Archives Audiovisuelles & Data

EGIDE - 14 Mars 2018

01

De la rareté à l’abondance

Hier :

Moyens de production rares et onéreux (35mm, caméra)

Moyens de diffusion rares et centralisés (chaine unique, cinémas)

Aujourd'hui :

Moyens de production facile et peu onéreux (smartphones)

Moyens de diffusion nombreux et décentralisé (internet, chaines ,)

Une Abondance qui pose des enjeux de :

Traçabilité, provenance

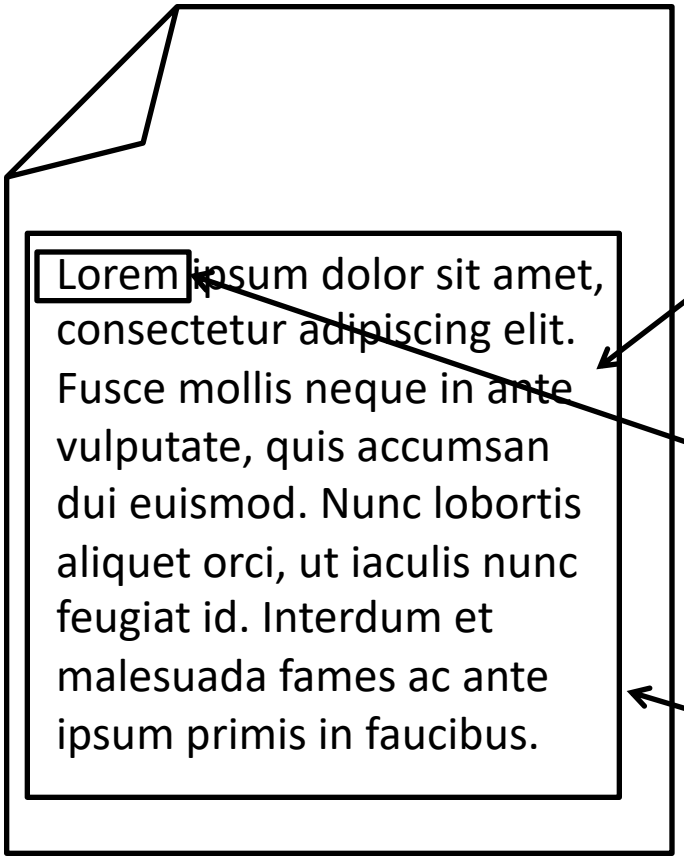
Mise en contexte, vérification de l'information

Importance de la donnée d'accompagnement :

02

Du document à la donnée

Document vs donnée



Un **document** est un ensemble logique, fini d'**informations** dont les limites peuvent être définies par des caractéristiques physiques.

Le fait que « Le premier mot du texte soit Lorem » est une donnée.

Le fait que « ce texte soit extrait d'un texte de Cicéron » est une donnée, qu'on appelle métadonnée (donnée sur la donnée)

Hier :

Prima du document :

On observe :

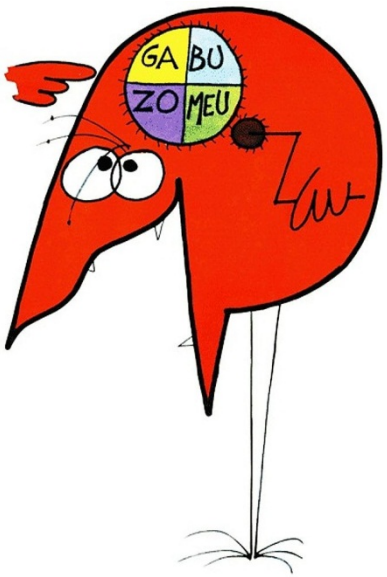
→ Cloisonnement des collections

→ Des systèmes documentaires tournés vers peu d'usages

→ Opacité des collections

Aujourd'hui :

Prima de la donnée :



Objectifs :

→ Décloisonner les collections

→ Libérer de nouveaux usages

→ Transparence des données

→ Résilience des systèmes documentaires

Exemple : Numériser une collection audiovisuelle

Un processus au coût maîtrisé, industrialisé

une équation de la problématique

l'importance aussi de la donnée documentaire

Une collection sans matériel d'accompagnement, sans données :

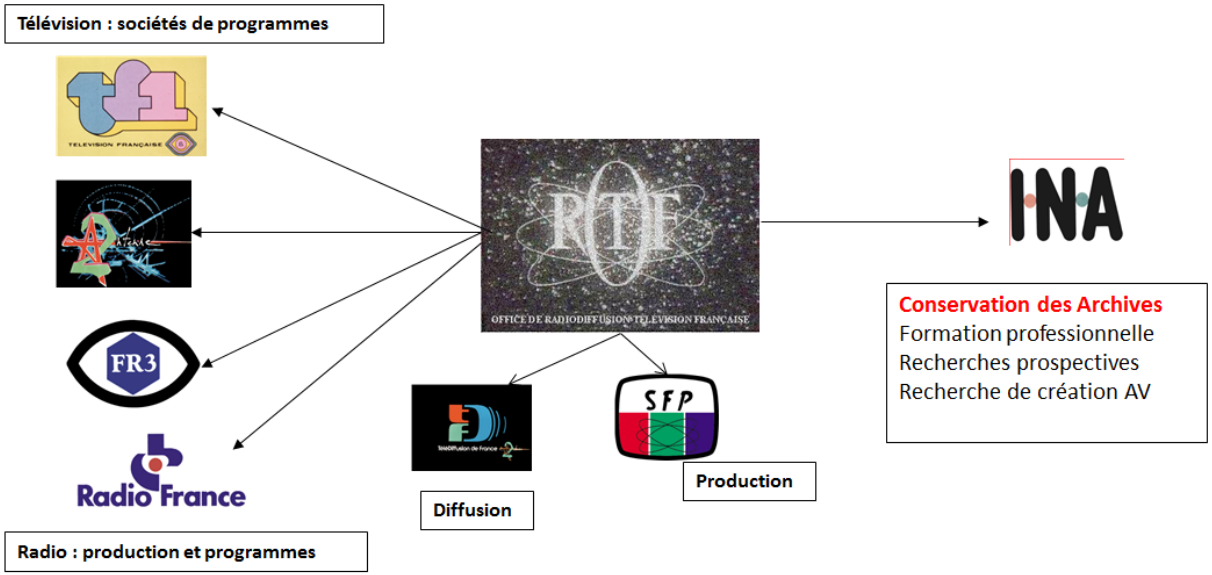
→ moins de valeurs

Car le temps de description documentaire nécessite d'énormes ressources.

Historiquement les collections de l'Ina ce sont deux fonds principaux :

1 / LE FONDS DES ARCHIVES PROFESSIONNELLES

L'éclatement de l'ORTF en 1974 : Archives de l'ORTF + TF1 + Antenne 2 + FR3 + Radio France : Le début des archives Patrimoniales



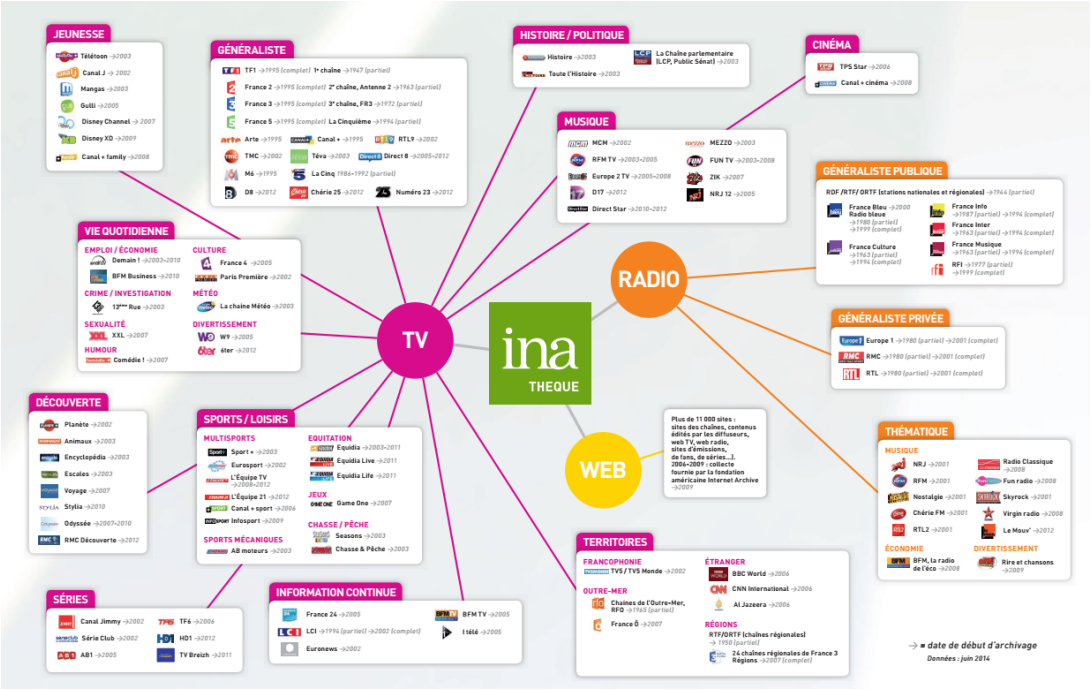
→ 2 000 000 d'heures

→ usage : une vente des archives pour réutilisation dans de nouveau programmes

2 / LE FONDS DU DEPOT LEGAL

Création du dépôt légal de l'audiovisuel français en 1992 :

