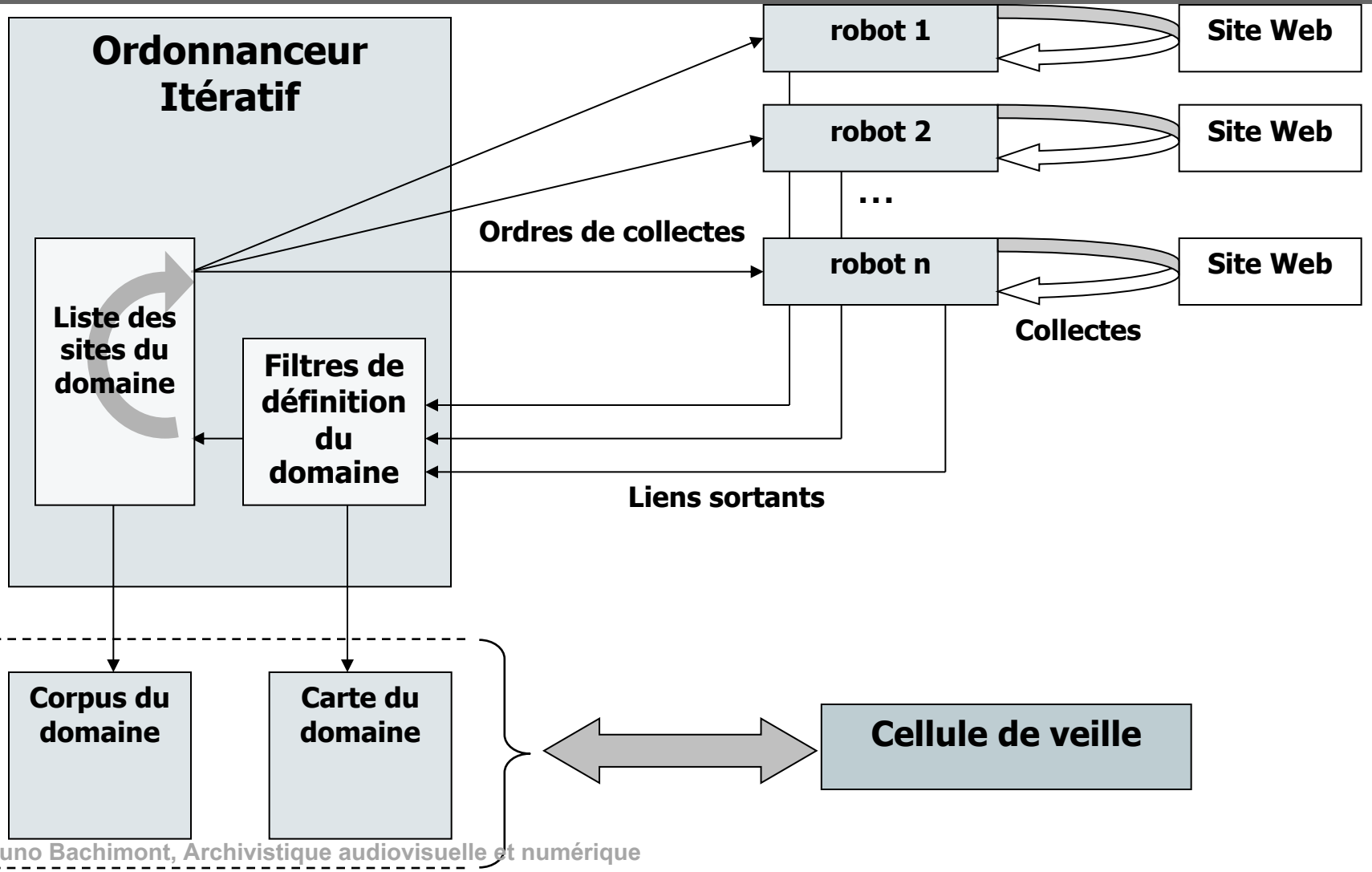
1. Définition

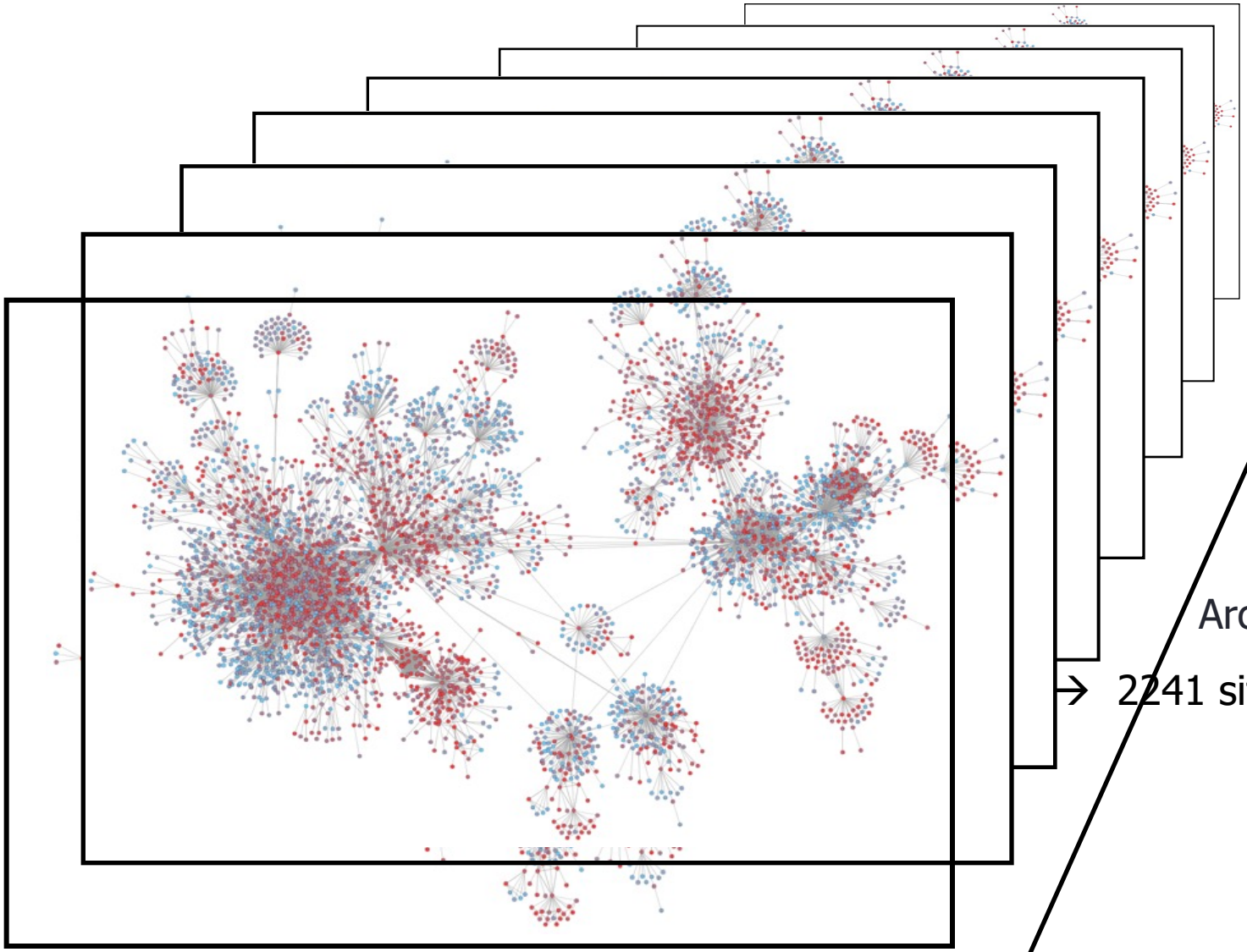
Liste de sites

blabla.com
fdsdgfg.fr
dfsqffqqdsf.fr
blabla.com
dfsqffqqdsf.fr
blabla.com



Prospection itérative du domaine





Archivage

→ 2241 sites

Prospection

Principe

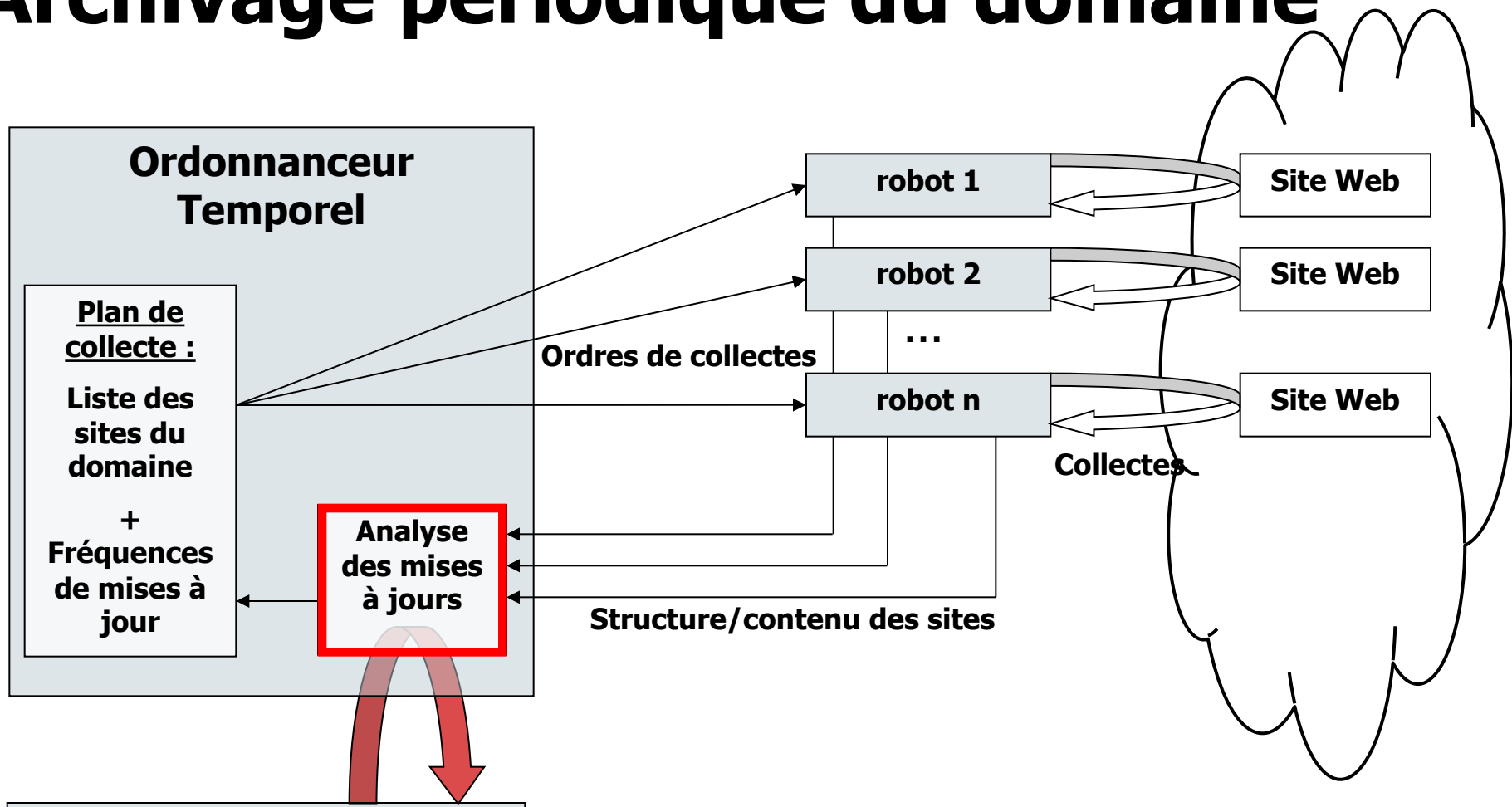
1. Définir, catégoriser et faire évoluer une liste de sites pertinents

2. Archiver ces sites à des intervalles de temps adaptés

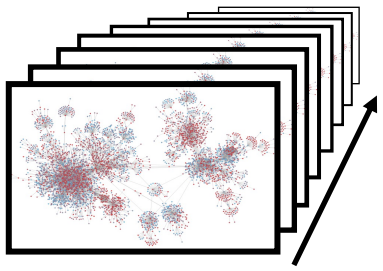
3. Mettre en place une consultation de cette archive

