#dhnord2021

Sur l’hétérogénéité des données

Pratiques consacrées à l’homogénéisation des données des différents fonds d'archives audiovisuelles (projet CROBORA)

Shiming SHEN (Chercheuse associée à l’INA, SIC.Lab Méditerranée, Université Côte d'Azur, shiming.shen@univ-cotedazur.fr)

Novembre 2021

Problème rencontré

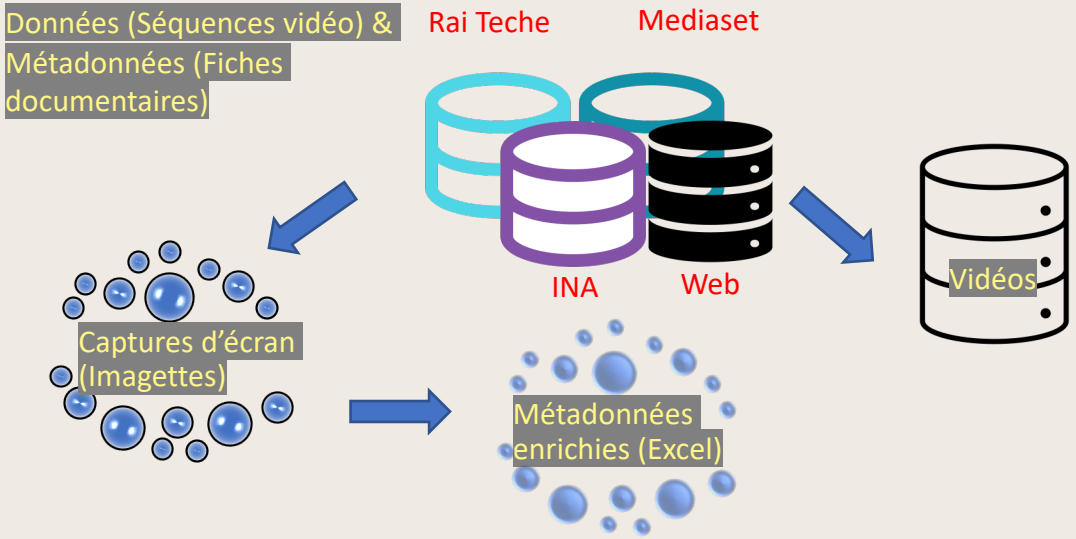
Nature des données

Dans le cadre du projet CROBORA, on collecte les ***réutilisations d'images d'archives dans les émissions télévisuelles et sur le Web*** racontant la construction de l'Union européenne. Le travail de recherche vise à collecter les éléments suivants :

1. **Les métadonnées existantes** produites par les documentalistes de l'INA, Rai Teche et Mediaset + **métadonnées issues du Web**,
2. **Des captures d'écran/images fixes** des images d'archives incrustées dans les vidéos,
3. De nouvelles **métadonnées** décrivant **les images fixes** produites par les chercheurs du projet,
4. **Les vidéos**

Hétérogénéité

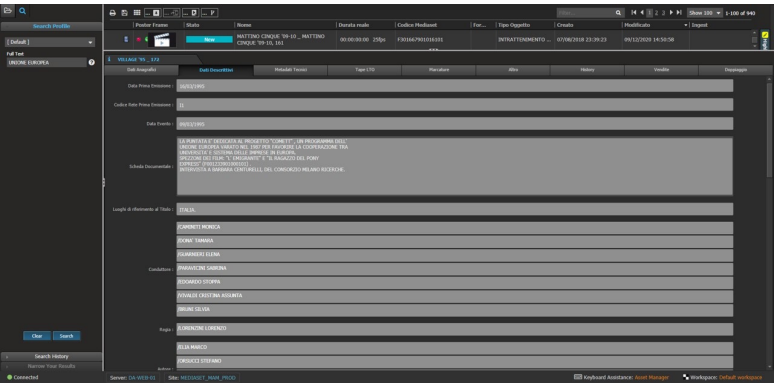
1. **Hétérogénéité** des **institutions** des fonds d'archives impliquées, ce qui conduit à des structures différentes des métadonnées: INA, RAI TECHE, MEDIASET, Web et DLWEB
2. **Terrain international** (France et Italie),
3. **Plusieurs chercheurs** impliqués, qui collectent et indexent les données différemment