100 chaines tv, 30 chaines radio et 11 000 sites web média **captés 24h/24**



→ 14 174 000 heures au titre du Dépôt légal

→ usage : consultation par les chercheurs

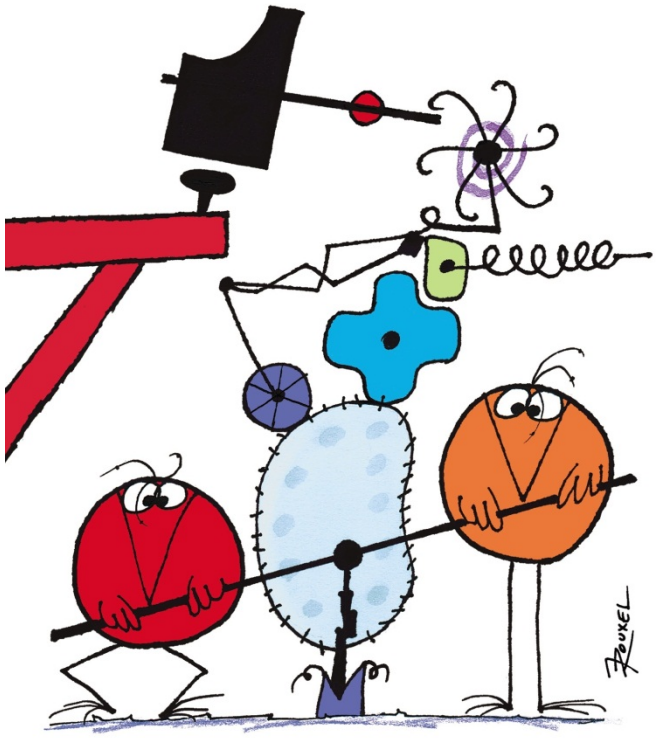
Deux Fonds distincts

2 logiques différentes

2 logiciels différents

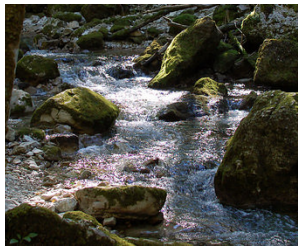
→ peu d'interopérabilité

→ outil documentaire tourné vers le matériel
et adapté à un usage précis

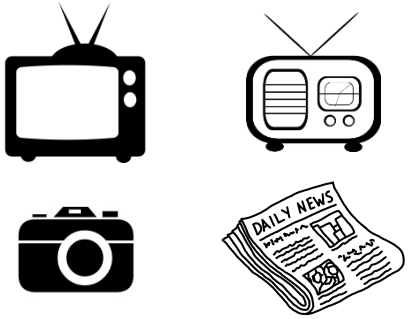


(recherche ou vente)

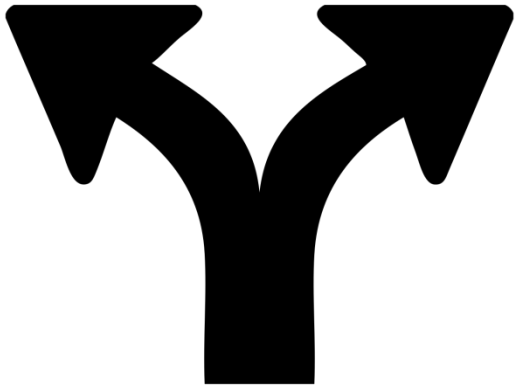
Les problèmes posés



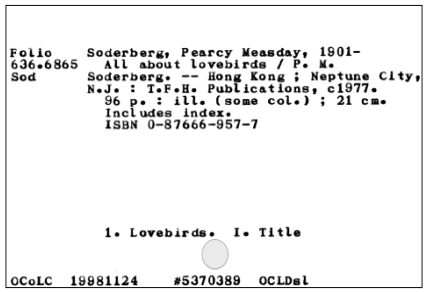
Stock vs flux



Différents médias



Evolution en parallèle



Folio Soderberg, Pearcy Measday, 1901-
636.6865 All about lovebirds / P. M.
Sod Soderberg. -- Hong Kong ; Neptune City,
N.J. : T.F.H. Publications, c1977.
96 p. : ill. (some col.) ; 21 cm.
Includes index.
ISBN 0-87666-957-7

1. Lovebirds. I. Title

OCoLC 19981124 #5370389 OCLDsl

Description rigide

Projet en cours à l'Ina : Le lac de données

Des données + des référentiels
+ Une interface de recherche suivant
le profil d'utilisateur

Objectifs :

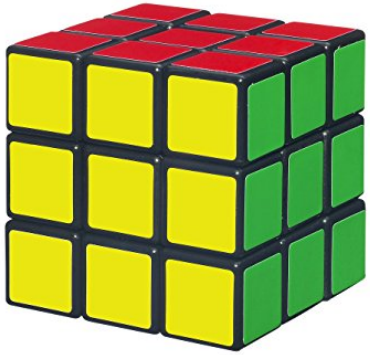
→ Réunir les données des différentes bases

→ Créer des liens entre collections

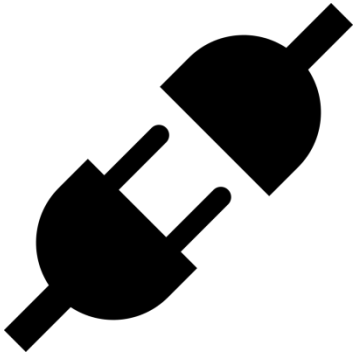
→ Favoriser de nouveaux usages

Les points importants

Souplesse



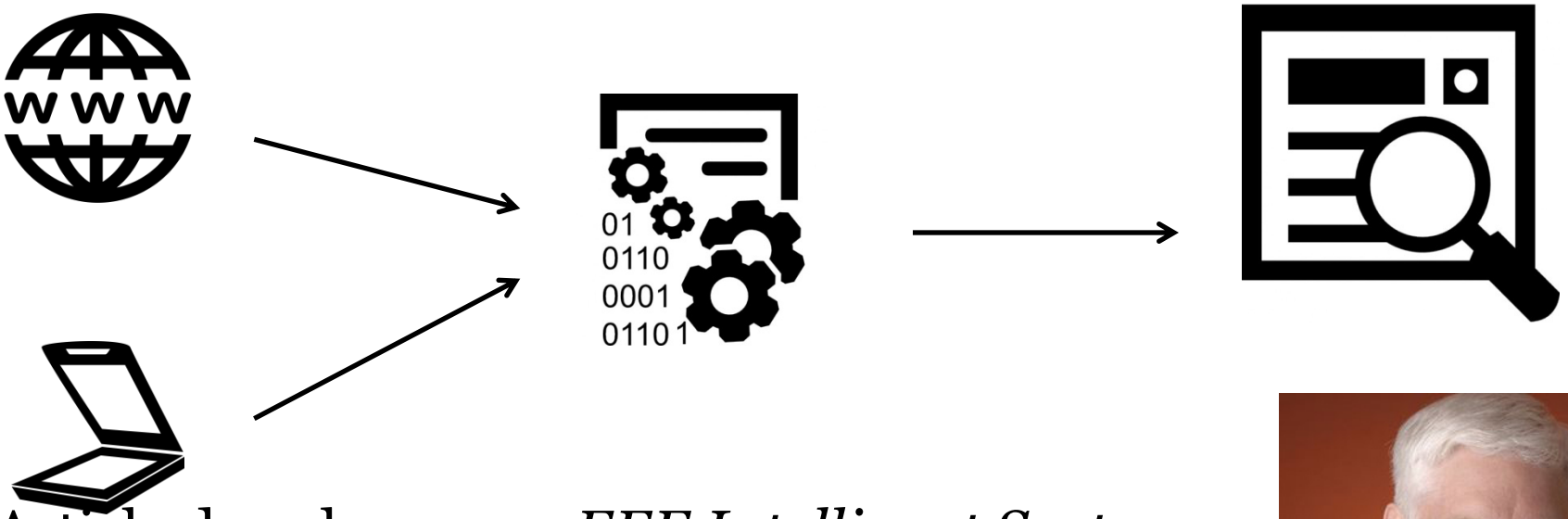
Cohérence



Interopérabilité

REMISE EN CAUSE DU TRAVAIL DE CATALOGAGE ET D'INDEXATION ?