4. Proposer des enrichissements pour l'analyse de cette archive

Archivage périodique du domaine



2. Archivage

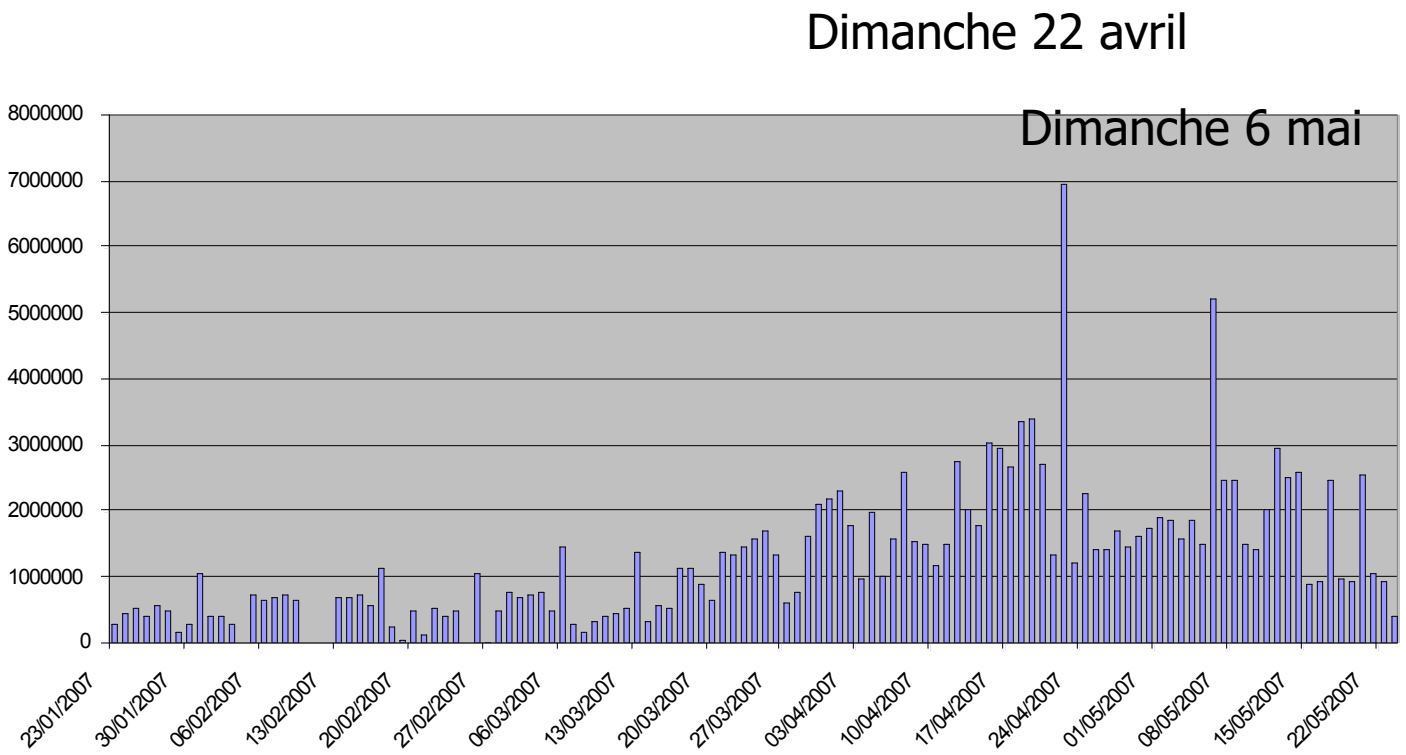


- multiples granularités
- collectes différentielles
- stocké en DAFF

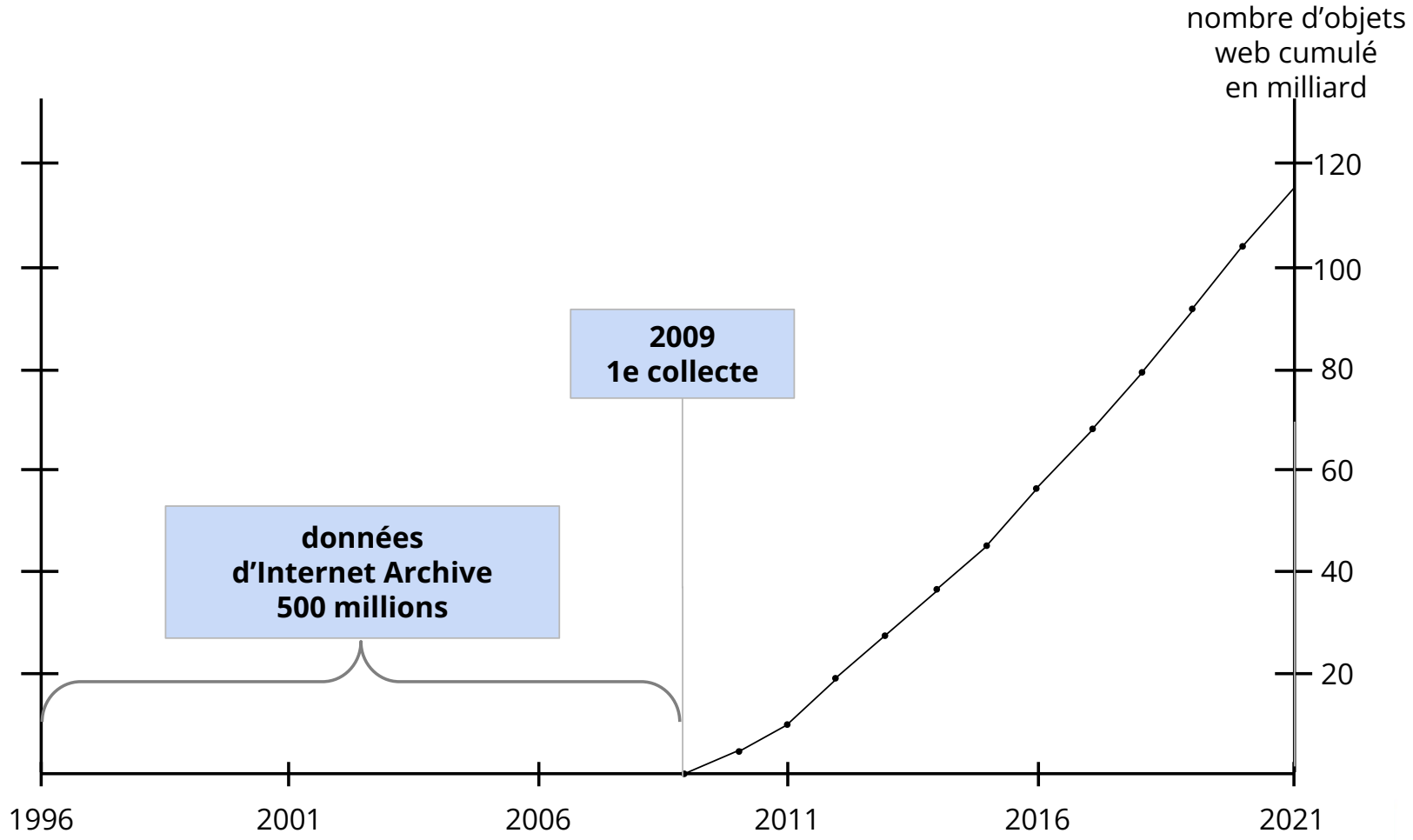
Campagne des présidentielles 2007

- 6 mois de collecte à rythme croissant
- 158 millions d'objets archivés
- Environ 1 To de stockage DAFF
- 28 000 vidéos, 517 Go

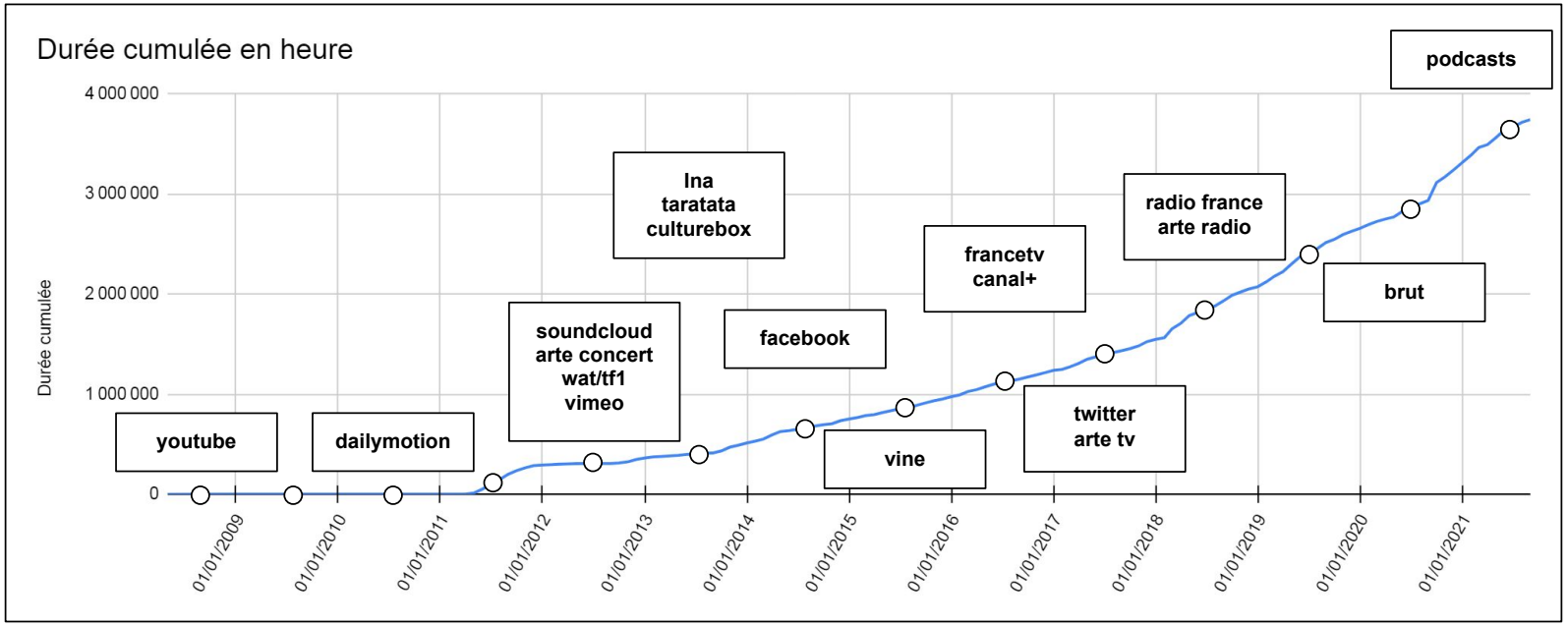
Exemple de collecte événementielle: élections présidentielles 2007



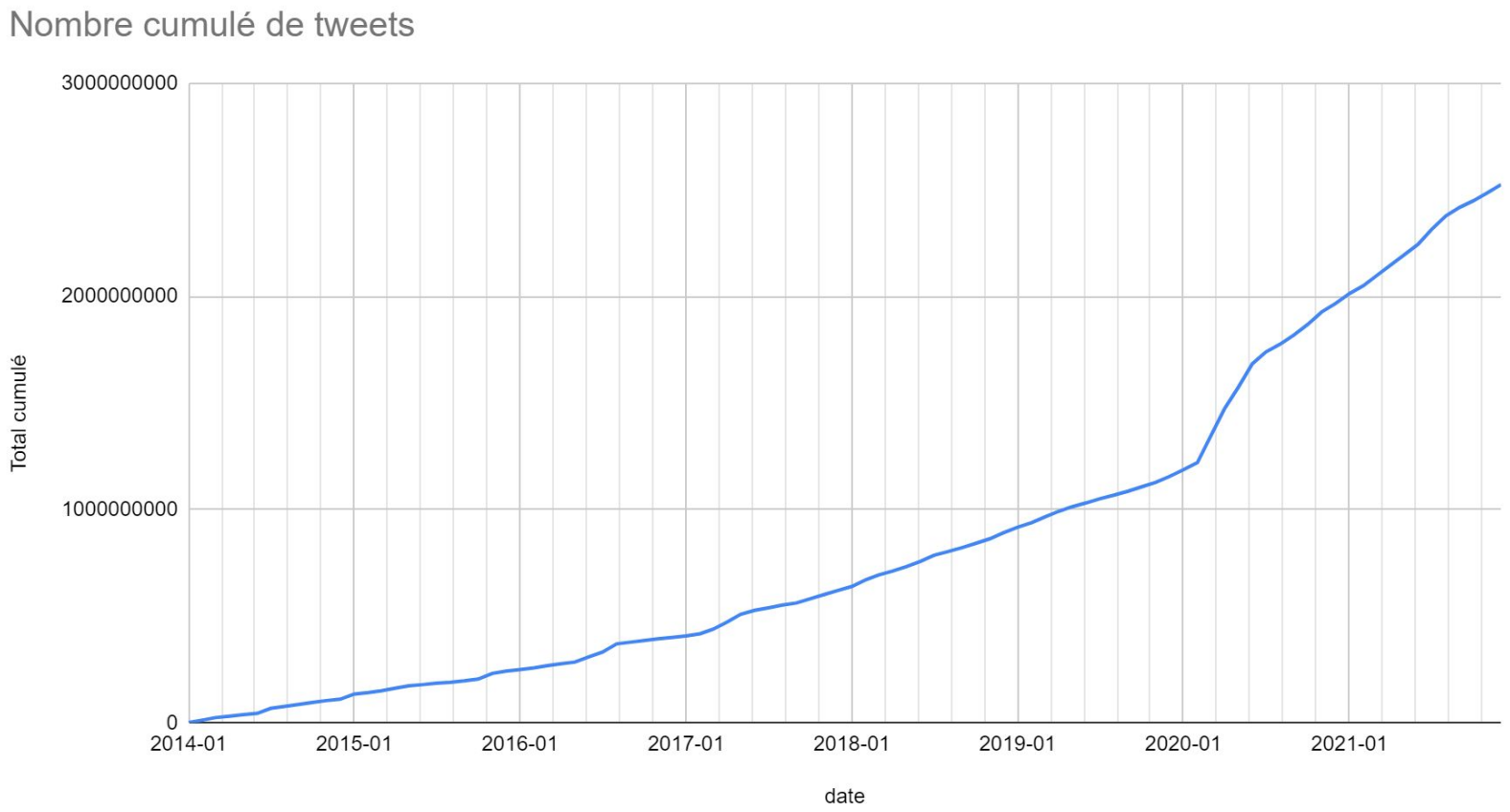
Evolution sur les années



Evolution audio / vidéo



Évolution Twitter



Plus sur DAFF : Digital Archiving File Format

- Format de fichier de type container de records
- Simple et auto-décrit
- Test d'intégrité intégré
- Agnostique au protocole (http(s), ftp, autre)
- Déduplication complète

DAFF : data / metadata

daff data records

daff metadata records

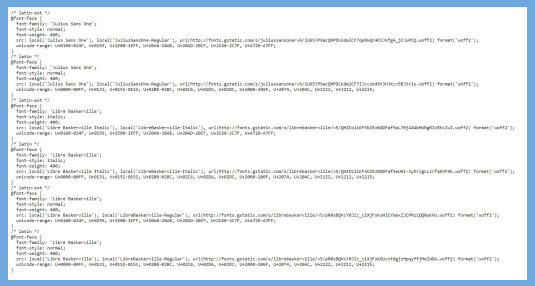
sha_256: e4ba78b2c0034f

```
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
```

sha_256: babc4e3130004def6



```
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
```

DAFF, en chiffres

- 129 milliards records
  - 116 milliards de métadata
  - 13 milliards de data
- Déduplication et compression:
  - 12 PB de données collectées et archivées
  - 2.5 PB de stockage
- DAFF tools
  - gestion des fichiers
  - vérifications d'intégrité

Le Streaming

Le *Streaming*

- Radio et télévision sur le Web

Qu'est-ce que le *streaming* ?

