IA x INA

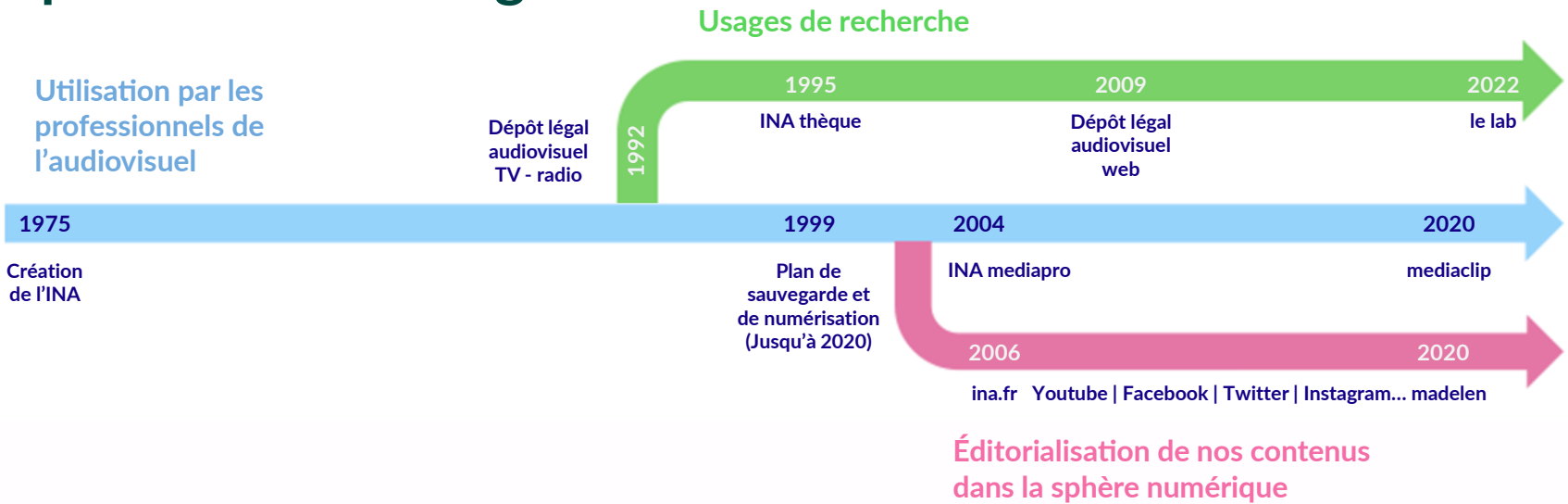
De nouvelles clés de navigation ?

Eléonore Alquier, directrice adjointe data et technologies - INA
Xavier Lemarchand - directeur coordination et intégration IA - INA
Journée d'étude Mediadix – Paris Nanterre – 13 décembre 2024

01.

→ Les enjeux de l’IA pour l’INA

Accroître la capacité de description des fonds audiovisuels pour tous ses usages



Accroître la capacité de description des fonds

Chiffres clés (fin 2023)

Dépôt légal audiovisuel tv-radio :
- **101** chaînes TV, **82** chaînes radio
- **25 millions d'heures**
- (+ 1,5 millions d'heures par an)

Dépôt légal audiovisuel web :
- **5,8 millions d'heures**

Fonds archives professionnelles :
- **2,5** millions d'heures

→

± 30 millions d'heures = 3400 ans 24/7…

La description des fonds est un défi crucial pour s'y orienter, les rendre découvrables, les analyser, les exploiter.

Jusqu'alors, les descriptions proviennent

3

du traitement documentaire

Écriture de chapôs, résumés, choix des descripteurs…

2

du travail de catalogage

Synchroniser, aligner et ajouter des données selon notre modèle (chaînes, titres, horaires, producteurs, typologies, génériques…)

1

de l'import de données externes

Données prévisionnelles de diffusion (Plurimédia, diffuseurs), données réelles de diffusion (Médiamétrie), données descriptives (diffuseurs, producteurs, agences de presse — Onclusive…)

➔ Mais la **description fine des contenus ne concerne qu'une partie des fonds**

Enjeux des outils d'IA de description automatique

- accroître nos capacités de description des programmes **en finesse, en richesse, en exhaustivité**.
- **rendre découvrable** une immense partie de nos fonds
- **ouvrir de nouveaux champs d'analyse et de recherche** (Dépôt légal) et **d'exploitation** (Fonds pro)

02.