Exemple : Signature du traité de Maastricht

Fonds d'archives différents → Métadonnées différentes

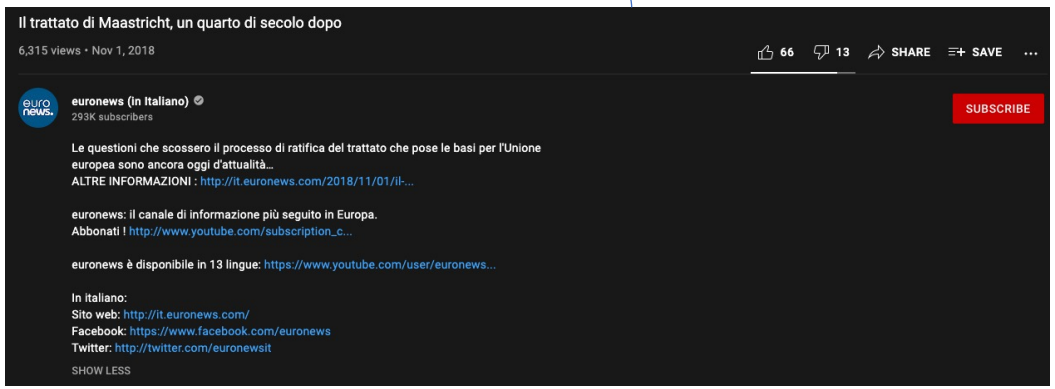
MEDIASET (Italie)



INA (France)

ID Notice	2846386.001.028
Titre propre	[Union européenne : les crises]
Titre collection	TF1 20 heures
Titre programme	TF1 20 heures : [émission du 30 Mai 2005]
Chaîne de diffusion	TF1
Date de diffusion	30/05/2005
Jour	lundi
Statut de diffusion	Première diffusion
Heure de diffusion	20:26:27
Heure de fin de diffusion	20:27:44
Durée	00:01:16:14
Genre	Journal télévisé
Type de description	Sujet
Générique	JOU,Leenknegt Florence
Descripteurs	Union européenne ; histoire ; crise ; vie politique ; traité
Chapeau	Reportage. Evocation des grands moments de crise dans l'histoire de l'Union européenne depuis la signature du Traité de Paris en 1950. Commentaire sur images d'archives.
Date d'enregistrement	30/05/2005
Société de programmes	Télévision Française 1
Producteurs	Producteur, Boulogne Billancourt : Télévision Française 1, 2005
Extension géographique	National
Sélection DL	O
Numéro DL	CL T 20050530 TF1 16h
Base	Dépôt Légal TV
Fonds	Télévision Francaise 1 Actualités
Titre matériel	[Journée de captation TF1 du 30 mai 2005]

WEB (Youtube)



Chercheurs différents → Captures d'écran différentes

(Même s'il s'agit d'une même séquence vidéo !)

INA

MEDIASET

WEB

Chercheurs différents → Indexation différente

	Personnalité	Événement	Lieu	Illustration
INA	Roland Dumas	Construction européenne		Signature
MEDIASET	Politico		Europa	Riunione
WEB		traité de Maastricht	Maastricht	Document, Signature

Solutions

Objectif Suivre les réutilisations d'une même archive audiovisuelle à travers différents contextes

1. Des pratiques contrôlées/partagées

Workflow/Protocole établi *à priori*
(Mais Mise à jour/Interaction en permanence)

These information (i.e. EU, EU-CITED and IRRELEVANT) are encoded in the field **video-as-subject** in the Excel file
Ex.:
video-as-subject = EU

T5: For videos marked as "EU", **create a snapshot for every sequence of archives** reused in the video (in each video there may be more than one sequence);
For videos marked as "EU-CITED", create a snapshot only for the sequence of archives reused in the video which contains EU representations.