- Dans le monde de la production…
  - ~~Flux audio/video, accessible "immédiatement"~~
  - ~~Le client ne peut pas copier le flux~~
- Deux types de *streaming*
  - Extraits *streamés*
  - *Streaming live* → flux "infinis"

➔ Deux approches différentes

Extraits *streamés*

- Durée finie
- Identique à chaque consultation
- Unicast à 100%

➔ Peut être assimilé à du téléchargement

- Protocole spécifique (RTSP, MMS, ICY, …)
- Débit ≈ temps réel

➔ Prise en charge possible par le robot

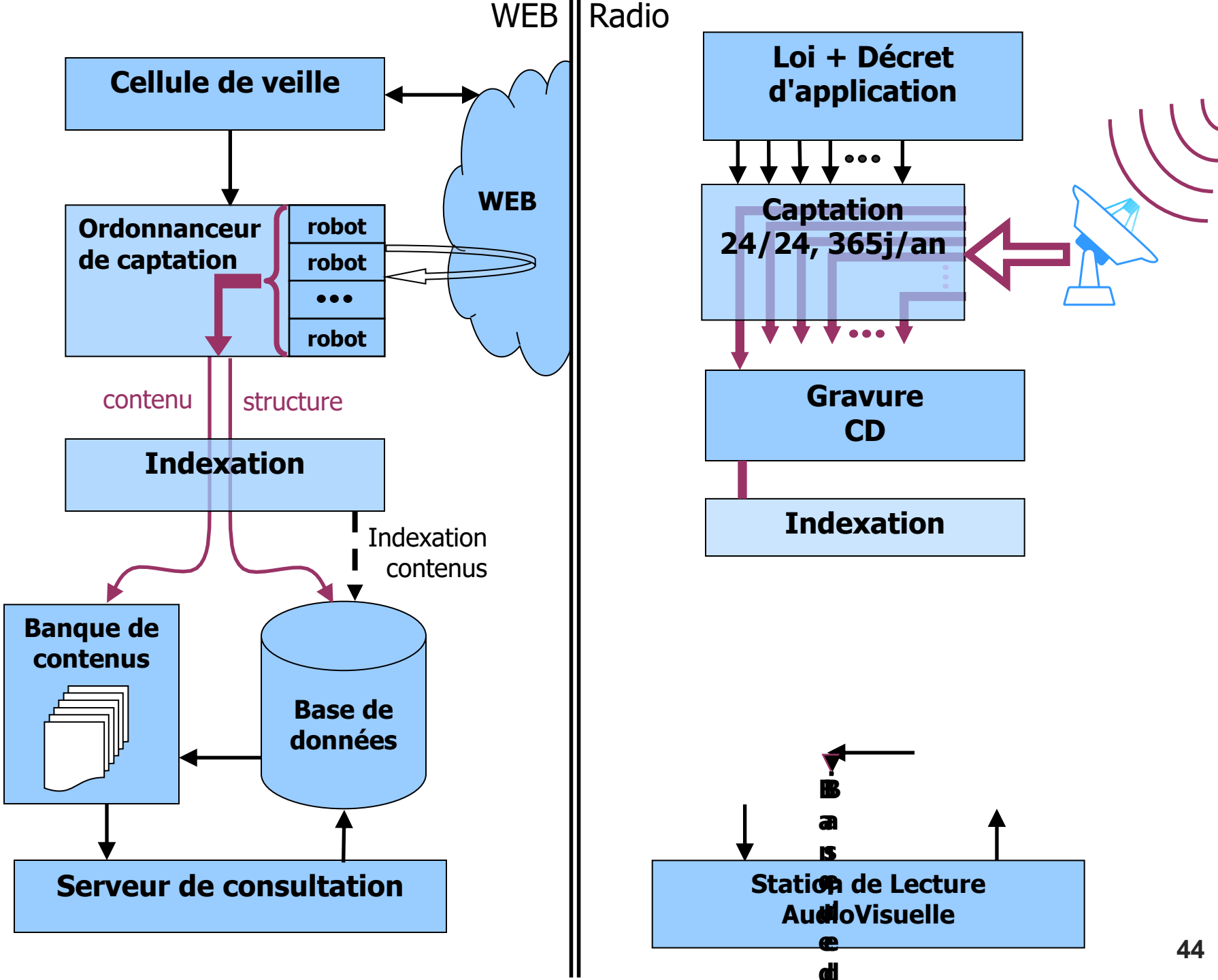
Streaming live

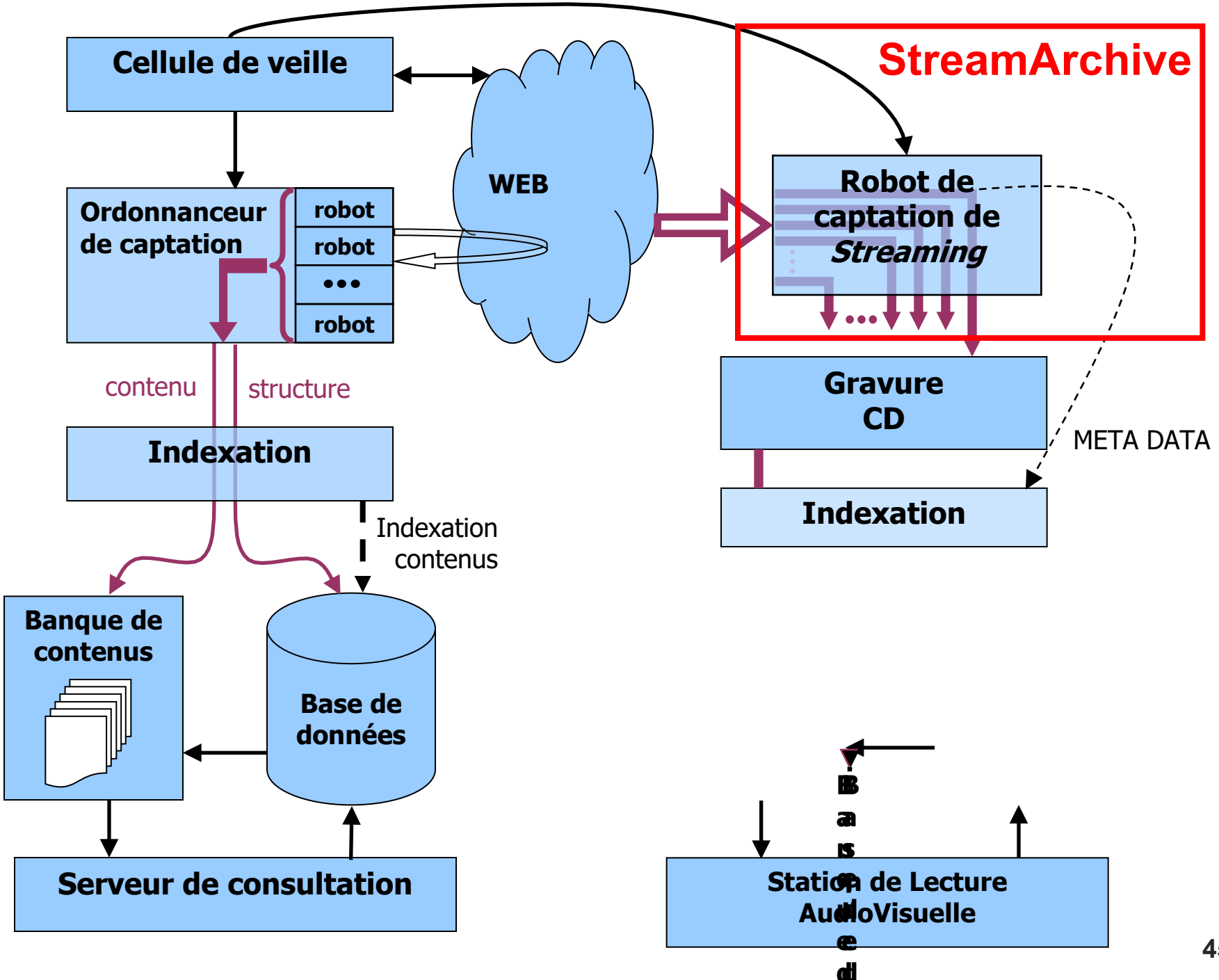
- Potentiellement infini
- Différent à chaque consultation
- Identique pour tous les utilisateurs

➔ Peut être assimilé à un flux radio/TV

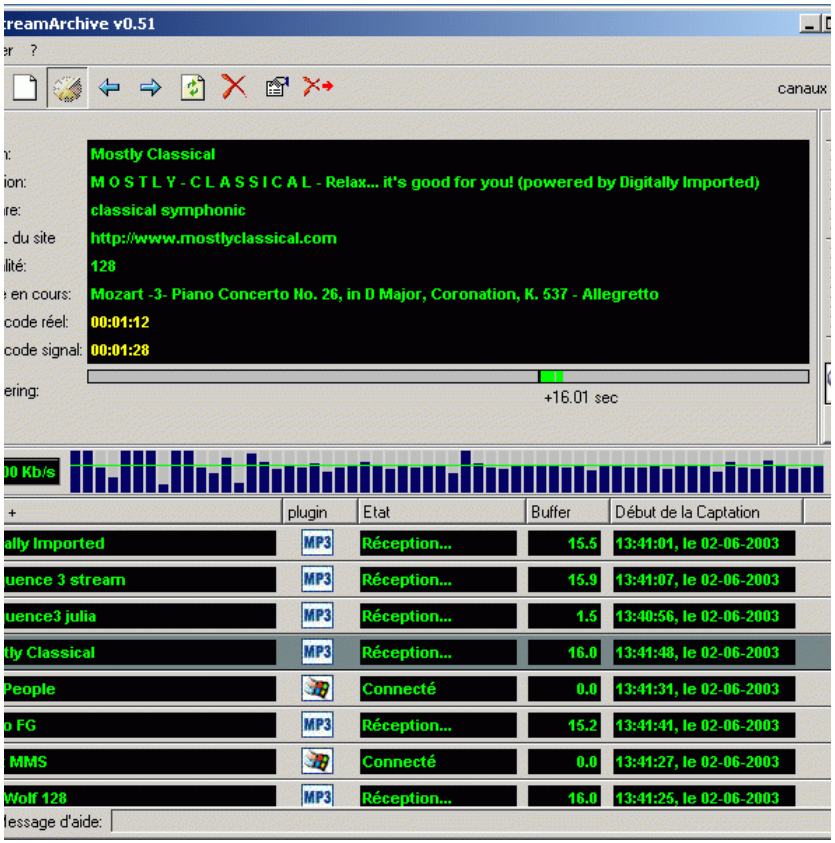
- Moins fiable (interruptions réseaux)
- Moins homogène (multiples formats)
- Plus volatile (changement d'adresse, …)

➔ Traitement spécifique dans la chaîne Web

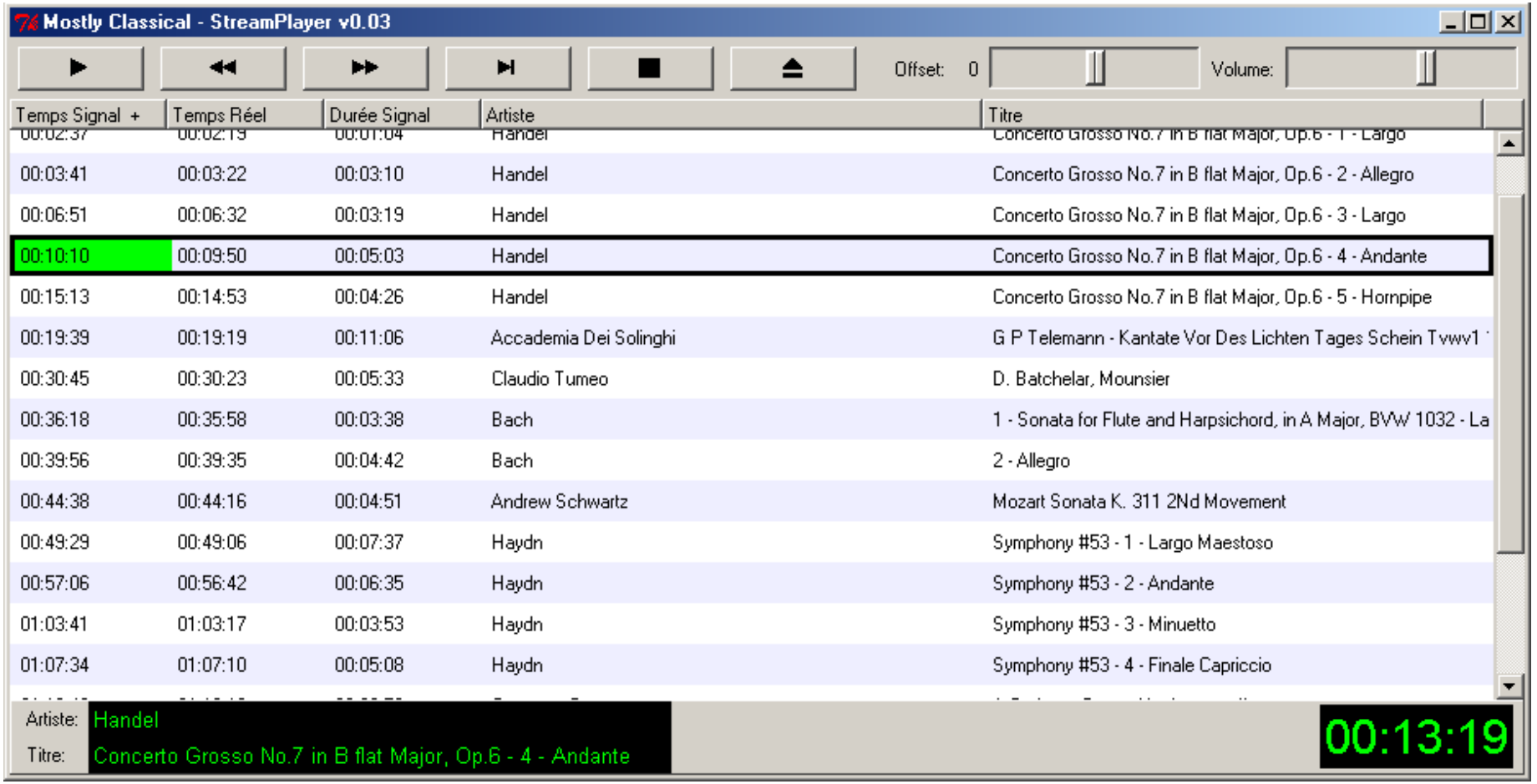

WEB
Radio
Cellule de veille
WEB
Ordonnanceur de captation
robot
robot
robot
contenu
structure
Indexation
Indexation contenus
Banque de contenus
Base de données
Serveur de consultation
Loi + Décret d'application
Captation 24/24, 365j/an
Gravure CD
Indexation
Station de Lecture AudioVisuelle


Cellule de veille
StreamArchive
WEB
Ordonnanceur de captation
robot
robot
•••
robot
Robot de captation de Streaming
contenu
structure
Gravure CD
META DATA
Indexation
Indexation
Indexation contenus
Banque de contenus
Base de données
Serveur de consultation
Station de Lecture AudioVisuelle

StreamArchive: captation radio



StreamPlayer – Interface de conultation



Principe

1. Définir, catégoriser et faire évoluer une liste de sites pertinents

2. Archiver ces sites à des intervalles de temps adaptés

1. Proposer des enrichissements pour l'analyse de cette archive

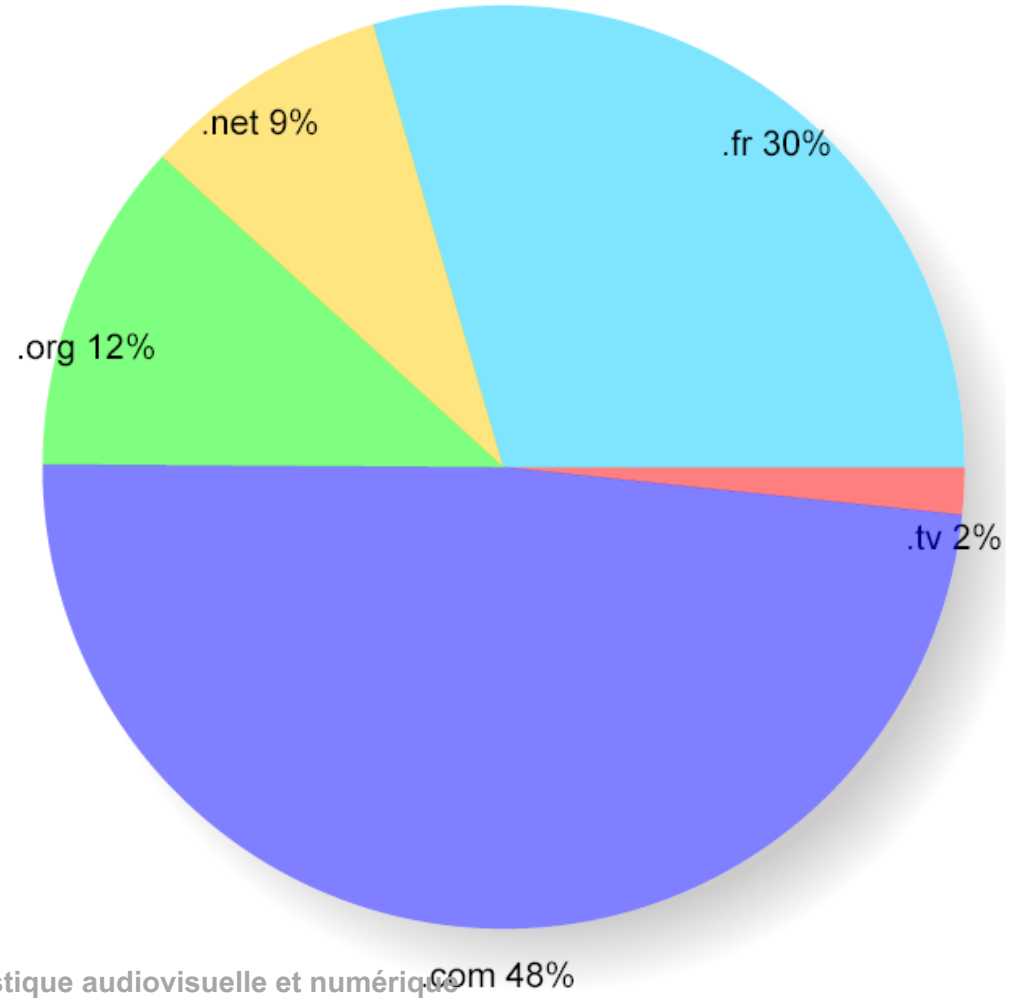
2. Mettre en place une consultation de cette archive