→ Un premier cas limité : la segmentation automatique LCI et CNEWS

Ambition des *projets* IA… en 2019

Générer des données pour améliorer le service à nos usagers internes et externes

Démarche itérative :
- Tests d'outils
- Elaboration de cas d'usages avec les équipes de catalogage et de documentation
- Organisation de projets portés par des pilotes cadres « métier »

Effet d'opportunité dans un contexte de mise en place d'un système d'information centralisé et d'un modèle de donnée unique

Cas d'usage : *segmenter automatiquement* une journée de programme de chaînes d'information continue

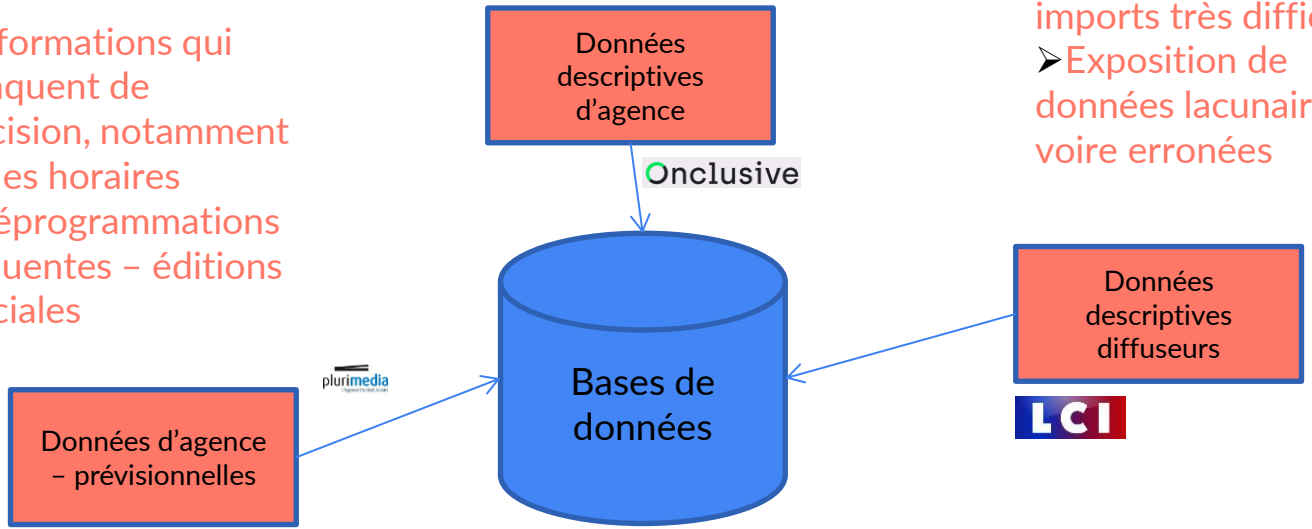
Cadre : mission de dépôt légal portant sur les programmes radiodiffusés, télédiffusés et sur le web média.

1/Identifier les programmes diffusés avec des éléments d'identification de type catalogue pour constituer la bibliographie nationale.

2/Enrichir la description des programmes pour des usages de recherche et pour des usages professionnels (BtoB et BtoC) sur un périmètre plus restreint lié à nos conventions avec l'audiovisuel public ou à des mandats de distribution.

Agréger plusieurs *sources* de données

➢Informations qui manquent de précision, notamment sur les horaires
➢Déprogrammations fréquentes – éditions spéciales



➢Réconciliations des imports très difficiles
➢Exposition de données lacunaires voire erronées

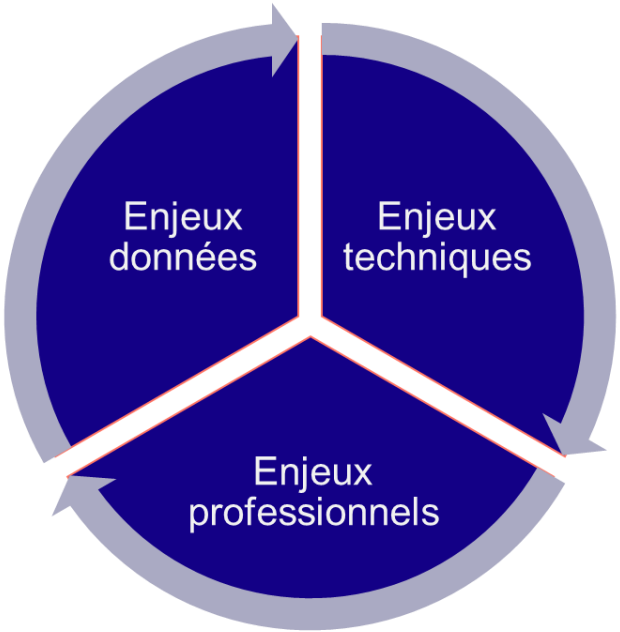
Logos des titulaires actuels des marchés de fourniture de métadonnées