L'indexation automatique du contenu ou la remise en cause du catalogage



Article dans la revue : *EEE Intelligent System journal*

mars 2009

"the unreasonable effectiveness of data"

Peter Norvig

Responsable de la recherche chez Google

La folksonomie
ou la remise en cause du référentiel

Article sur le blog : *Clay Shirky's Writings About the Internet*

Printemps 2005

"Ontology is Overrated: Categories, Links, and Tags"

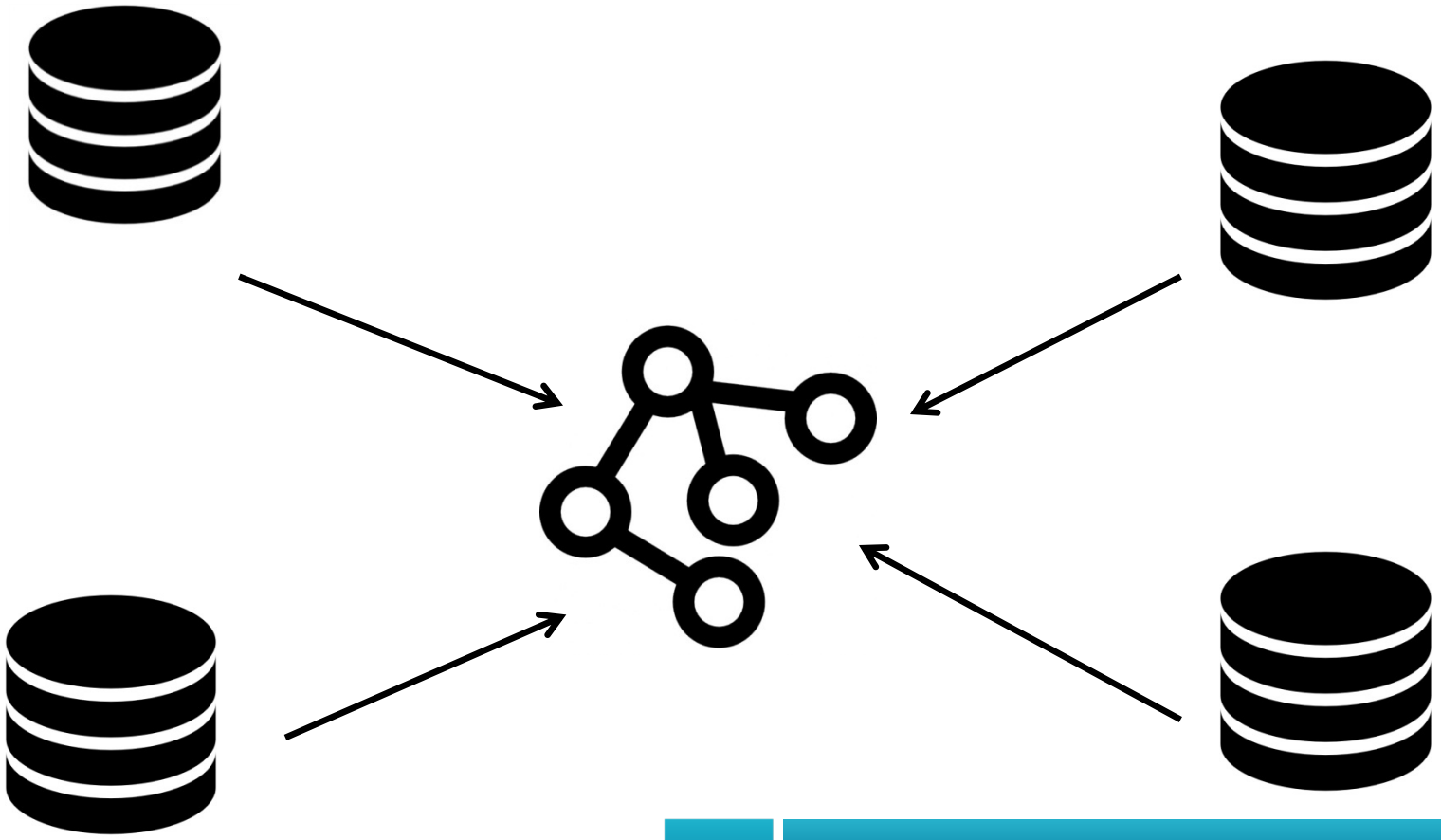
Clay Shirky

Journaliste spécialisée dans le numérique

ina.fr

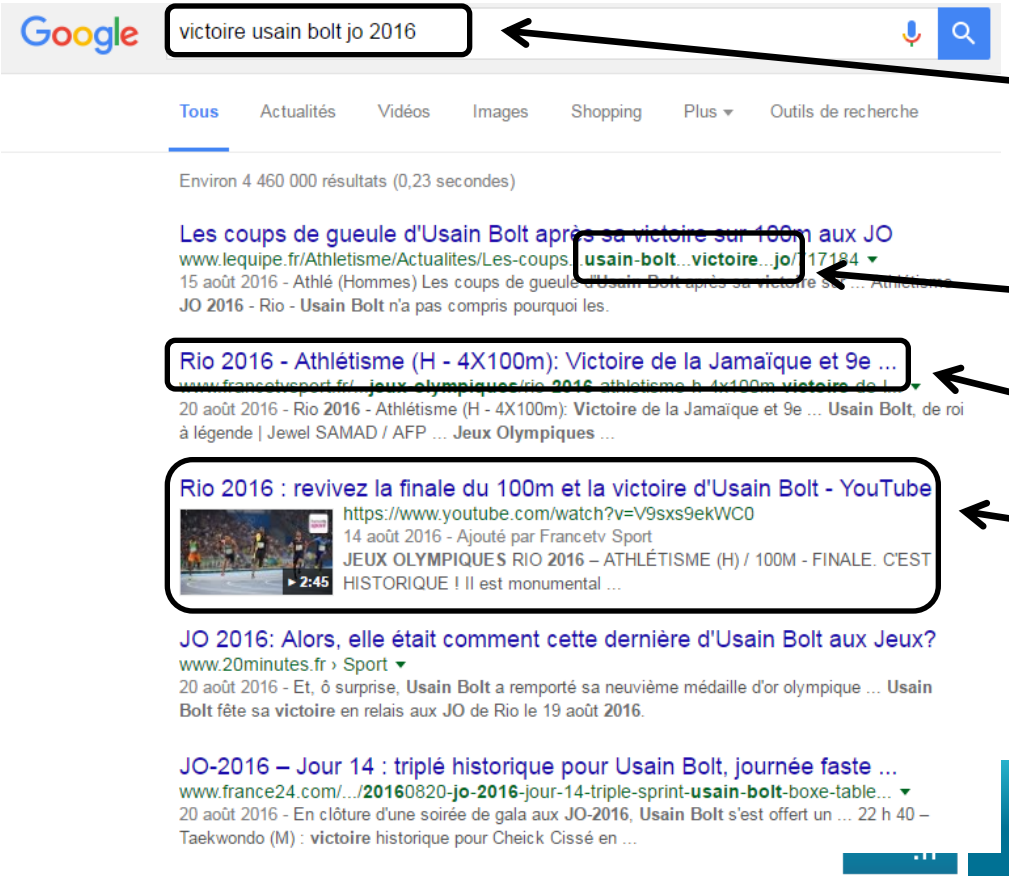
Mettre en cohérence les données

ou la nécessité de disposer de référentiels structurés



Hiérarchiser les données

ou la nécessité de disposer de métadonnées pertinentes



Ma recherche

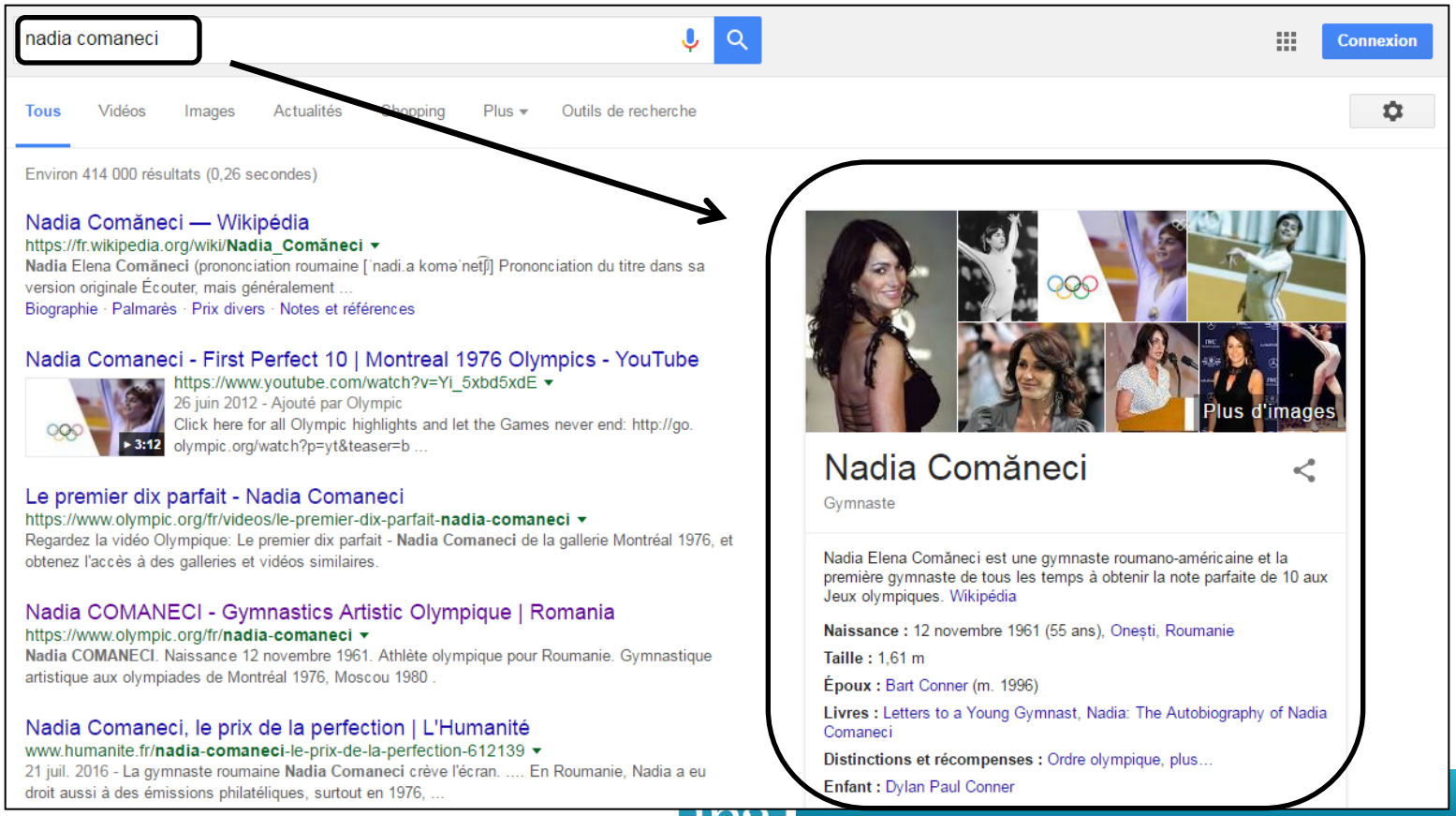
Les mots dans l'URL

Les mots dans le titre

Le format et la provenance du document

Organiser les données

ou la nécessité de disposer de données structurées

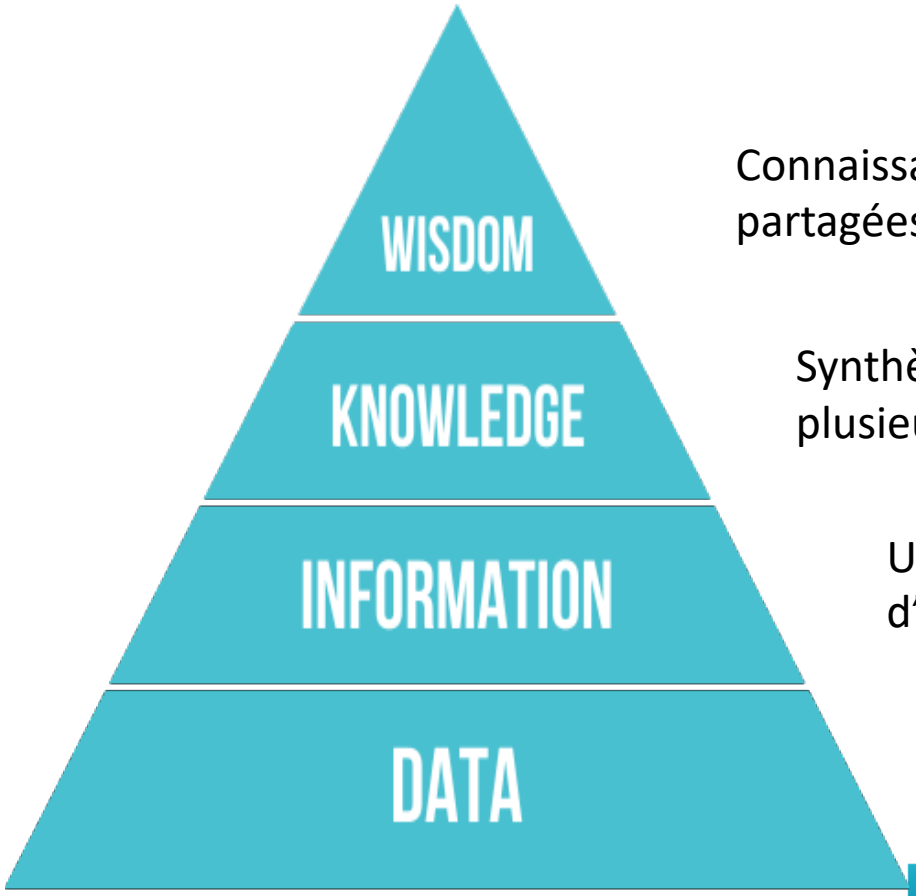


Le *knowledge graph* de Google

03

Typologie des données

Qu’est-ce-qu’une donnée ?



DIKW Pyramid

Connaissances assimilées et/ou connaissances partagées

Synthèse ou déduction obtenue à partir de plusieurs informations

Un ensemble organisé de données OU résultat d’un traitement sur un ensemble de données