Outils de visualisation

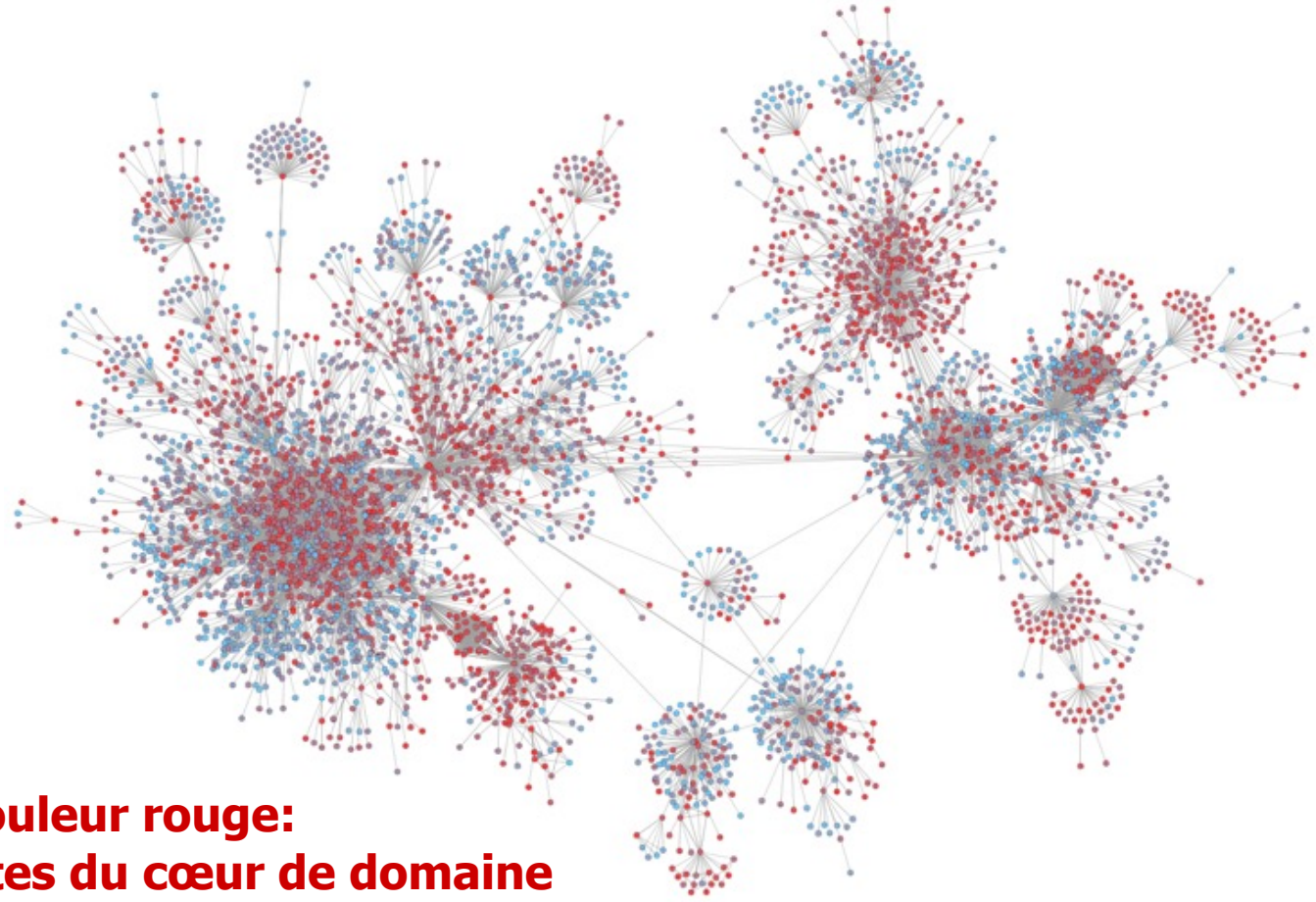
Topologie du domaine des sites médias

Visualisation de graphes : les sites sont représentés par des points reliés par des hyperliens.