Motivation **et** *enjeux* **du projet**

Utiliser les **images** et les **sons** pour générer les données
Traiter plus efficacement (qualité et précision) la **masse** d'images captées au titre du dépôt légal

Tester et constituer une **boîte à outils** de traitement automatique organisés en workflow de traitement

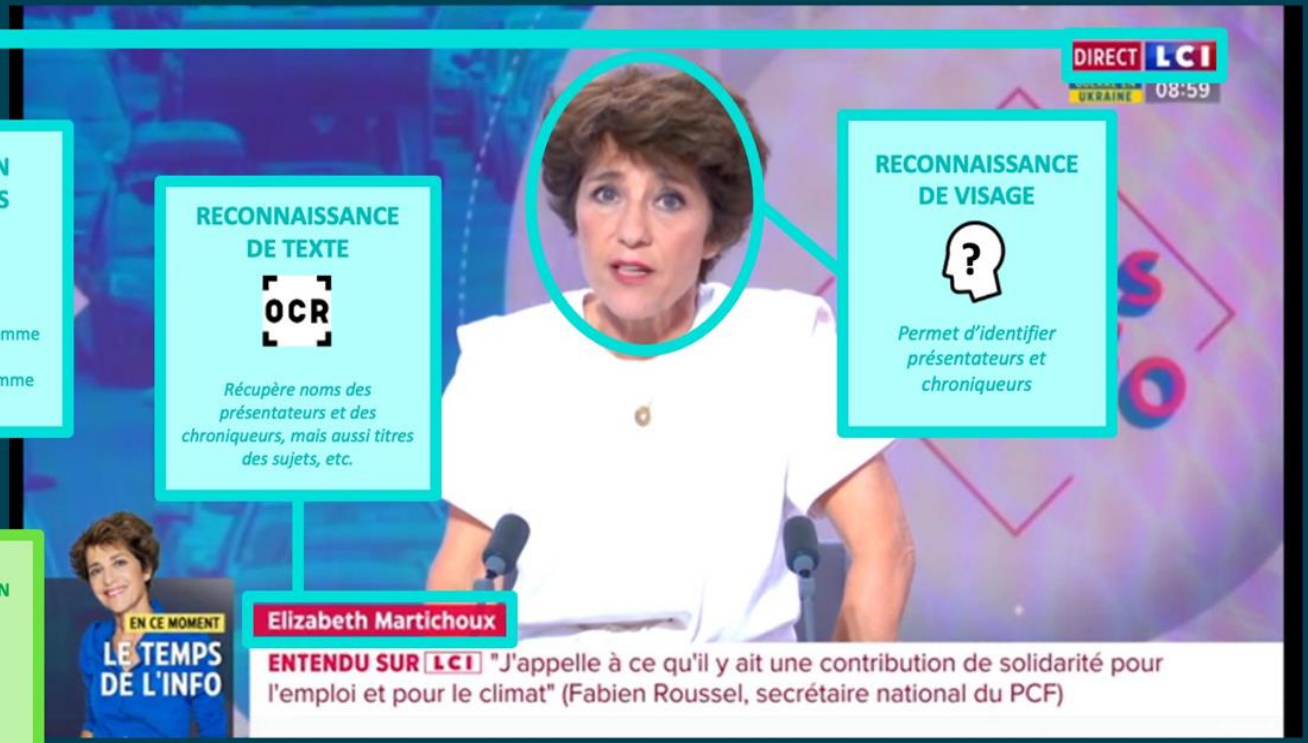
Déployer des outils reposant sur **l'analyse documentaire** appliquée aux techniques **d'apprentissage machine**



1 Segmenter une chaîne d'information : quels moyens?

Moyens informatiques

Moyens essentiels



DETECTION DES LOGOS

Avec logo = programme
~~sans logo~~ = plage interprogramme

RECONNAISSANCE DE TEXTE

OCR

Récupère noms des présentateurs et des chroniqueurs, mais aussi titres des sujets, etc.