Des faits, signaux et symboles formant une unité indépendante les unes des autres et non traitée

Typologie des données

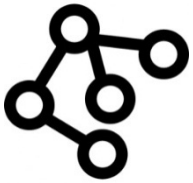
Contenu

- Nativement numérique
- Issu d'un processus de numérisation (OCR, Speech to text, reconnaissance de formes)

Métadonnées

Informations sur le document qui peut-être d'ordre :

- descriptive
- administrative (technique, traçabilité, juridique)
- structure



Référentiels

Informations de référence permettant d'assurer une cohérence entre les données

Données d'usage et contributions

Toutes les données laissées par les utilisateurs : logs, traces, contributions sur les réseaux sociaux en rapport avec les données

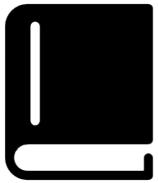
Données générées /secondaires

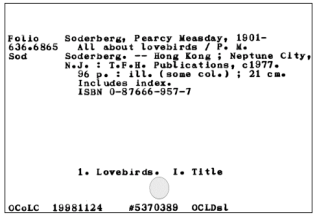
Données obtenues à partir de l'exploitation des autres types de données : données quantitatives, extraction, *machine learning*, inférence…

De la notice à la donnée

Anatomie d'un changement de granularité

Seule la description est numérique




Folio Soderberg, Pearcy Measday, 1901-
636.6865 All about lovebirds / P. M.
Sod Soderberg. -- Hong Kong ; Neptune City,
N.J. : T.F.H. Publications, c1977.
96 p. : ill. (some col.) ; 21 cm.
Includes index.
ISBN 0-87666-957-7