Every snapshot is named and encoded as follows, ex.
DATE_TITLE_PROGRAM TITLE_BRODCASTER
230101_PaponCourDroitsDeLHomme_20h_TF1
Syntax for creating names:

1. DATE: DDMMYY
2. TITLE: Capitalize the first letter of each word
3. PROGRAM TITLE: Once the program title is settled, make sure that every member concerned is aware of and uses the same program title
4. BROADCASTER: Once the broadcaster name is settled, make sure that every member concerned is aware of and uses the same broadcaster name
5. GENERAL PRINCIPAL: No space is allowed; No letter with accent mark is allowed (Ex. à); No punctuation is allowed except "_" for separating the different parts of name

IF there are more than one sequence of archives reused in a video (which is often the case):

1. For the first sequence, the syntax is the same as above. Ex. 250101_ComparatifAgesDesRetraitesEnEurope_13h_TF1
2. From the second sequence, add "_Number" at the end. Ex. 250101_ComparatifAgesDesRetraitesEnEurope_13h_TF1_2; 250101_ComparatifAgesDesRetraitesEnEurope_13h_TF1_3...

Where to store and share your snapshots:

1. Create a shared space on OneDrive or any other cloud
2. Create a file for each channel (file should be named after channel name) or platform
3. Create two subfiles under each channel file (one is named as "Snapshots with ID"; the other is named as "Atlas_Channel name". Ex. Atlas_TF1)
4. Put all of the snapshots of one video into the same file (file name: **ID**), then put this file into the file "Snapshots with ID"; For the other file "Atlas_Channel name", just put all the snapshots into it

Why creating two subfiles for the same data:
It's just about two different ways of organization. Via the subfile "Snapshots with ID", we could know the relations between different snapshots since those from the same video

Laetitia Biscarrat 10:04 AM Jun 23

Juste une interrogation au cas où sur l'identification des captures: 230101_PaponCourDroitsDeLHomme_20h_TF1
Le fait d'utiliser les tirets bas il me semble que ça neutralise les termes dans Iramuteq. Bon, je vois mal en quoi passer les légendes sous iramuteq est pertinent mais je le signale car ça peut être le cas pour d'autres outils de dataviz que vous pensez utiliser peut-être?

Peppe Cavallari 9:42 AM Sep 29

Petite question : lorsqu'il s'agit d'un acronyme, p.e. CECA ou UE, faut-il écrire en majuscule seulement l'initiale (Ceca, Ue) on toute lettre composant l'acronyme ?

布朗 5:58 PM Today

Bonjour Peppe, je choisirais CECA/UE

2. Thésaurus partagé

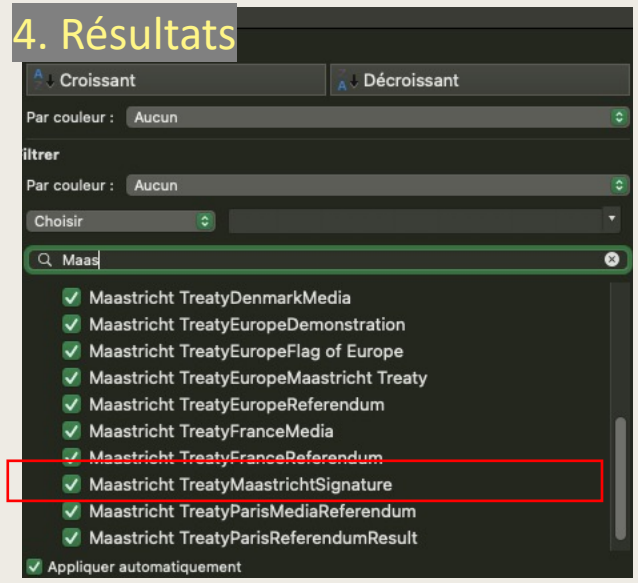
Un vocabulaire contrôlé permettant une formalisation de l'indexation

ENCODING-NAME	Lexicon	Description
Maastricht Treaty	Traité de Maastricht, Trattato di Maastricht	Concluded in 1992 between the 12 member states of the…
Maastricht		City and a municipality in the southeastern Netherlands
Signature	Scratch of pen, Firma	Name, written in someone's own characteristic way, often at the…

3. Indexation contrôlée

	Personnalité	Événement	Lieu	Illustration
INA		Maastricht Treaty	Maastricht	Signature
MEDIASET		Maastricht Treaty	Maastricht	Signature
WEB		Maastricht Treaty	Maastricht	Signature

4. Résultats



Toutes les réutilisations de la séquence sur la signature du traité de Maastricht