Réparation des TLD dans le domaine

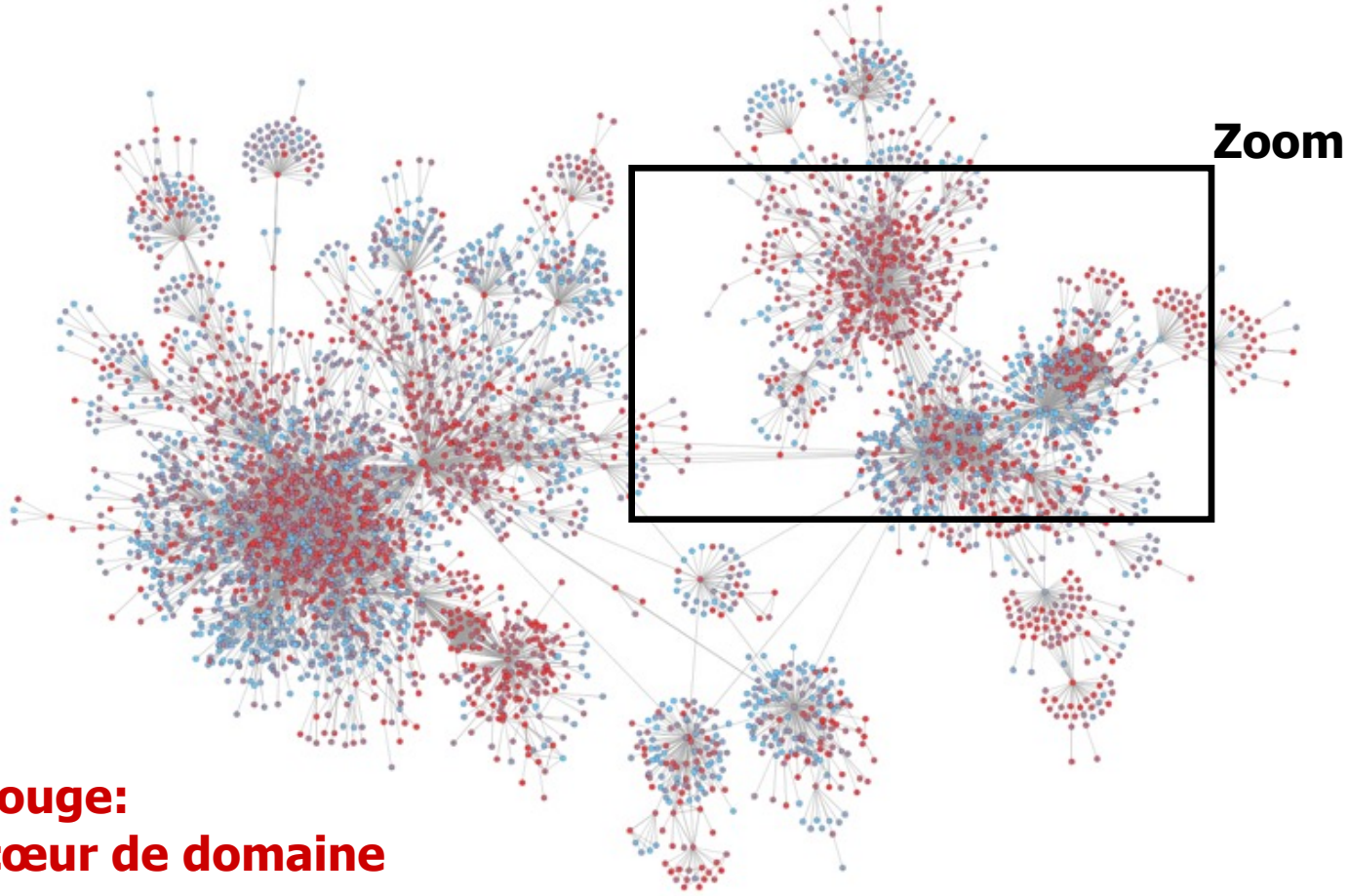


Topologie du domaine des sites médias



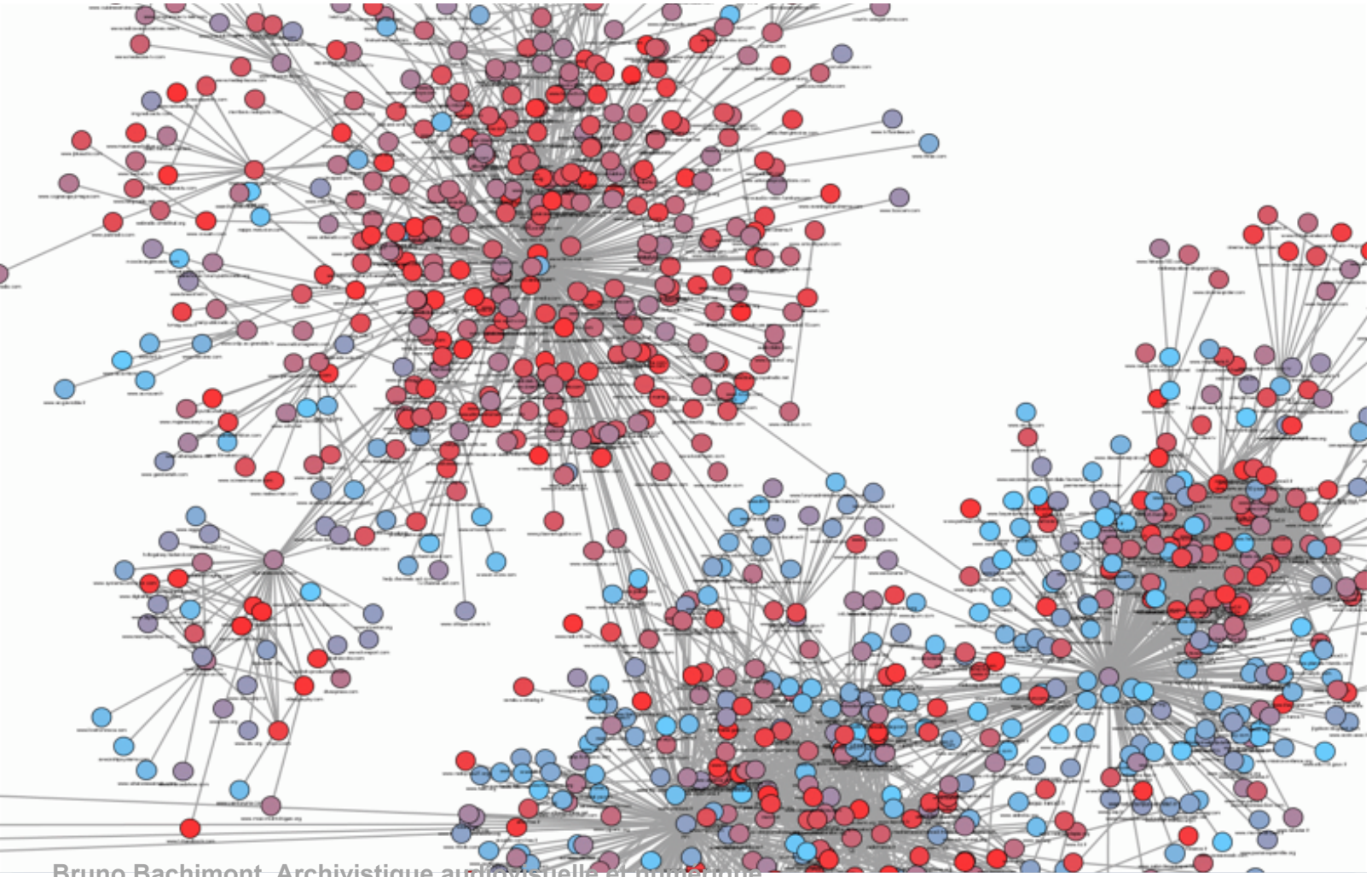
Couleur rouge:
sites du cœur de domaine

Topologie du domaine des sites médias

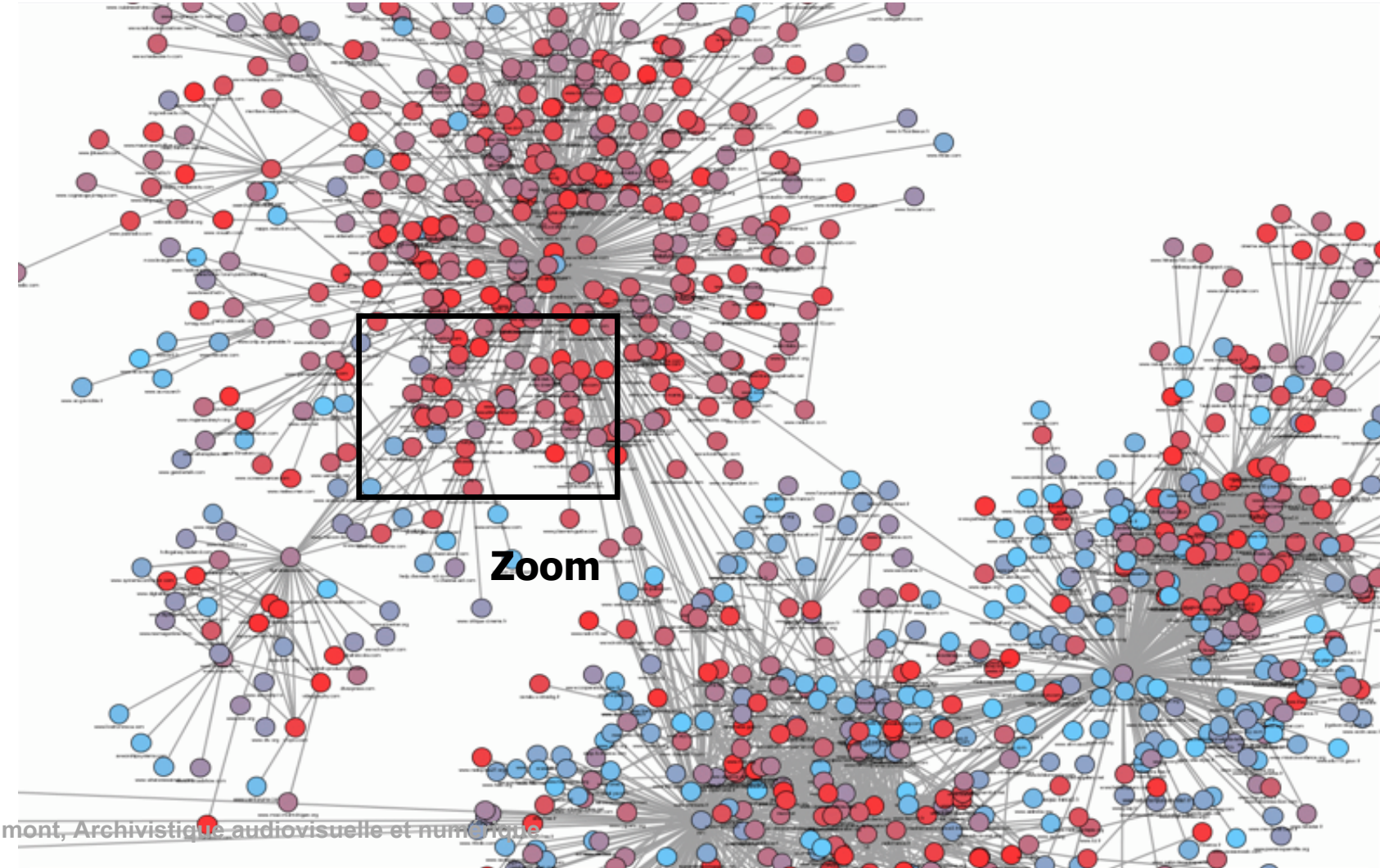


Couleur rouge:
sites du cœur de domaine

Topologie du domaine des sites médias



Topologie du domaine des sites médias

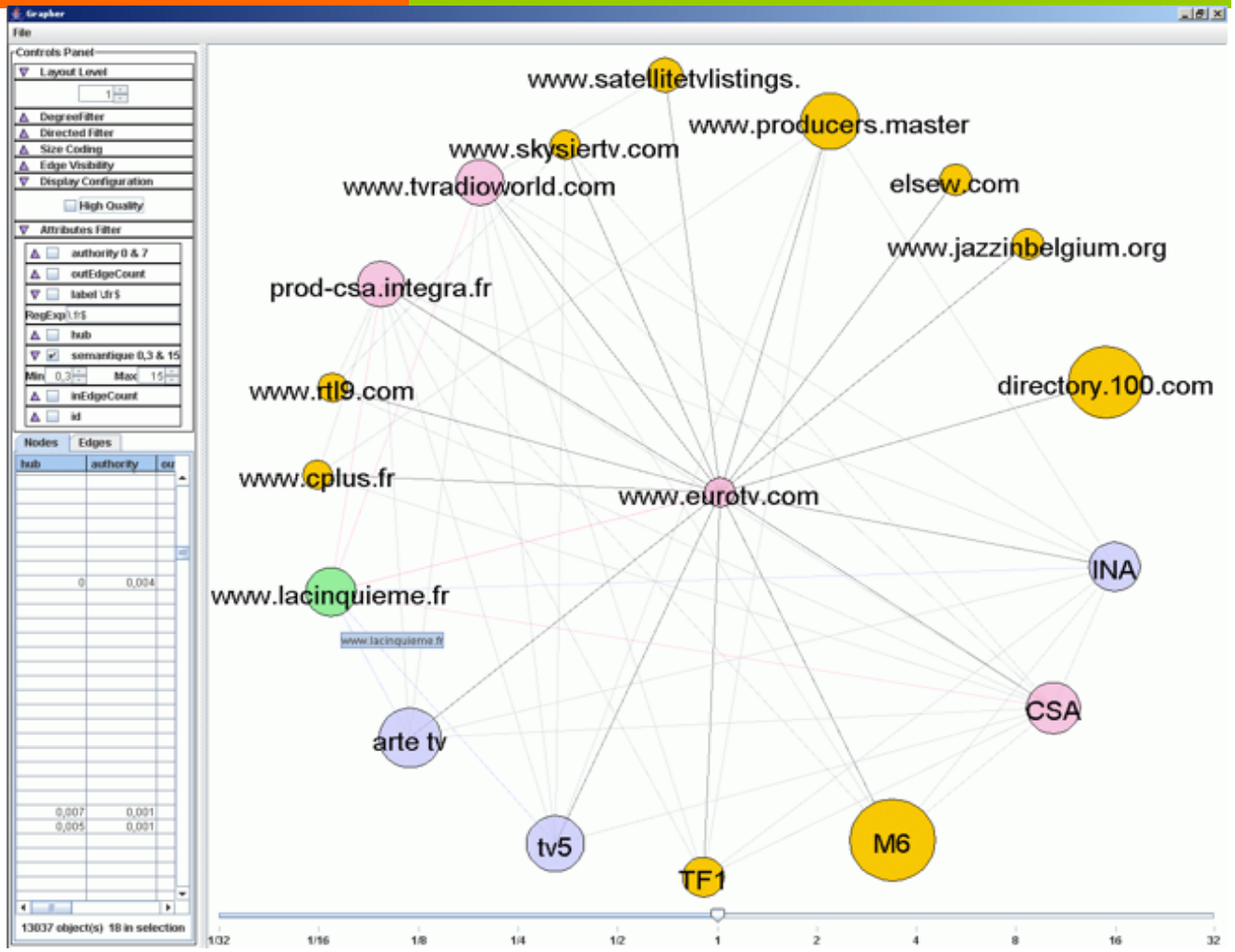


Topologie du domaine des sites médias

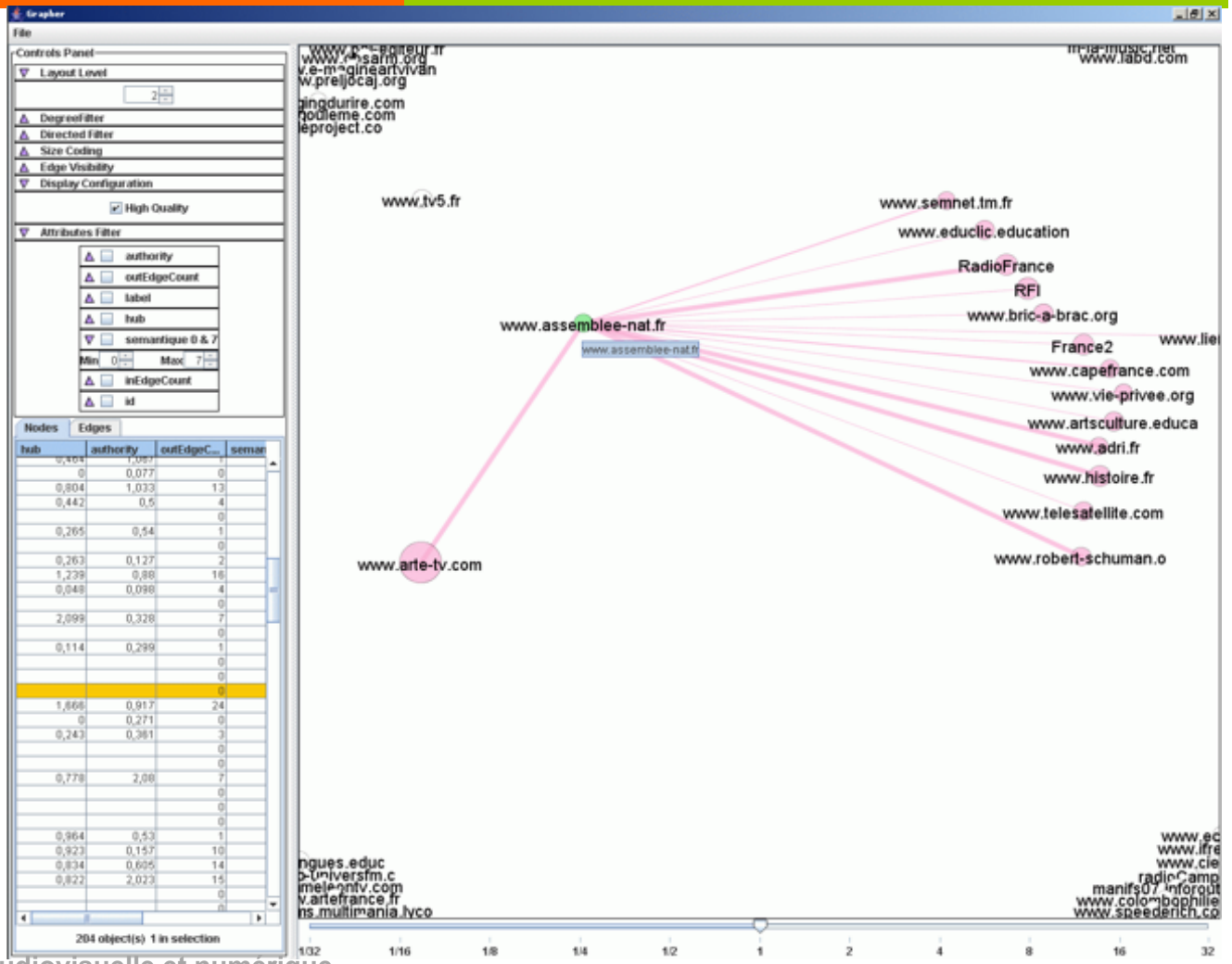
Connectivité des sites

Voisinage direct et positionnement des sites dans le domaine

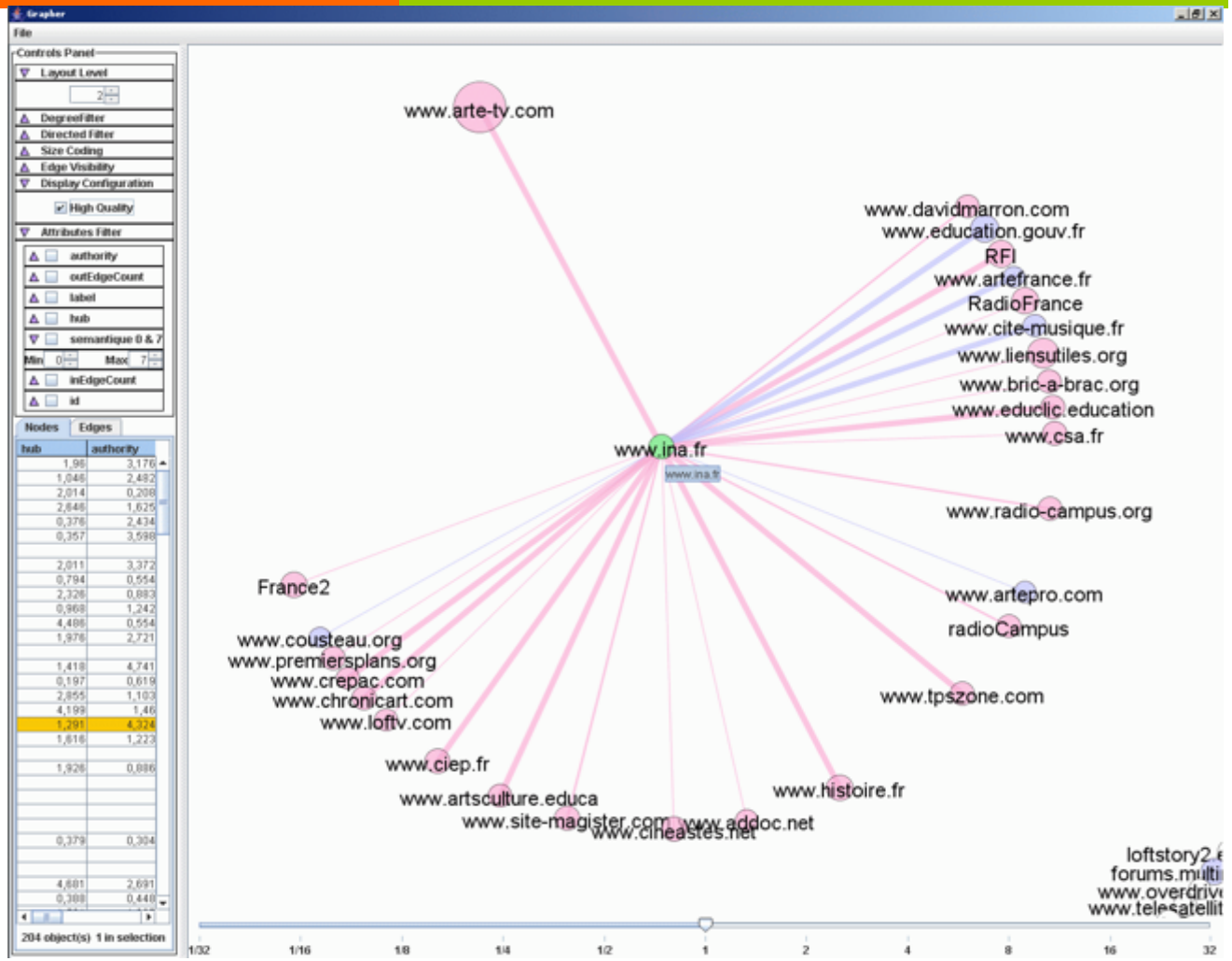
Connectivité des sites : eurotv.com



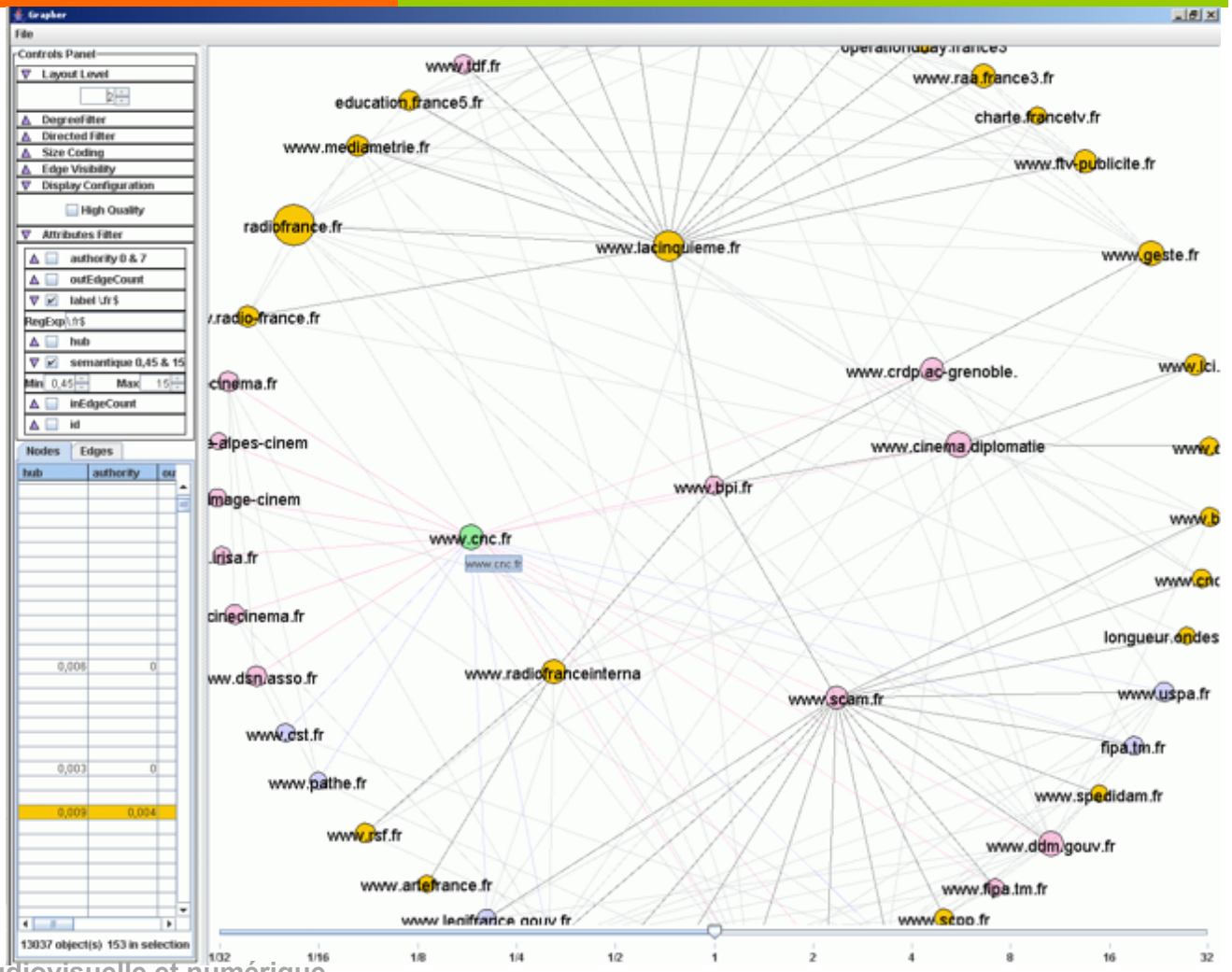
Connectivité des sites : assemblee-nat.fr