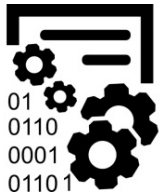
1. Lovebirds. I. Title

OCoLC 19981124 #5370389 OCLDal

Notice

La description d'un document forme lui-même un document, un tout cohérent et fini

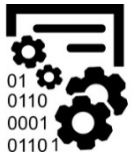
Contenu numérisé



Métadonnée

Chaque information est indépendante
Elargissement de la description (technique, juridique, structure…)

Contenu nativement numérique ou numérisé



Donnée

Fin de la hiérarchie entre les données (description, contenu, référentiels)
Diversification des données (données d'usages / contributions)

ina.fr

L'importance des données d'usage

Intérêt dans la recherche de documents :

Le nombre de vue -> une information importante pour le consommateur de vidéo

http

Les crêpes (avec beaucoup d'alcool) de Raymond Oliver | Archive INA

147 043 vues

857 47 PARTAGER

Ina Les Recettes Vintage
Ajoutée le 14 avr. 2017

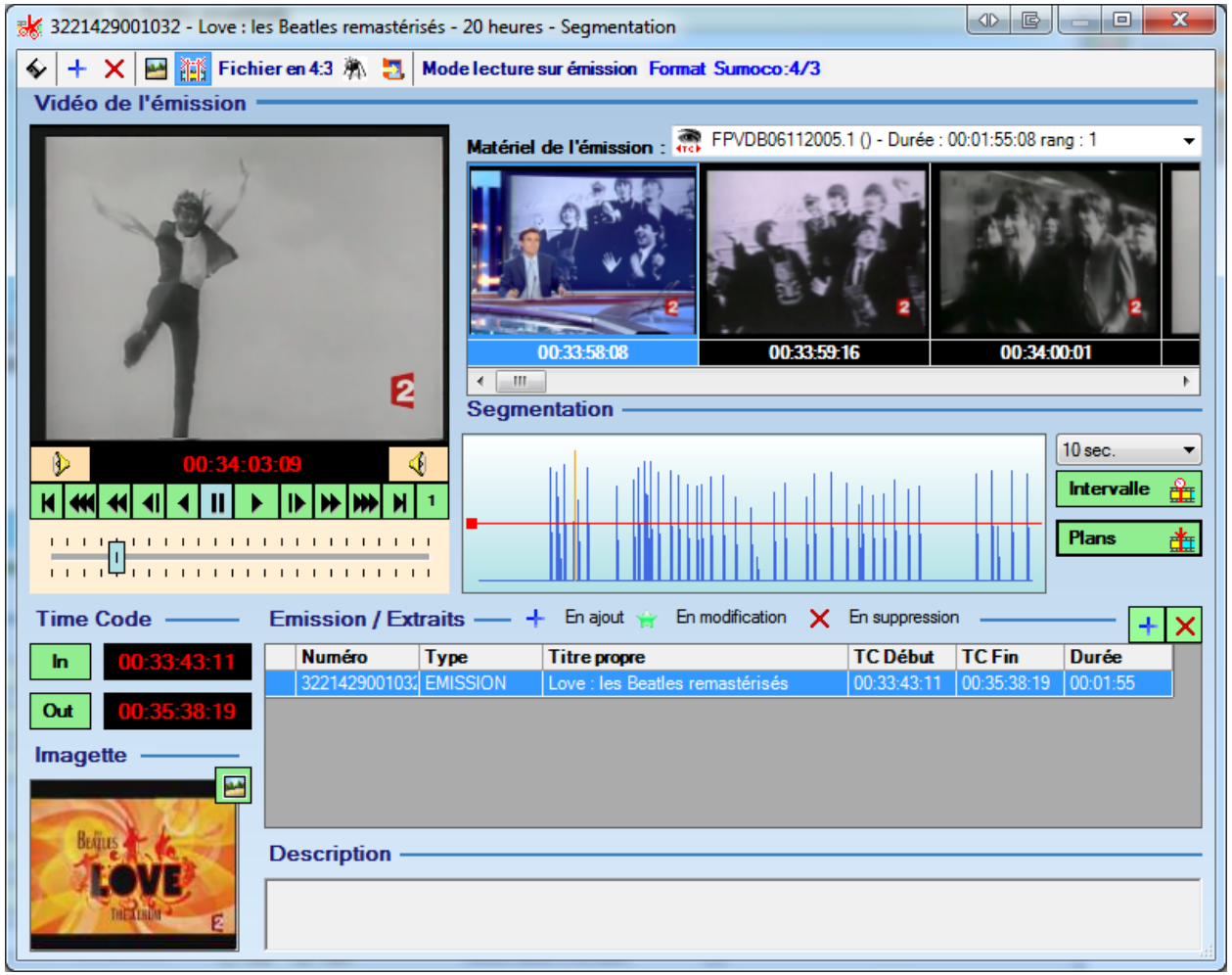
S'ABONNER 7,7 K

Abonnez-vous http://bit.ly/InaLesRecettesVintage
Art et magie de la cuisine
Raymond OLIVER et Catherine LANGEAIS préparent des crêpes, une recette demandée par les

Des modèles économiques basés sur l'utilisation de ces données d'usage :

Algorithme + données usages = Bizness juteux

Les données de contenu (signifiant / signifié)



Les données d’accompagnement

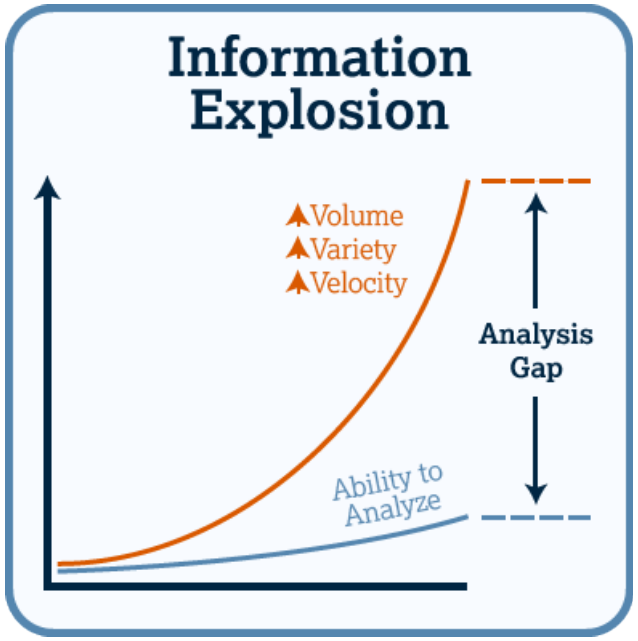
Les métadonnées

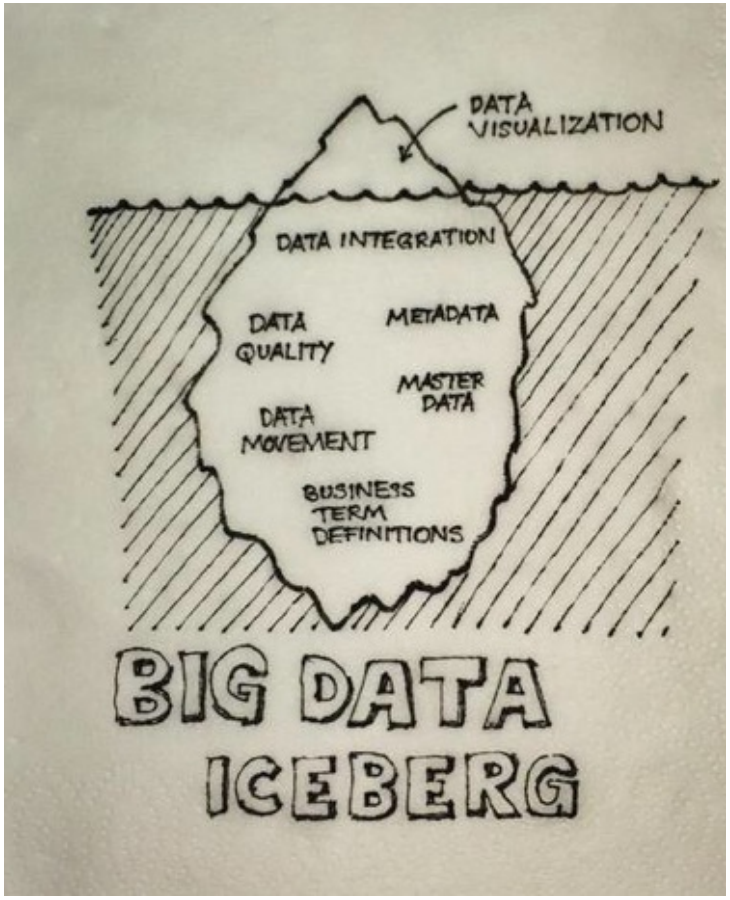
Les référentiels (thésaurus, listes contrôlées,…)

Les données générées automatiquement

Notion de Big Data

Explosion quantitative de la donnée numérique





Essai de définition

Le Big data désigne la capacité technologique à traiter de très grandes masses de données avec des infrastructures matérielles standards.