RECONNAISSANCE DE VISAGE

Permet d'identifier présentateurs et chroniqueurs

CLASSIFICATION D'IMAGE

Identifie des types d'images : les plateaux, reportages, etc...

INFOS DE PROGRAMMATION

Reconstitution chronologique de la programmation

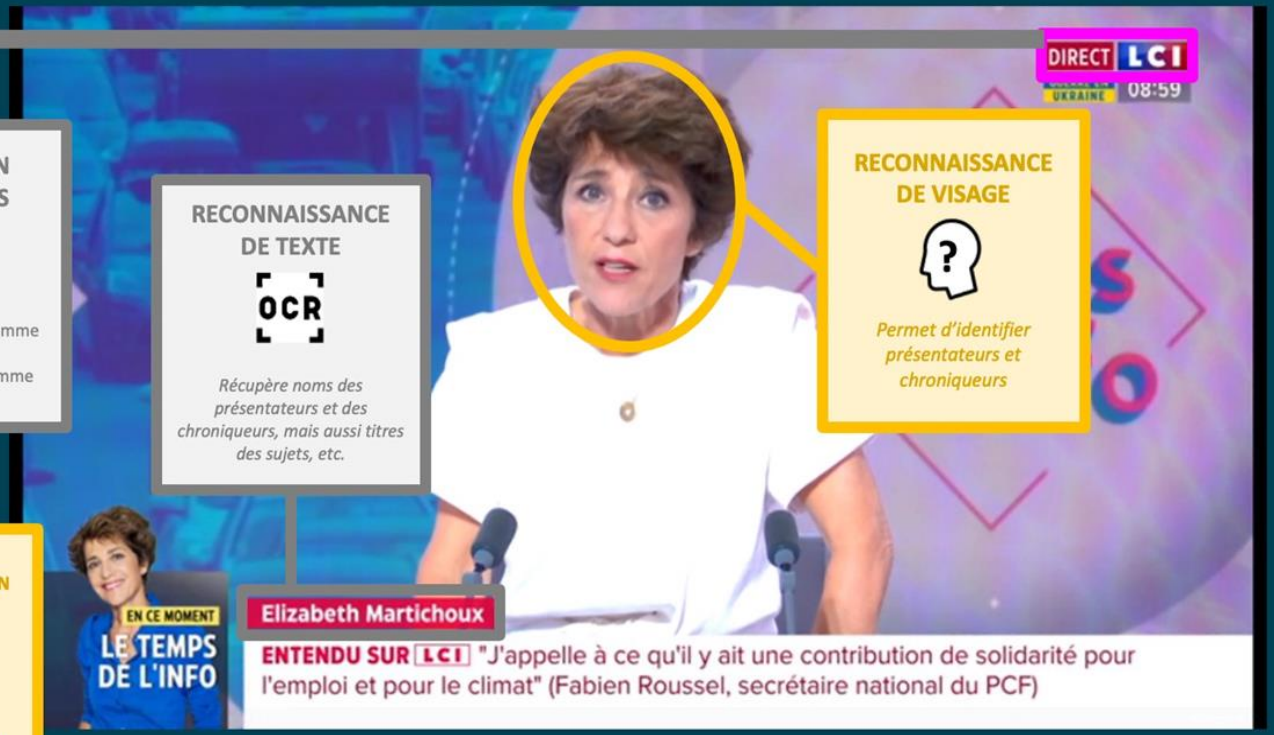
CLASSIFICATION D'IMAGE

Permet d'identifier les génériques et les éléments toujours identiques

2 Segmenter une chaîne d'information : quels moyens?

Maintenance occasionnelle

Maintenance régulière



DETECTION DES LOGOS

Avec logo = programme
~~sans logo~~ =
plage interprogramme

RECONNAISSANCE DE TEXTE

OCR

Récupère noms des présentateurs et des chroniqueurs, mais aussi titres des sujets, etc.

RECONNAISSANCE DE VISAGE

Permet d'identifier présentateurs et chroniqueurs

CLASSIFICATION D'IMAGE

Identifie des types d'images : les plateaux, reportages, etc...