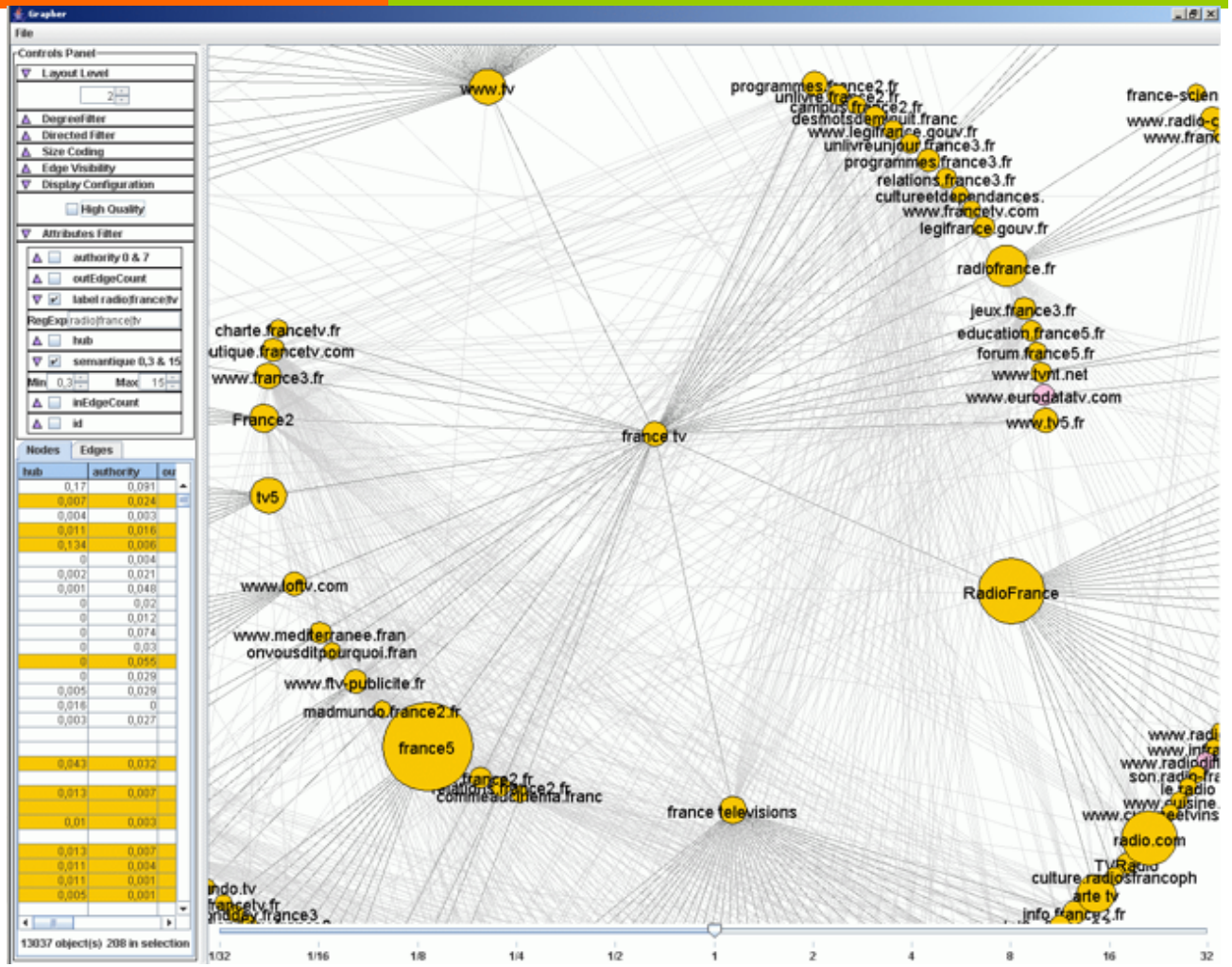
Connectivité des sites : ina.fr



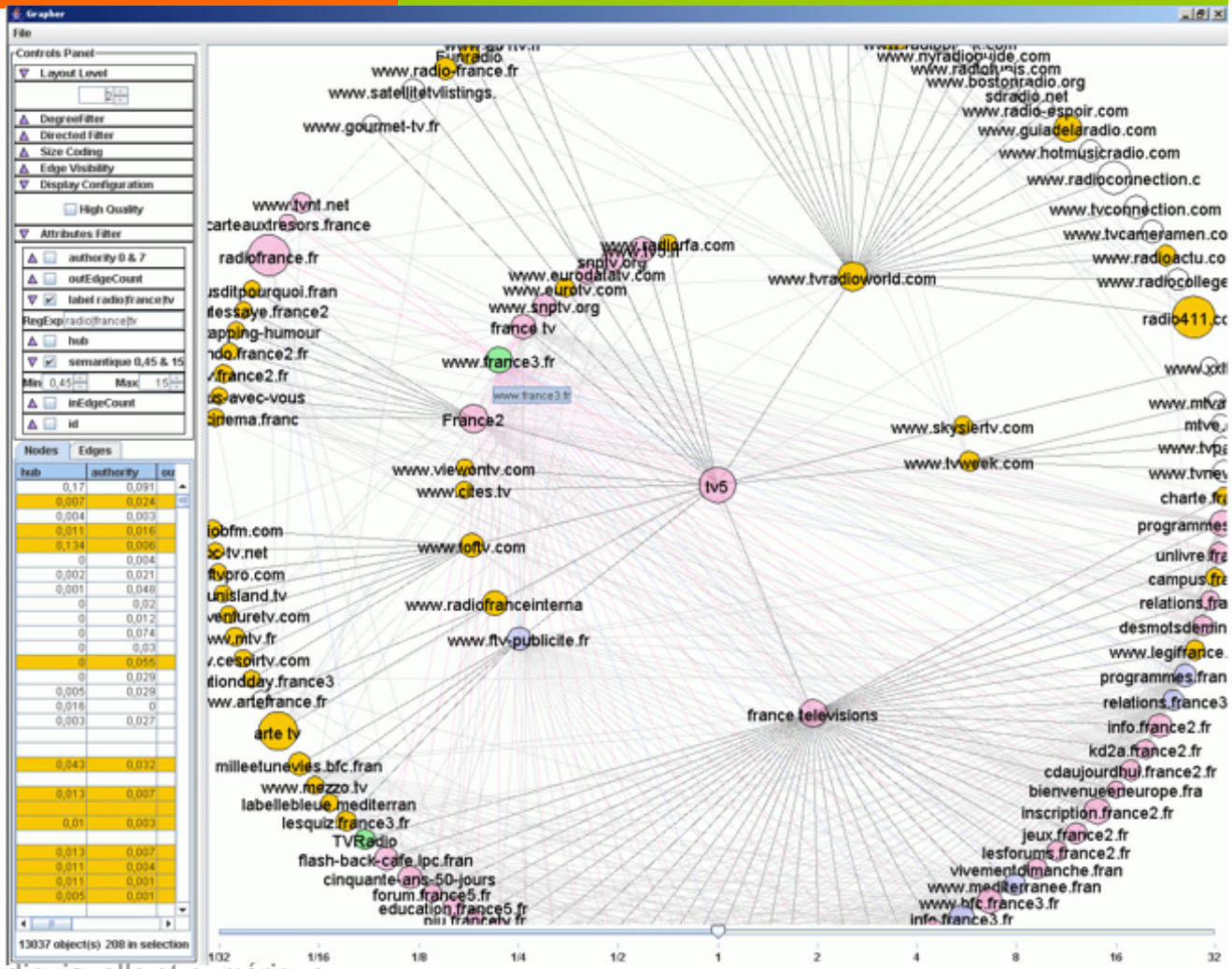
Connectivité des sites : bpi.fr



Connectivité des sites : francetv.fr



Connectivité des sites : tv5.fr



Principe

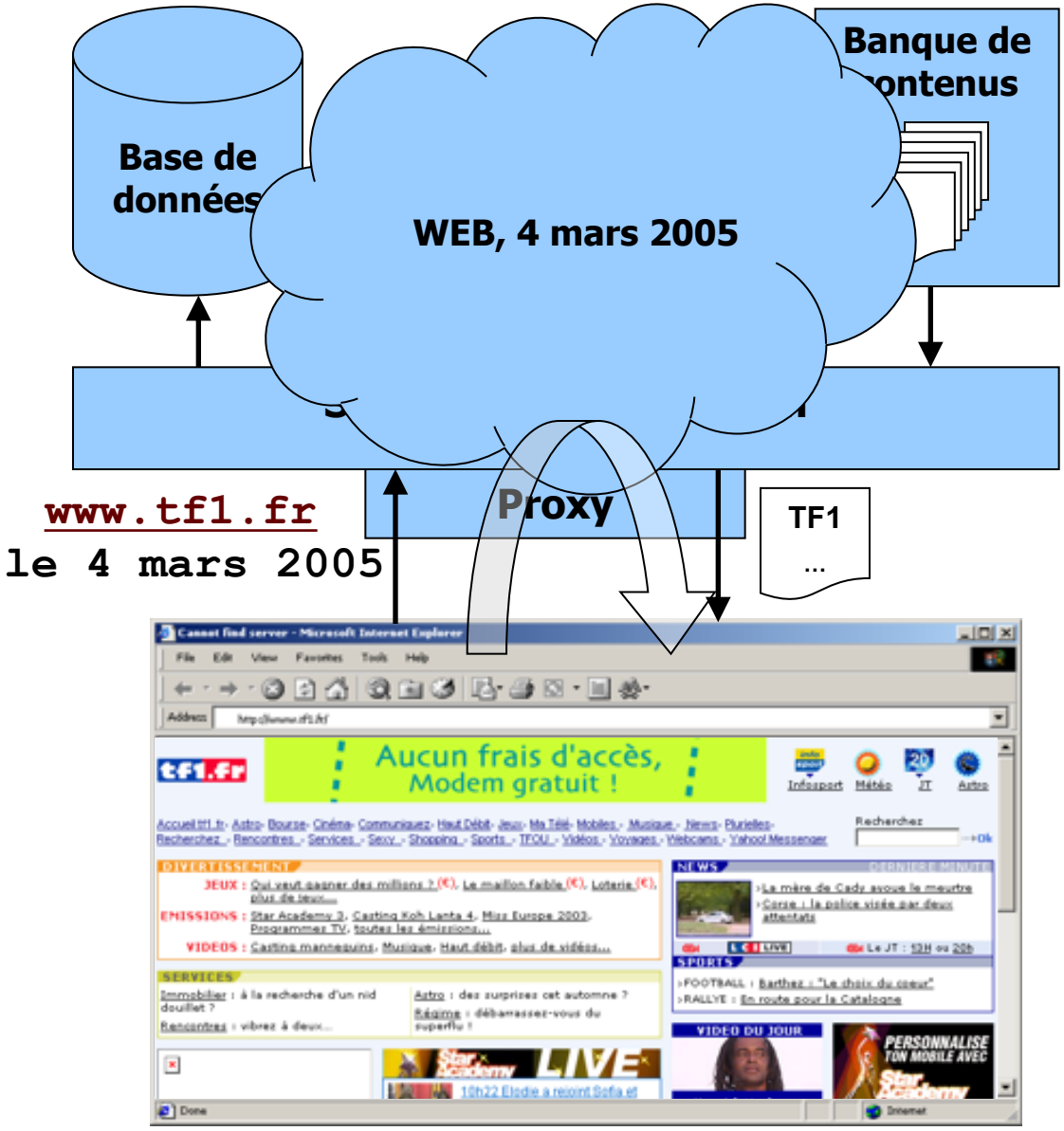
1. Définir, catégoriser et faire évoluer une liste de sites pertinents