Par extension, on désigne avec cet expression un mouvement général se caractérisant par un intérêt particulier du monde numérique pour la donnée.

05

Le Département Recherche & Innovation numérique de l'Ina

Objectif du département Recherche et Innovation de l'Ina

→ Nouveau outils pour les usages existants

Fournir des outils

Assister, préparer le travail de documentation, d'annotation, d'indexation, de description

Proposer des annotations, des index, des repères

→ Proposer de nouveaux usages / services

Recherche par similarité

Réalité augmenté

Le Département Recherche et Innovation de l'Ina

→ 5 étapes :

S'ouvrir

Imaginer

Structurer

Créer

Inspirer

→ 3 règles

Dot it fast (3 à 6 mois)

Share it more (connectés en interne et en externe)

Try it now (généraliser ou jeter)

Objectif des technologies 1/2

→ Détecter

Image : présence de visages, de textes, de logos, d'objets divers

Audio : présence de parole, de musique, d'applaudissements, de sons divers

→ Reconnaitre

Image : le visage de qui ? Qu'est il décrit ? Quel tableau ?

Audio : qui parle ? Qu'est-il dit? Quelle chanson ?

Image + audio : qui?

Objectif des technologies 2/2

→ Organiser

Image : regrouper les visages

Audio : musiques par genre

→ Structurer temporellement

Découpage des Jt en « Plateaux » et « Reportages »

Chapitrage

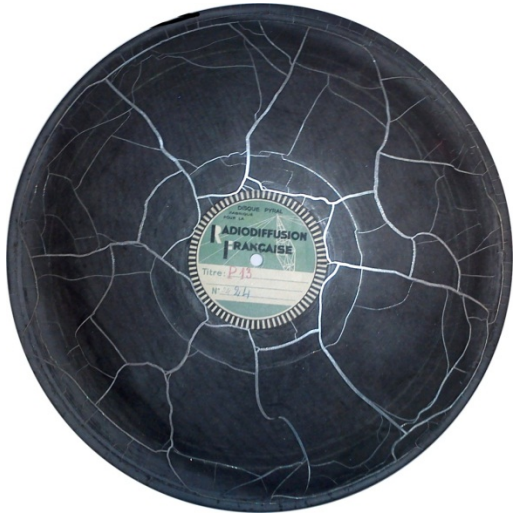
Détection des sections d'un morceau de musique (couplet –refrain)

Démonstrations sur l'Espace Recherche
http://recherche.ina.fr

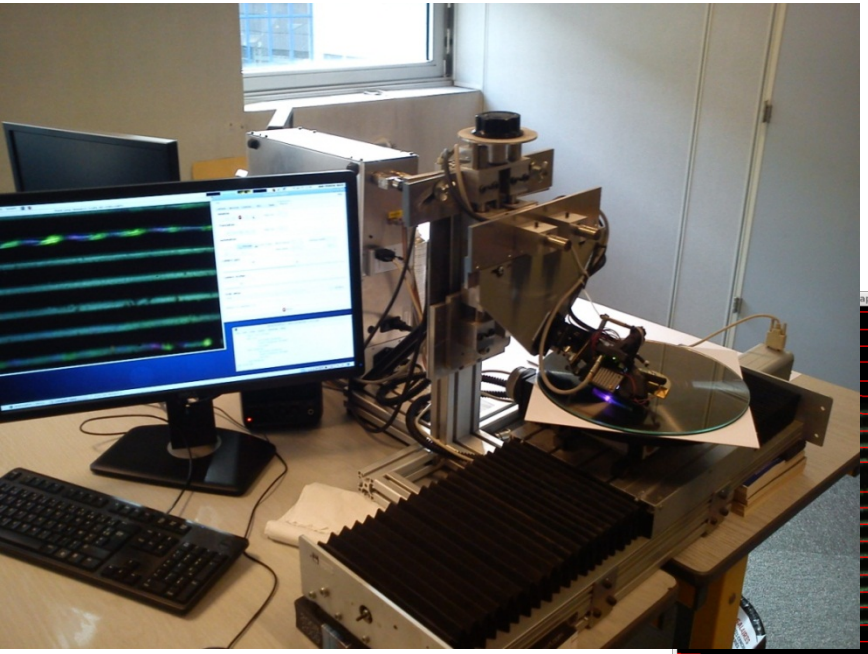
Exemple du projet SAPHIR, à l'Ina : récupération de données sur supports défectueux