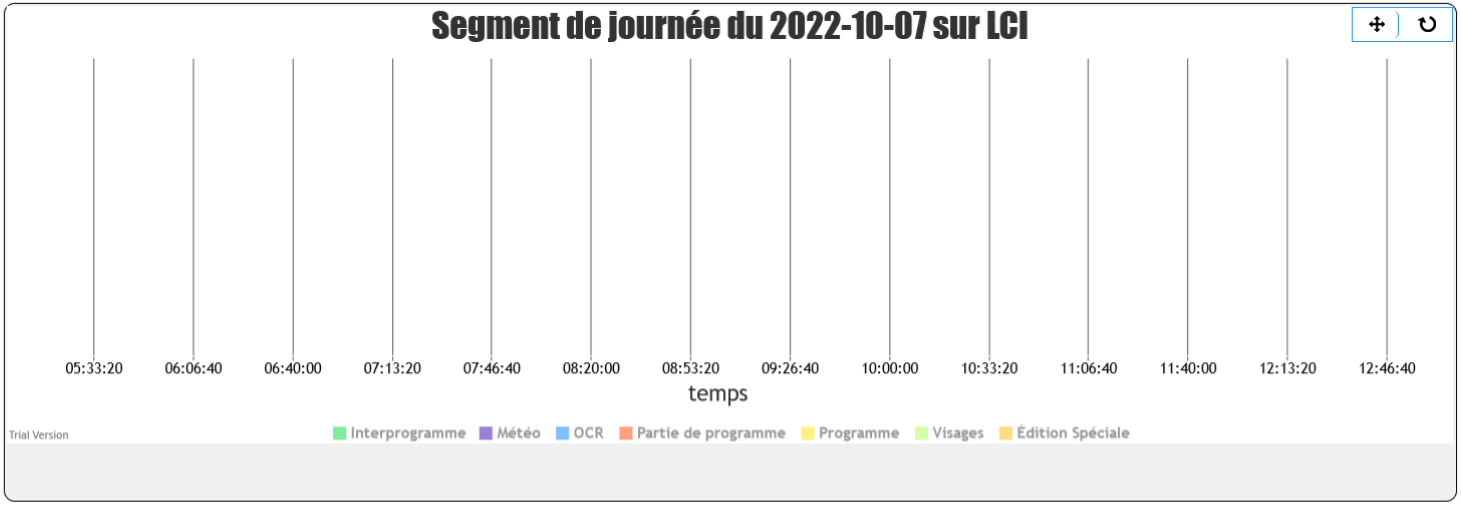
INFOS DE PROGRAMMATION

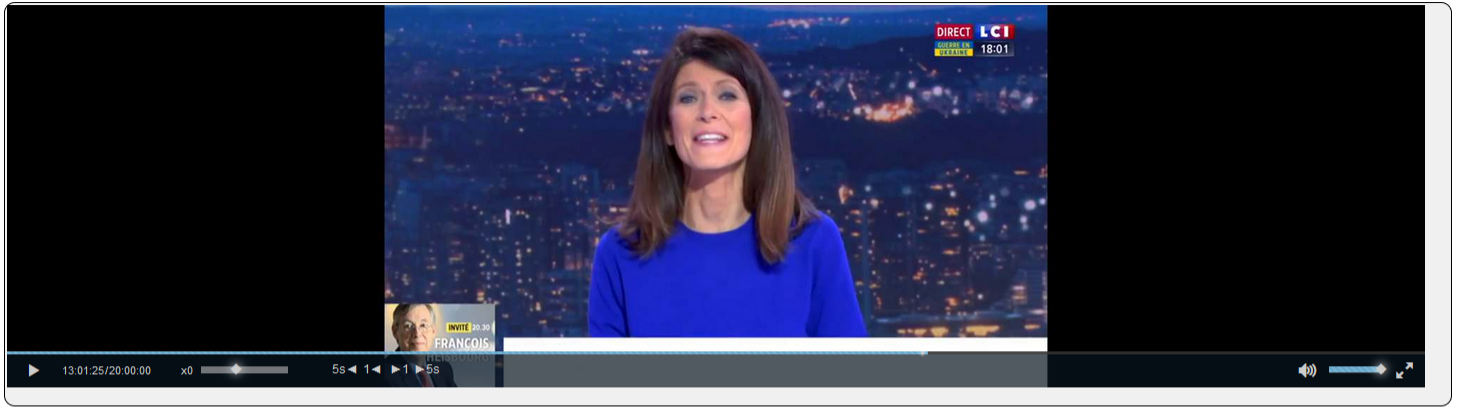
Reconstitution chronologique de la programmation