2. Archiver ces sites à des intervalles de temps adaptés

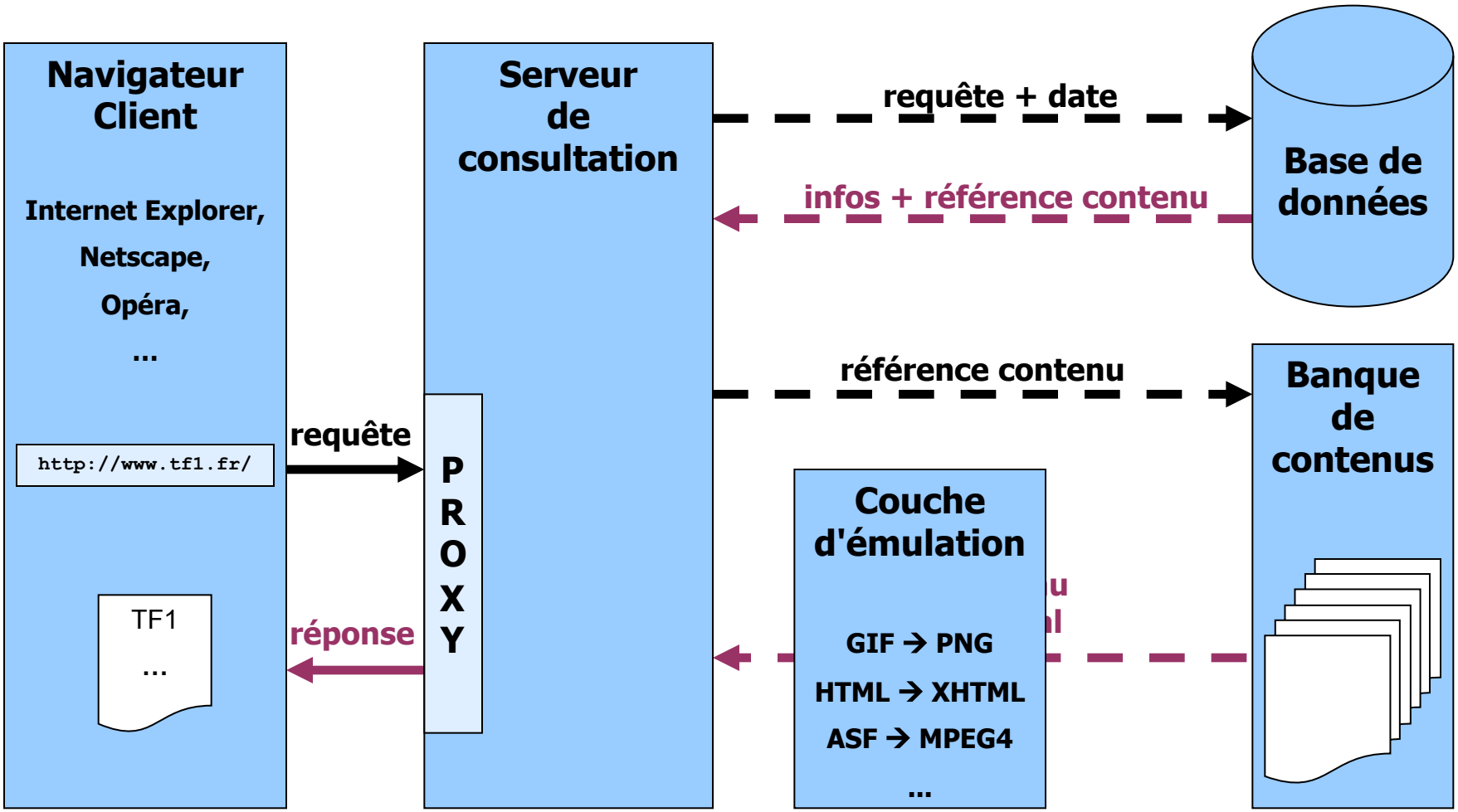
1. Proposer des enrichissements pour l'analyse de cette archive

2. Mettre en place une consultation de cette archive

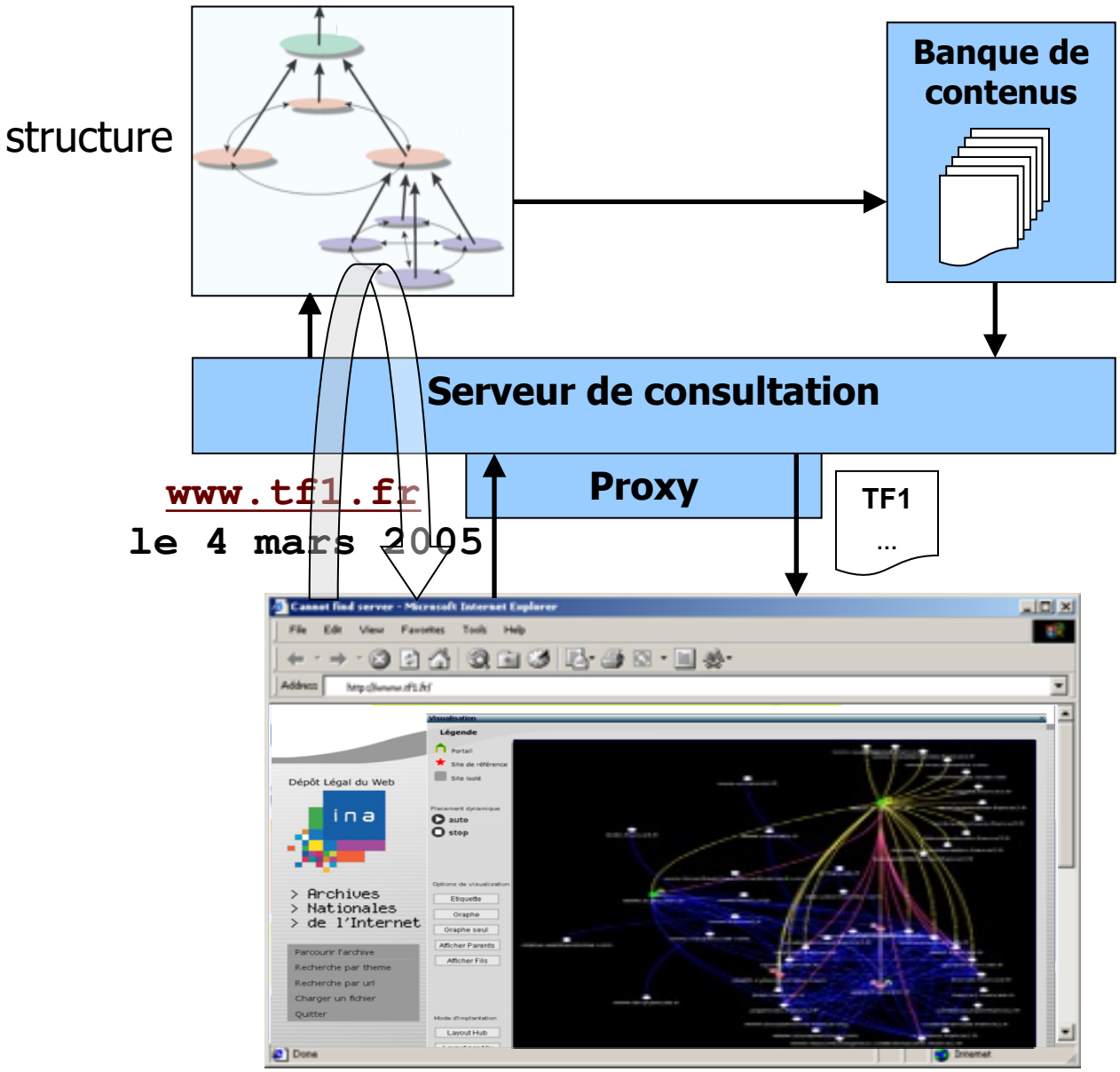
Consultation



Consultation

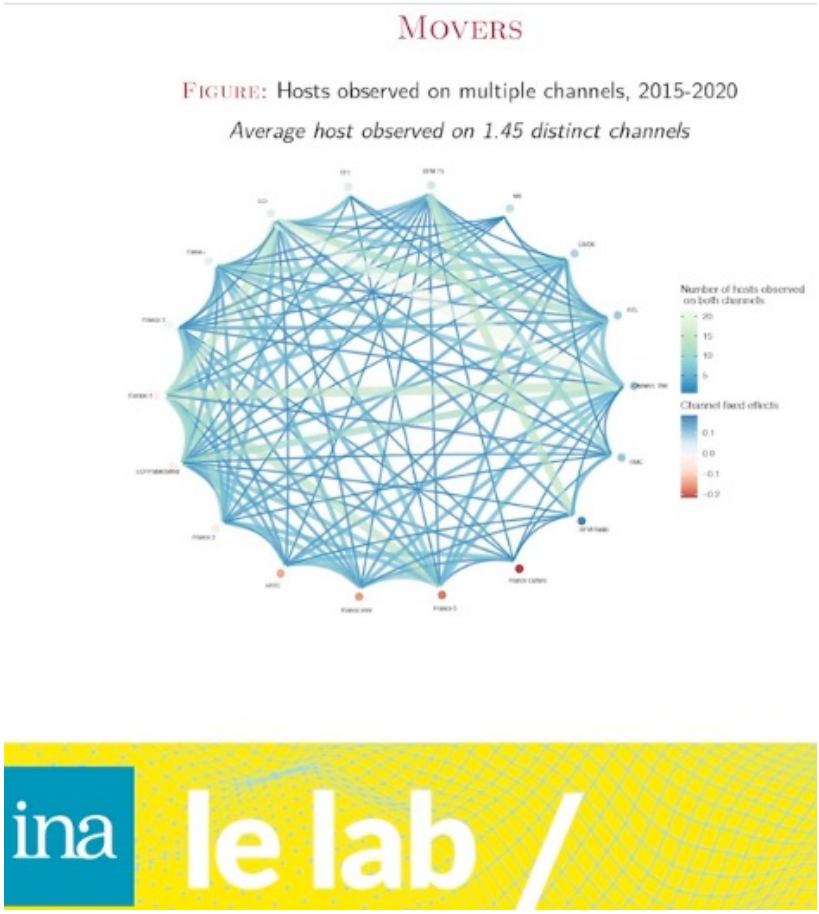


Consultation




Dépôt Légal du Web
Fichier Marques-pages Connection Affichage Préférences
Date de consultation : 2005-11-26 19:50:49
Versions Marques-pages Recherche Fichiers Historique
Versions disponibles
Version actuellement en ligne
Dernière version enregistrée
Versions antérieures
le 14-10-2005 à 20:45:24
le 24-11-2005 à 19:03:33
le 25-11-2005 à 16:54:13
le 26-11-2005 à 19:50:49
le 27-11-2005 à 12:24:24
le 28-11-2005 à 11:37:52
le 12-12-2005 à 02:56:52
France 5 - Chaine de télévision sur la découverte, l'emploi, la santé : documentaires, émissions, reportages, programmes
Précédent Suivant Recharger Stop http://www.france5.fr/ Go
france 5.fr
France 5 est une chaîne du groupe francetélévisions
France 2 France 3 France 4 RFO France Télévisions
Mercredi 14 décembre
ACCUEIL PROGRAMME LES SITES FRANCE 5 EMISSIONS THEMATIQUES LA CHAINE
SAMEDI 26
19:00 DOSSIER SCHEFFER
AVEC LE COMMISSAIRE DIVISIONNAIRE
FRÉDÉRIC PÉCHENARD
Pour en savoir plus »»
RECHERCHE
Dans France 5
Dans la grille des programmes
OK
AIDE !
TOUT FRANCE 5
Actu société
Arts et culture
Découverte nature
Emploi éco
Famille jeunesse
Histoire
Santé
Sciences
Vie pratique
PARTICIPEZ
Réactions, forums, contributions, témoignages,...
FRANCE 5 EDUCATION
Les nouveaux services Educatifs Multimédia de France 5
FRANCE 5 EMPLOI
BIEN VIVRE LE MONDE DU TRAVAIL
LES ZOUZOUS
LES RENDEZ-VOUS DU JOUR
09:10 "L'oeil et la main" : De la discrimination à la création d'entreprise.
09:40 "Cas d'école" : Ados et sexualité.
10:35 "L'atelier de la mode" : La tendance vintage.
11:10 "Question maison" : L'art nouveau, le viager...
12:00 "Silence, ça pousse !" : Le site de Siguyria, au Sri Lanka.
19:55 "Le journal du blogue".
20:05 "Ubik" : Marie Gillain, Tommy Lee Jones, Clarika...
» Bandes annonces | Tout le programme
09:40 - "Cas d'école"
La sexualité des ados. Reportages en ligne en vidéo.
En savoir plus
19:55 - "Le journal du blogue"
La parole est aux blogueurs ! Retrouvez la vidéo de l'émission en ligne.
En savoir plus
bandes annonces
sur France Télévisions
LES EMISSIONS
A vous de voir
Arrêt sur images
Avis de sorties
Brigade nature
C dans l'air
Le journal du blogue
Les amphis de France 5
Les maternelles
Les Zouzous
L'odyssée de l'espèce
NEWSLETTER
Je m'inscris ou modifie mes abonnements
Mon email ici
OK
»Voir la lettre d'infos de la semaine
Navigateur Plan de travail Editeur Texte Editeur HTML Cartographie
Done
none console

Exploiter les archives : vers la data Science



- **L'INA, qui accompagne déjà les chercheurs grâce à l'INA thèque, lance *le lab*, une nouvelle offre pour l'exploration et l'analyse des données issues des collections d'archives conservées par l'Institut. Des outils numériques et un ensemble de services pensés pour éclairer le temps long des pratiques et des discours médiatiques à la radio, à la télévision et sur le web.**

- L'INA, qui opère l'un des plus grands "data center" audiovisuels au monde et dont les collections s'enrichissent chaque année d'1,5 million d'heures de médias, veut stimuler la recherche à partir de ses données. C'est ce que permettra ***le lab***, un incubateur de projets de recherche centrés sur l'exploitation de données dans tous les champs des sciences sociales.

https://www.ina.fr/actualites-ina/l-ina-lance-le-lab-la-data-media-au-service-des-chercheurs