20 000 disques concernant la mémoire radio de 1930 à 1960

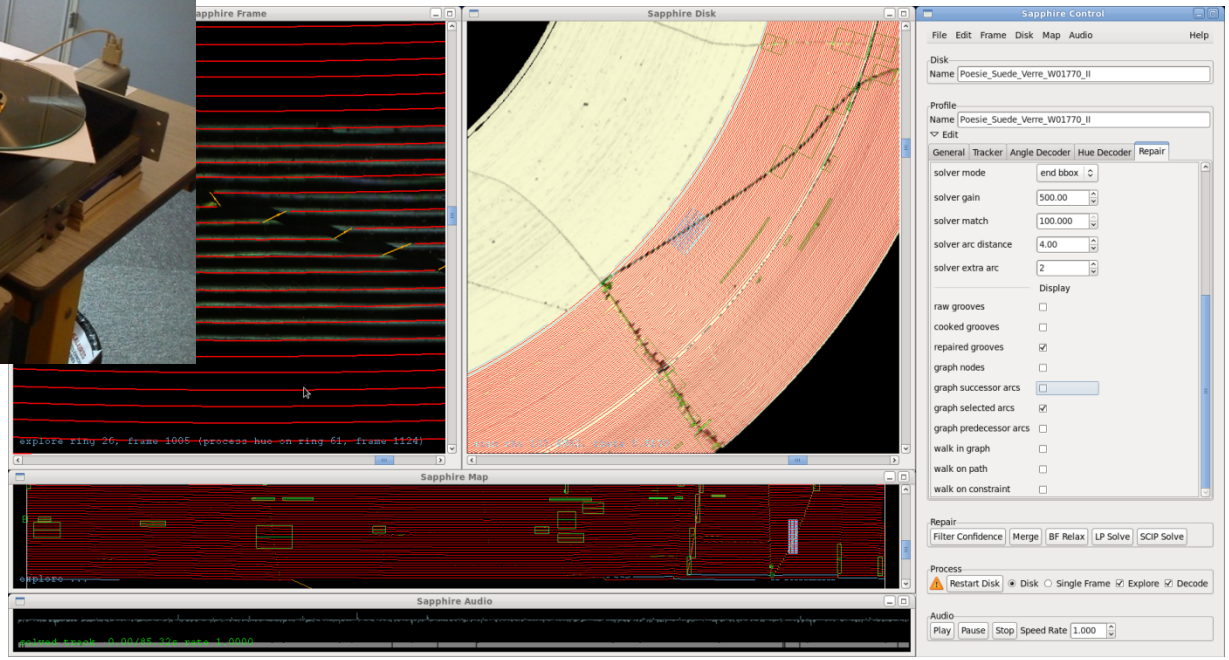
Lecture avec contact impossible



Lecture de l'angle du sillon par la couleur réfléchie et



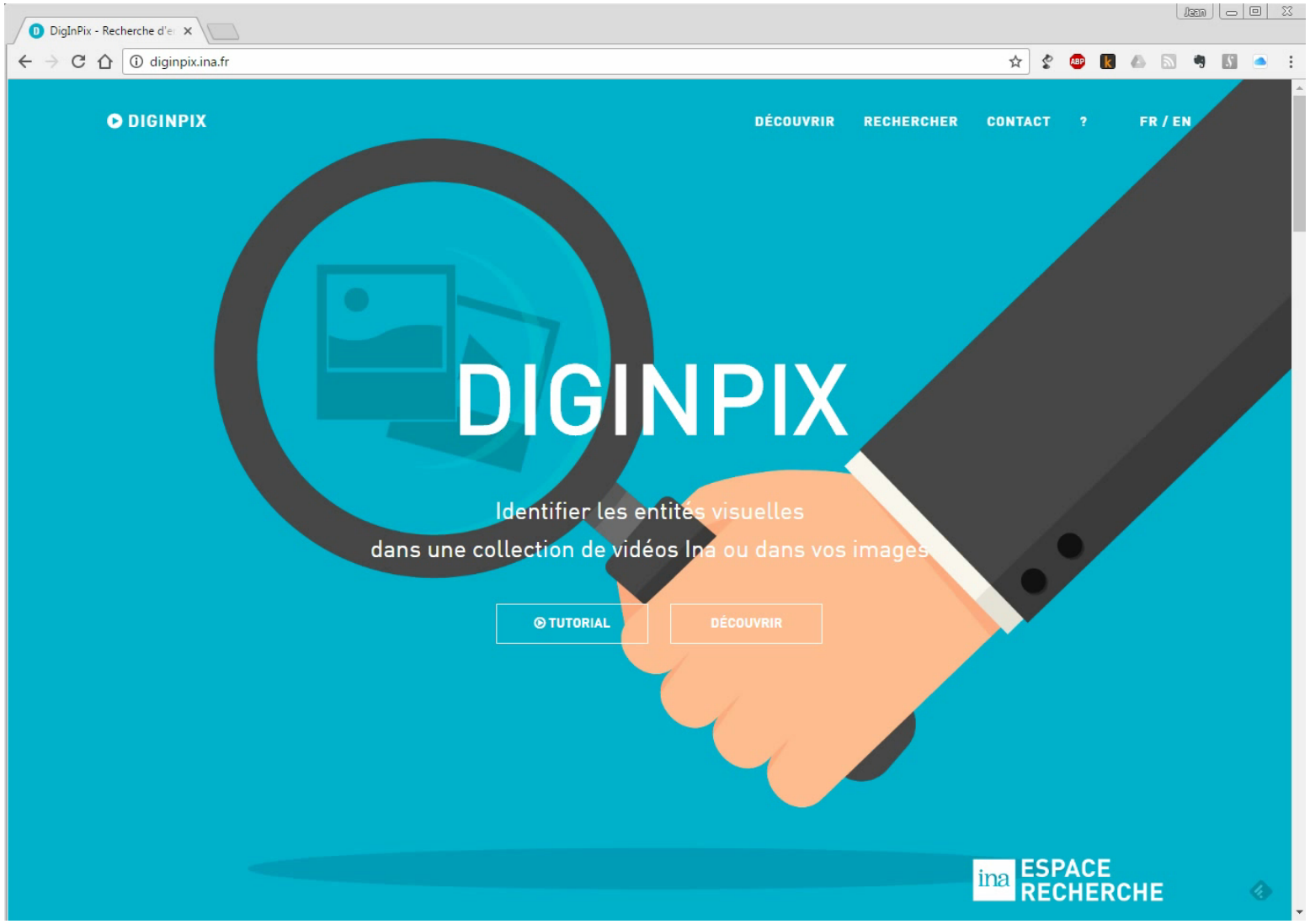
Décodage et reconstitution de disque fracturé



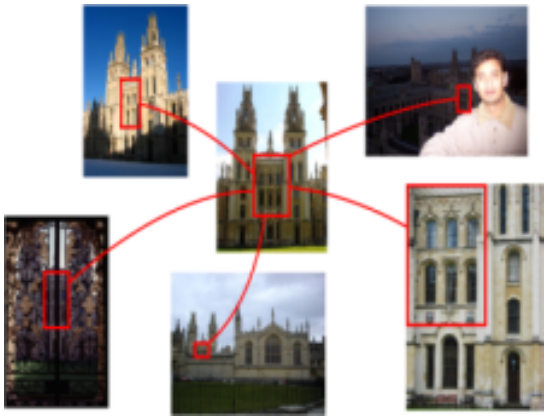
SIGNATURE : signature de vidéos

Empreinte numérique, fingerprinting

DIGINPIX
Reconnaissance de peintures dans des archives vidéo


DigInPix - Recherche d'e
diginpix.ina.fr
DIGINPIX
DÉCOUVRIR
RECHERCHER
CONTACT
?
FR / EN
DIGINPIX
Identifier les entités visuelles
dans une collection de vidéos Ina ou dans vos images
TUTORIAL
DÉCOUVRIR
ina ESPACE RECHERCHE

Recherche rapide de parties d'images (bâtiments, logos…)



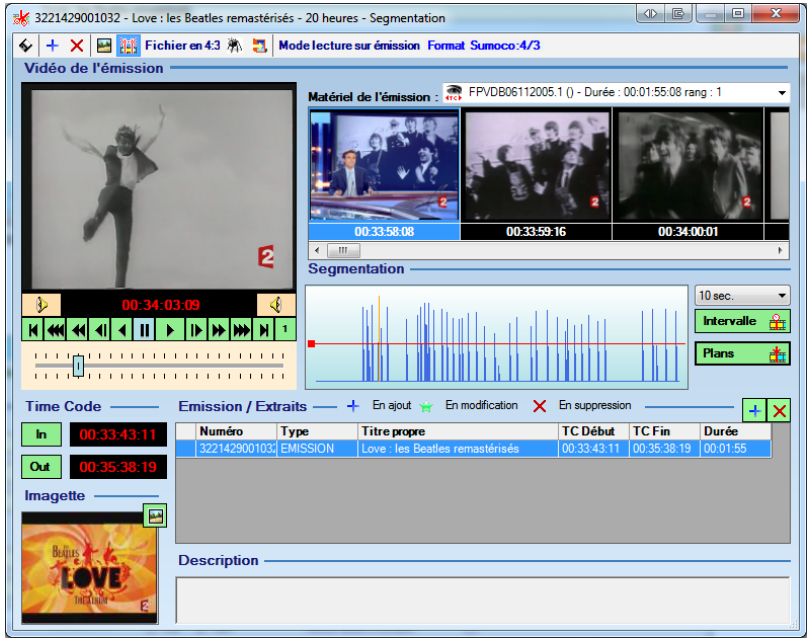
Détection de « concepts » visuels

Reconnaître Londres, Prague et Paris

Paris, France | Prague, Czech Republic | London, England

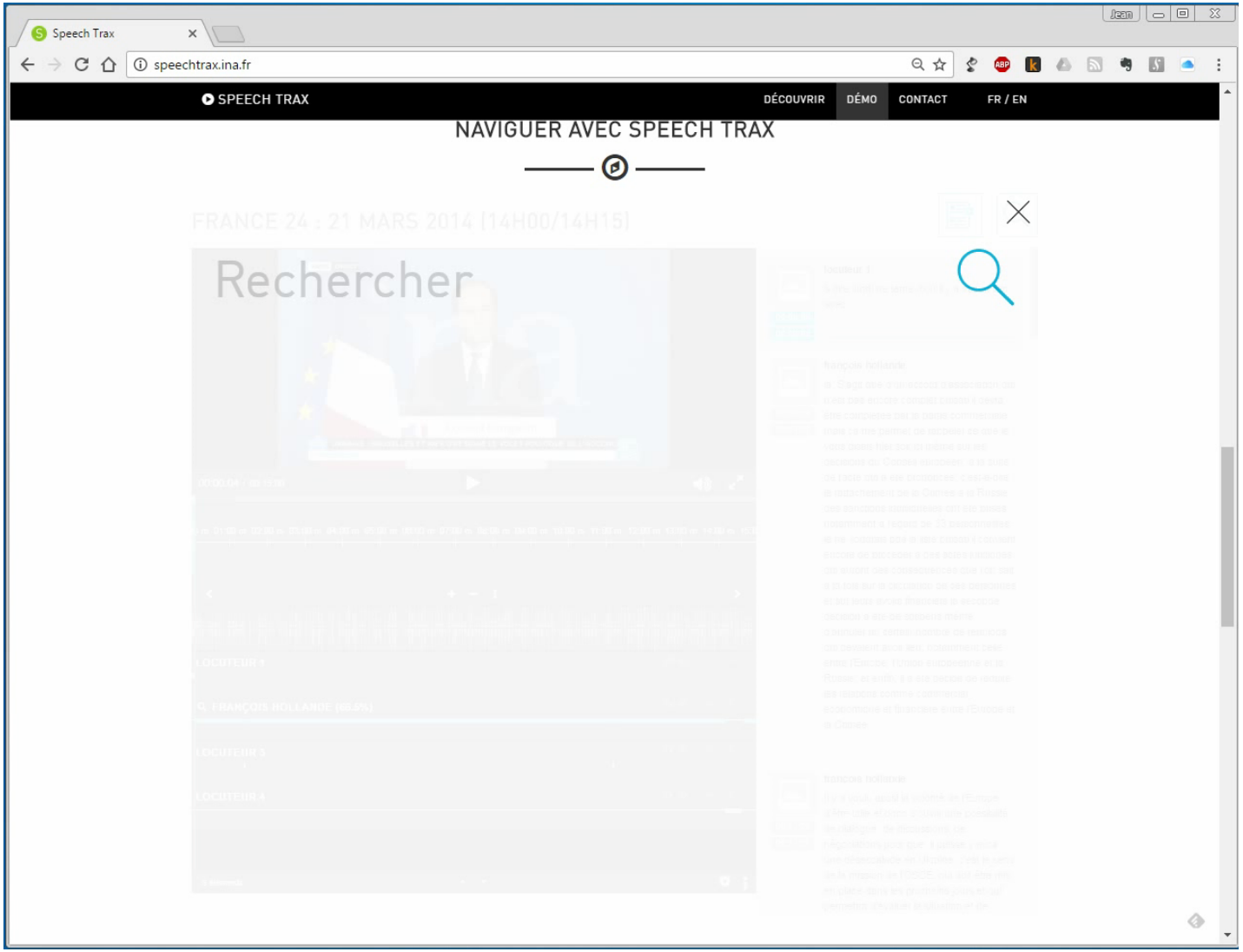
Source : *What makes Paris look like Paris ?* [Doersch *et al.*, 2012]