CLASSIFICATION D'IMAGE

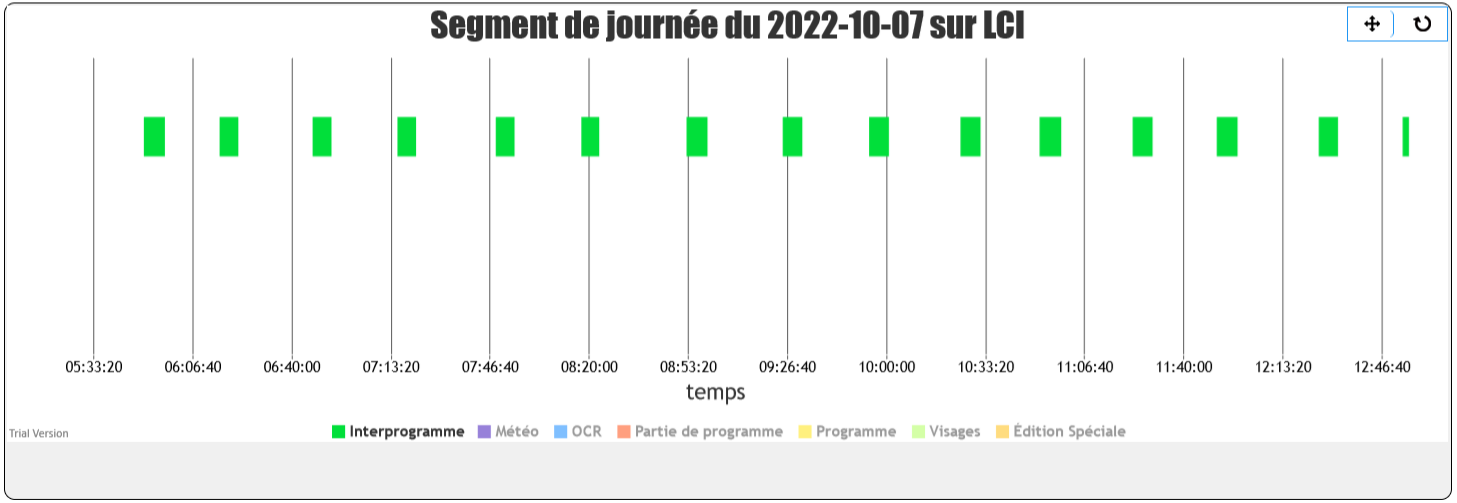
Permet d'identifier les génériques et les éléments toujours identiques