ina le lab /

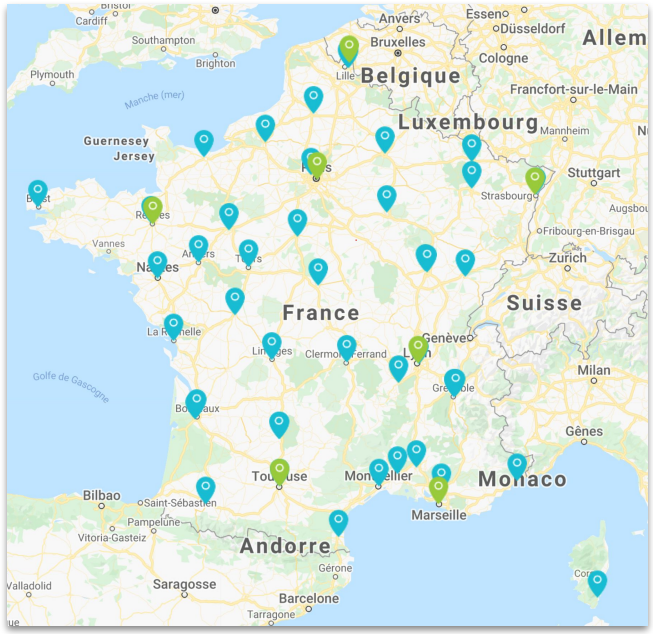
Consultation DL-Web

Consultation experte

7 lieux
Inathèque
Délégations régionales

Consultation autonome

50 PCMs
Bibliothèques partenaires

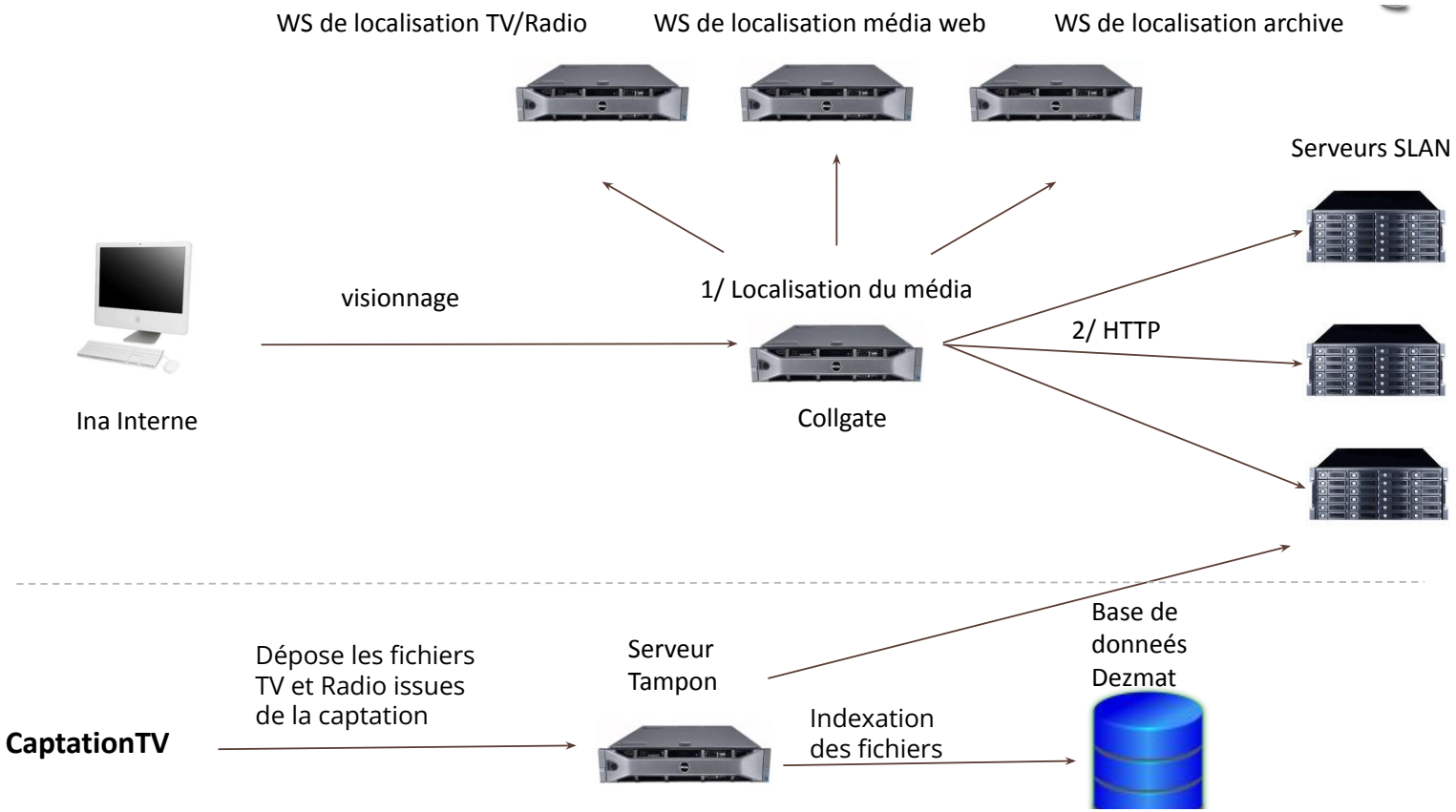


Conclusion sur l'approche de l'Ina

Un projet d'ingénierie d'ampleur

- Webmedia :
  - client-lourd basé sur electron
  - un navigateur web avec dimension temporelle
  - déployé dans les lieux de consultations et chez les personnels INA
- 1 dizaine d'applications web
  - moteurs de recherche
  - accès aux sources nominées
  - guide et tutoriels
  - statistiques
- Construite autour de librairies
  - WebComponent spécifiques
  - Vortex & Harbor : serveur applicatif et consultation

Une architecture complexe



Conclusion sur le statut de l'archive du Web

Trois objets en confrontation

- Les réalités socio-culturelles
- Le web tel qu'il était
- L'archive comme telle et telle qu'elle le représente
- Enjeu :
  - L'objectivité du Web ne se révèle que par l'objet « archive »
  - Celle du réel par celle du Web
  - Les objets ont leur régime d'objectivité propre
    - Les dégager pour eux-mêmes
    - montrer les décalages entre eux.

« l'archive »:

- Un construit qui correspond à un enregistrement partiel du donné
- Un formatage lié aux contraintes de stockage, de rejouabilité et d'accessibilité.
- L'archive qui n'est :
  - Ni une trace
  - Ni une donnée
  - Mais un enregistrement…

La trace

- Une proposition de caractérisation :
  - Une trace est un élément dont le statut causal est en rupture avec son environnement.
- Exemple :
  - Trace comme résidu : des substances qui restent, mais elles n'ont rien à faire là ; elles appartiennent à une histoire qui n'est pas celle du contenu qui contient ces traces ;
  - Trace comme empreinte ou indice : la trace a été produite, causée, par un actant qui n'est plus là, actant qui introduit une causalité qui n'est pas celle de l'environnement (branche cassée par le passage d'un gibier, trace de pas, etc.).

La donnée

- Une proposition de caractérisation :
  - Une donnée est une proposition affirmant un fait, dont l'expression est formalisée et l'interprétation normalisée.
- Exemple :
  - L'enregistrement élémentaire d'une base de données (!)
  - Un triplet RDF
  - Une cellule d'un tableau Excel
  - Etc.

L'enregistrement

Ni une trace (car déjà une donnée)

- L'enregistrement n'est pas un reste, mais une transformation qui engendre un nouvel objet selon un format propre :
  - Signal électrique à partir de vibrations sonores ou lumineuses
  - Codes numériques à partir de cellules (natif)ou signaux (numérisation)
- Même produit de manière mécanique et automatique, l'enregistrement est un arbitraire imposé à une origine.

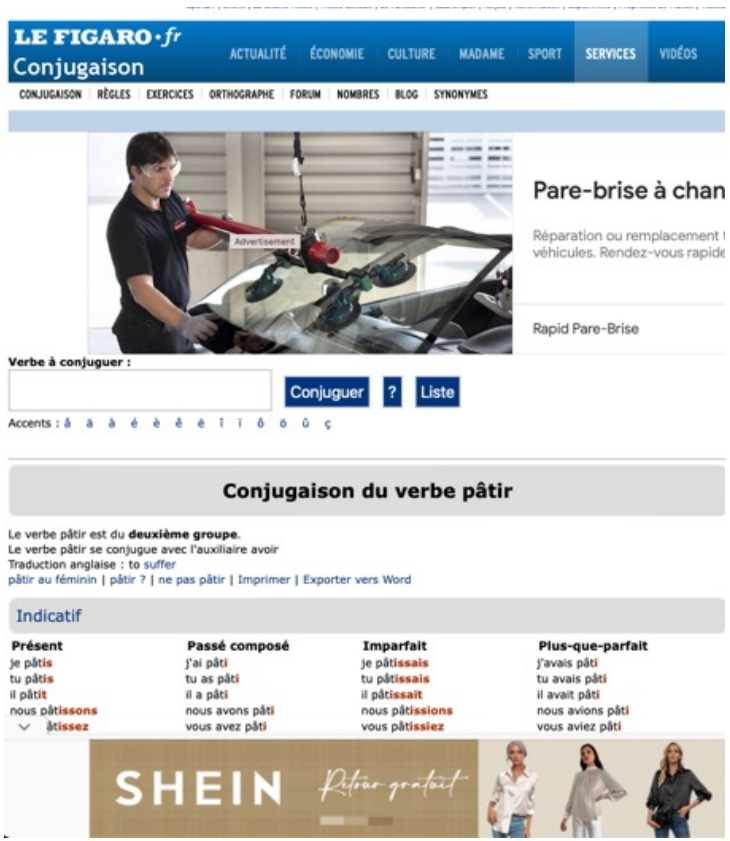
Ni une donnée (car encore une trace)

- L'enregistrement n'est pas une pure convention ni un pur décret :
  - On se le donne à partir d'un donné, d'un fait.
- Même définie en fonction de la manipulabilité attendue et de la résolution espérée, la donnée issue d'un enregistrement emporte avec elle un résidu, une trace de son origine.

Le web ré-inventé

- Dégager l'objectivité propre de l'archive constituée pour appréhender les autres ;
- Cette dernière négocie son objectivité selon différents régimes :
  - Ressemblance (analogie, archétype et objectivité comme vérité d'après nature)
    - Elle présente une apparence : ça ressemble au Web !
  - Conséquence physique (causalité physique, mesure et objectivité mécanique)
    - Elle est obtenue par enregistrement et captation
  - Conséquence formelle (codage, déduction et objectivité du jugement exercé)
    - On la fait parler et on l'analyse comme la base d'information pour retrouver le web, mais aussi ce dont il est la manifestation.
- Ces régimes révèlent par leurs discontinuités l'objectivité du Web passé et des réalités manifestées par ce dernier :
  - Rupture sémiotique dans la ressemblance
  - Rupture technique dans la conséquence
  - Rupture déductive dans le codage.

Rupture sémiotique

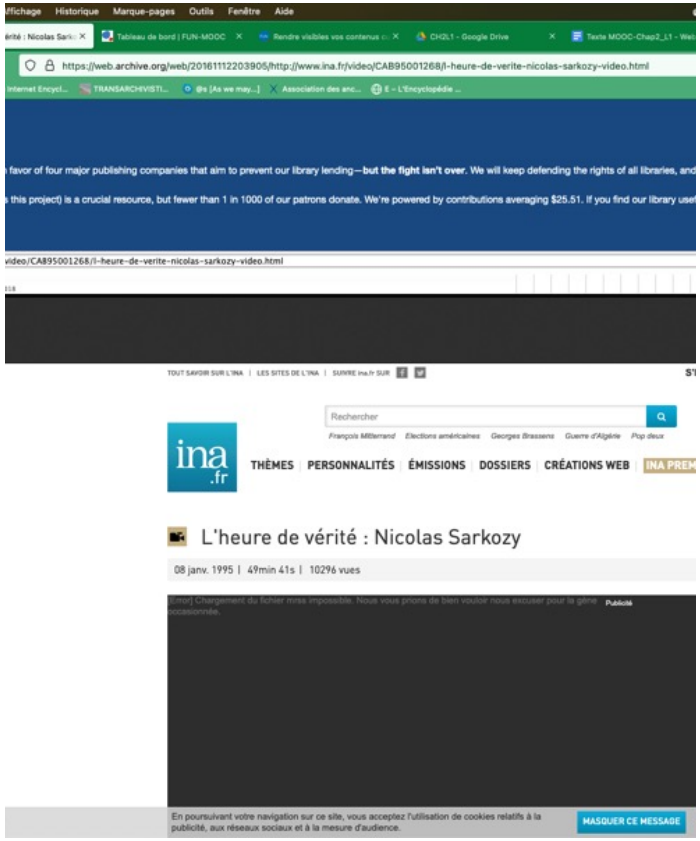


- On reconnaît des styles graphiques différents, dénotant une rupture ou hétérogénéité dans la ressemblance :
  - ce qui ne se ressemble pas n'appartient pas à la même objectivité
- Hétérogénéité du web avec lui-même
- Surmonter le présentisme de l'apparence et de la ressemblance

Rupture sémiotique, bis repetita

- Le style sémiotique dénote vis-à-vis des styles actuels :
  - Paradigme de la ressemblance qui permet de dégager des catégories graphiques temporelles.
  - Ces catégories s'assimilent des vérités d'après nature, le type de site qu'on reconnaît dans leur temporalité malgré tous les efforts qu'ils faisaient pour se démarquer les uns des autres.
- Hétérogénéité entre les époques du Web

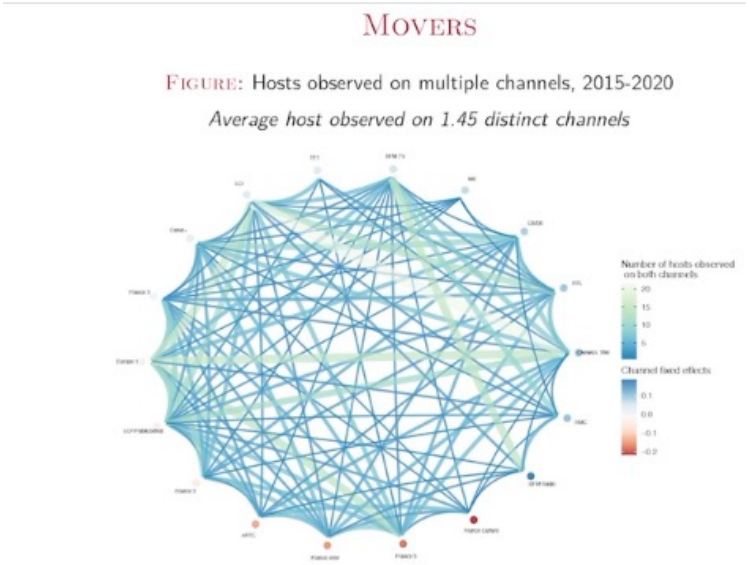
Rupture technique



- Rupture dans la chaîne technique des formats de décodage : le player ne sait pas rejouer les formats d'origine
- Ce qui n'est pas rejoué n'appartient pas à la même objectivité
- Exemple :
  - ina du 4 février 2016 sur la Wayback Machine
- Hétérogénéité de l'archive et du Web
- Surmonter le présentisme du numérique

Rupture déductive

- Analyse et interprétation de l'archive comme base de données
- Le jugement exercé s'exprime sous la forme d'un traitement calculatoire et déductif mais repris sous l'angle du narratif final.
- Hétérogénéité du monde et du web
- Surmonter le présentisme de la donnée



ina le lab /

Exploiter ces ruptures pour montrer la rupture du temps passé…

- Apparat critique à construire :
  - Contexte socio-historico-culturel : le Web n'est pas une réalité isolée,
    - quoi / pourquoi ?
    - Empathie historique, pas d'anachronisme psychologique
  - Contexte socio-technique :
    - qui archive, comment ?
    - Biais techniques, historicité technique
  - Contexte algorithmique : traitement analytique des données,
    - Pour qui, Pour quoi ?
    - Historicité des données à ré-établir ;
- Éviter que les études du Web ne basculent dans un présentisme des problèmes et des données

Noème du Web

- Ni un
  - Ça a été (Barthes et la photographie analogique)
  - Ça a été manipulé (bb et le numérique)
- En effet :
  - C'est calculé et reconstruit dynamiquement à l'instar du numérique
  - C'est reproduit à partir de traces muées en données pour opérer la reconstruction : on l'a ré-inventé (au sens des archéologues).
- Donc, quelque chose comme :
  - Ça a été ré-inventé (trouvé et reconstruit).

Conclusion : historicité du Web ?

- 3 niveaux de réalité :
  - Une réalité dont le Web serait une manifestation d'une manière ou d'une autre : c'est pour cela qu'on l'interroge ;
  - Le web tel qu'il existait un moment donné : trace ou émanation de la réalité précédente ;
  - L'archive qu'on en a tiré : trace ou émanation du web tel qu'on l'a capté.
- L'archive est une trace de trace qu'on utilise comme donnée pour reconstituer / appréhender la réalité considérée.
  - potentiellement une lessiveuse temporelle, plongeant le passé dans un présentisme des données et du numérique ;
  - Enjeu de mobiliser les ruptures (sémiotique, technique, logique) pour rétablir l'historicité du Web.