DESIGN PRINT
Automatisation de la segmentation

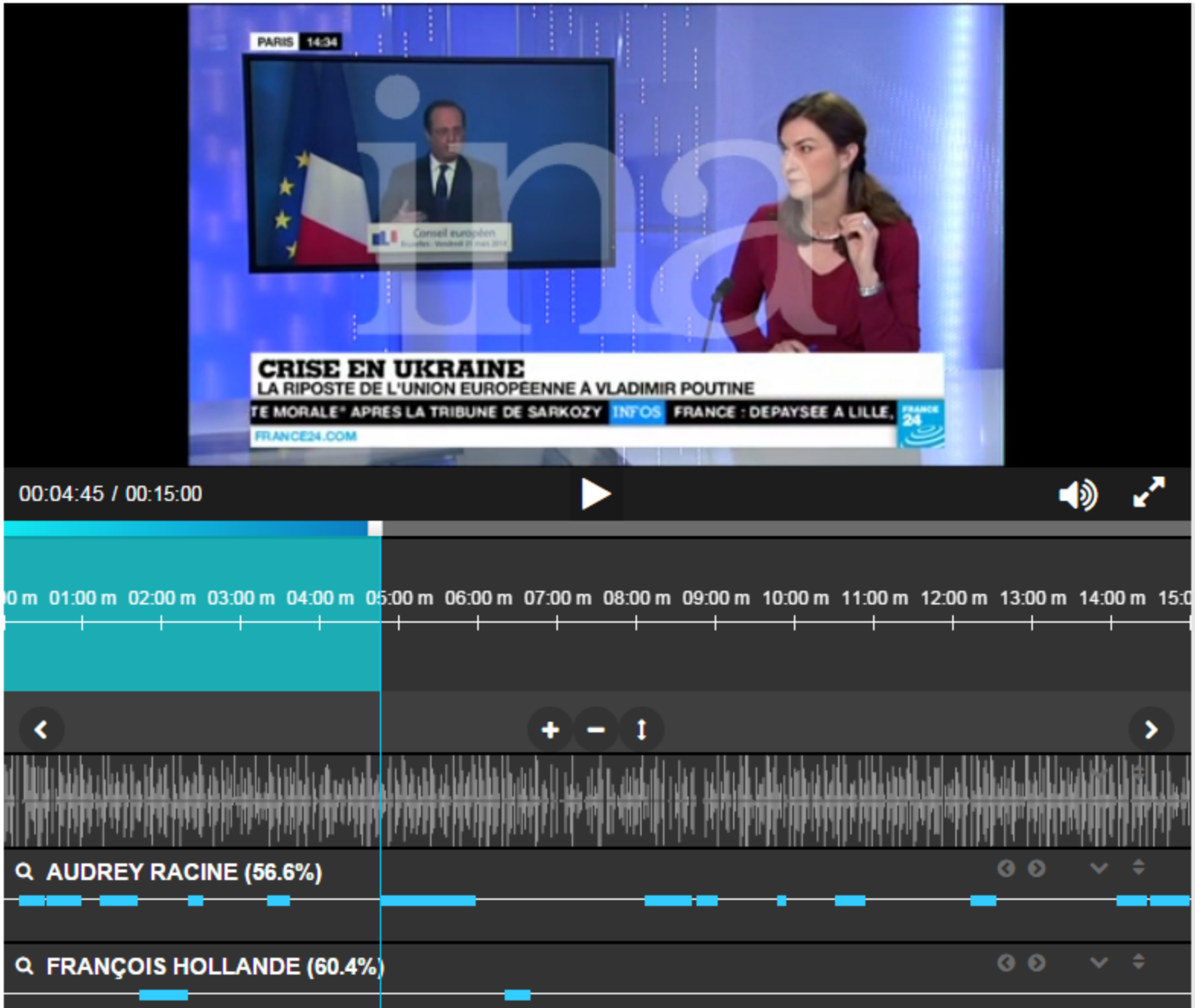


La segmentation des fichiers numériques (pose de TC in/out) est une opération essentiellement manuelle, laborieuse et chronophage. L'Ina a mis au point un système de segmentation automatique qu'il a testé sur les journaux télévisés de 1989. Le projet a consisté à imaginer une interface qui permette de valider, corriger et capitaliser ces repérages temporels automatiques.

SPEECH TRAX
Identification de voix célèbres dans des archives vidéo


Speech Trax
speechtrax.ina.fr
SPEECH TRAX
DÉCOUVRIR
DÉMO
CONTACT
FR / EN
NAVIGUER AVEC SPEECH TRAX
FRANCE 24 : 21 MARS 2014 (14H00/14H15)
Rechercher

FRANCE 24 : 21 MARS 2014 (14H30/14H45)



audrey raci

00:04:45
00:05:57

la peur juste
parlera dans
focus mais e
annoncé la f
avec la Russ
toutes les m
déclaration a
français de l
actuellement
et au bord d
a également
France était
avions de ch
renforcer sa
baltes des sa
pas vraimen
président Vla
la dernière m
Crimée à la
d'application
plus tôt sans
Parlement ru
l'unanimité le
les faits, tout
mort. Dans l'
échanges de
l'opposition e
socialistes a

UTILISATION :

→ Pour les documentalistes :

Proposer une première documentation brute et imparfaite des contenus indexés par l'Ina pour amorcer le travail documentaire

→ Pour les clients et usagers de l'Ina :

Proposer des parcours d'archives obliques reposant sur les interventions orales de personnalités du paysage audiovisuel français

DATASET
Ouvrir les données Ina pour la Recherche

→ Mise à disposition de la communauté scientifique et technologique d'un corpus de documents audiovisuels issus de ses collections, de fiches documentaires et de métadonnées associées à ces documents.

Ce corpus est destiné à la mise au point, l'expérimentation et l'évaluation d'outils de recherche et d'analyse de contenus multimédias dans un strict cadre de recherche scientifique.

http://dataset.ina.fr

04

La mutation du métier de documentaliste audiovisuel

La mutation du métier de documentaliste audiovisuel

→ L'éditorialisation prends une place majeure

Le documentaliste n'est plus qu'un médiateur, il enrichit le document brut par des données.

- exemple de data + archives (tour de France) :

La mutation du métier de documentaliste audiovisuel

→ Automatisation de certaines taches chronophages

Indexation automatique : lancer des calculs par lot

→ Bâtiments

→ Personnalités

→ Logos

→ Segmentation

→ Lien donnée


C'EST TOUT
POUR AUJOUR
D'HUI

Merci de votre attention

Délégation Ina Nord
(Hauts de France / Normandie)
8 Allée de la Filature 59000 Lille

Clément Mouly
Responsable Documentaire
cmouly@ina.fr