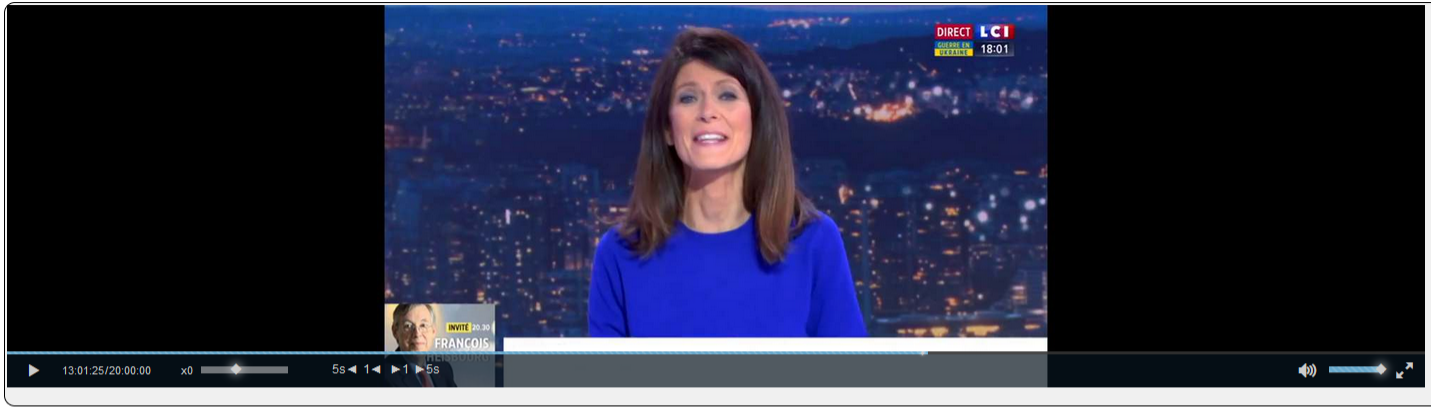
Exporter la journée



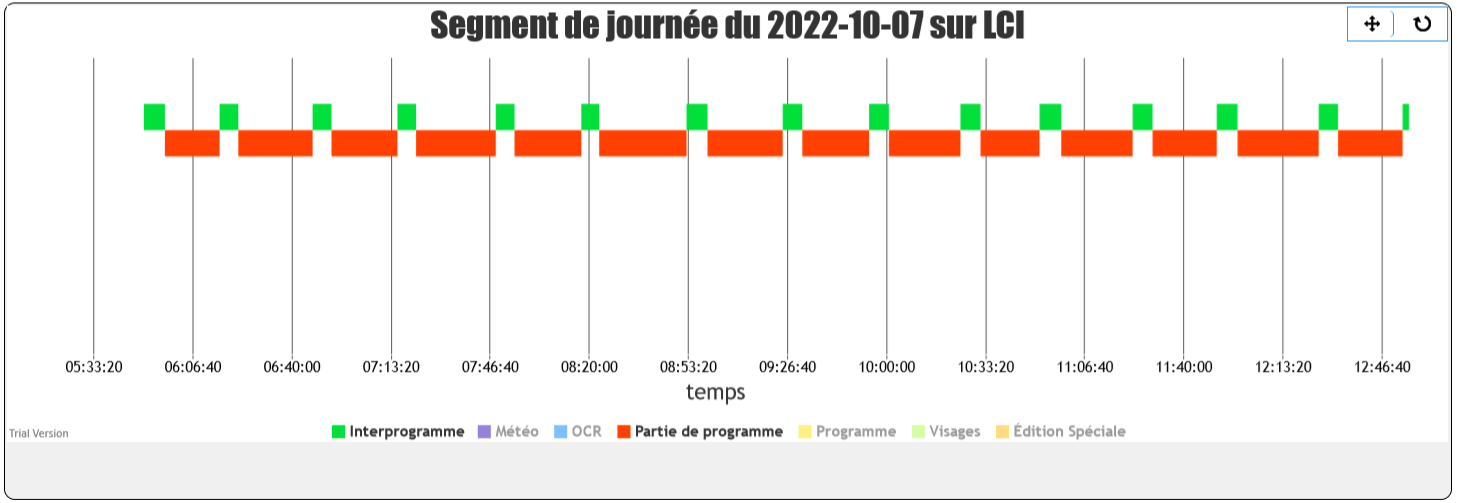


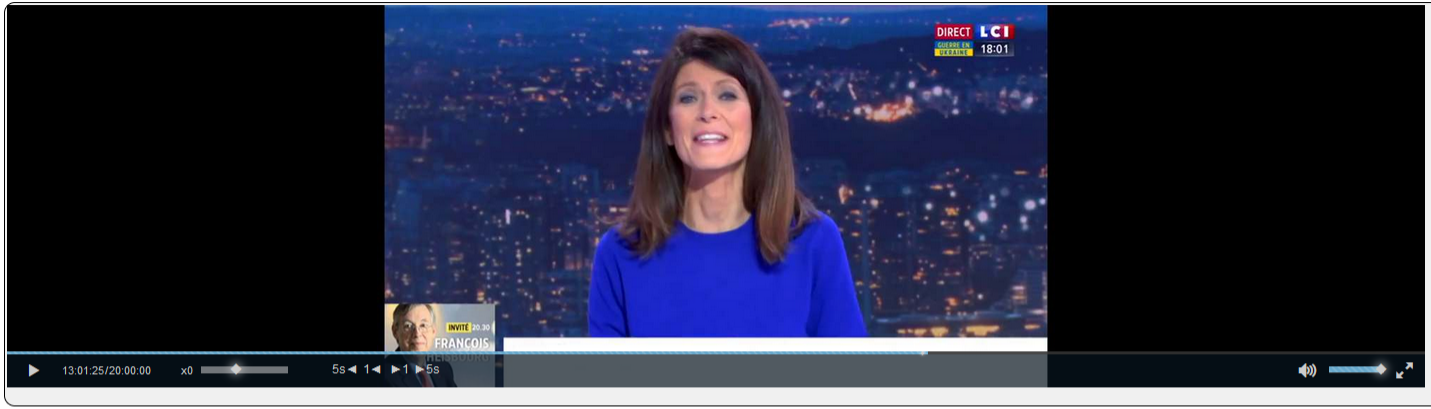
Exporter la journée



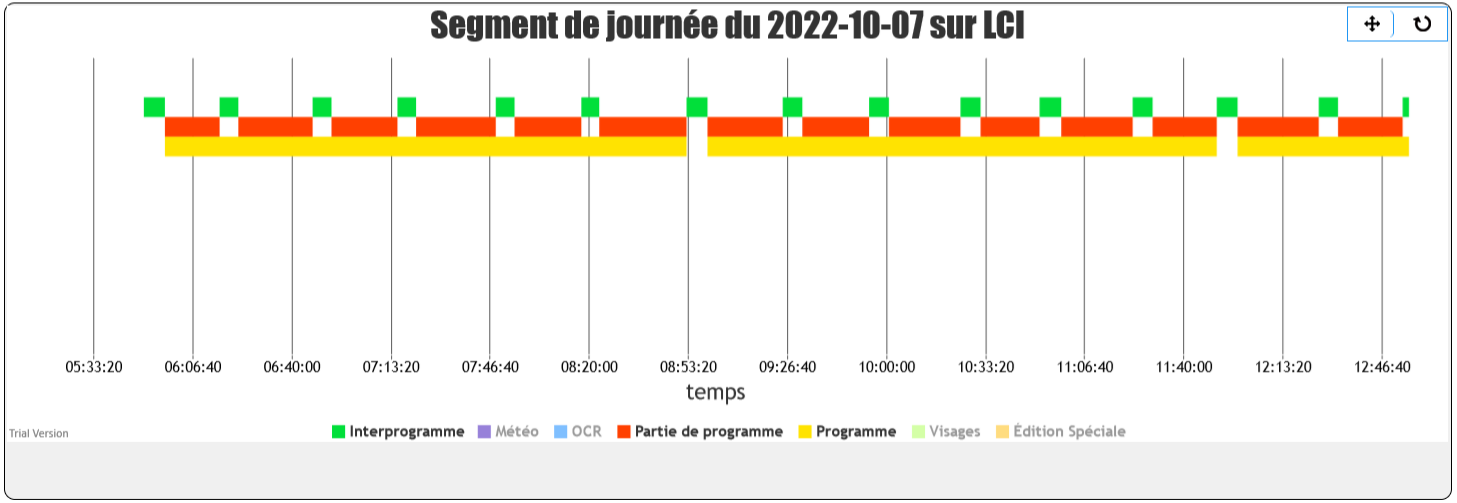


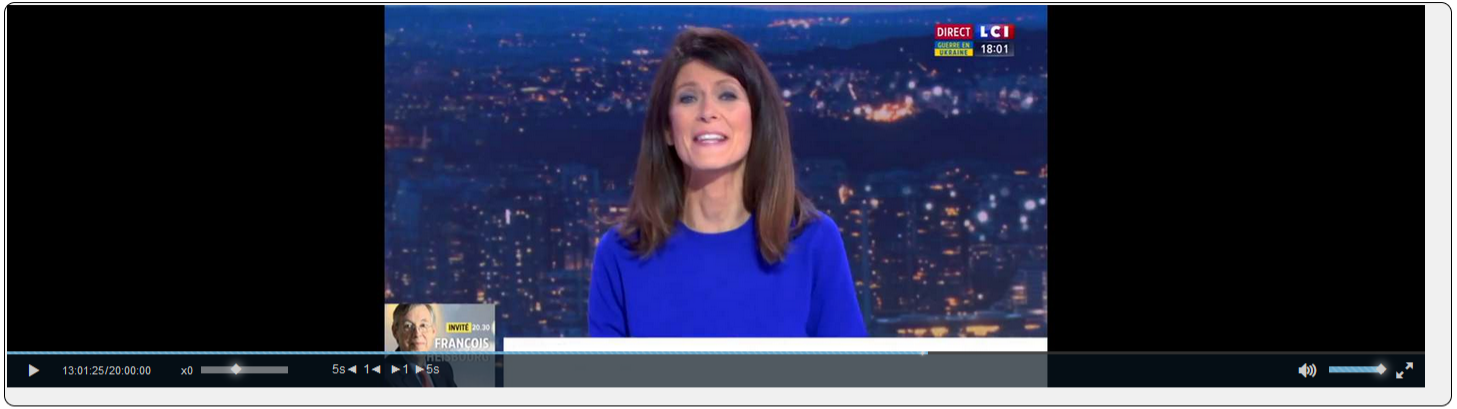
Exporter la journée



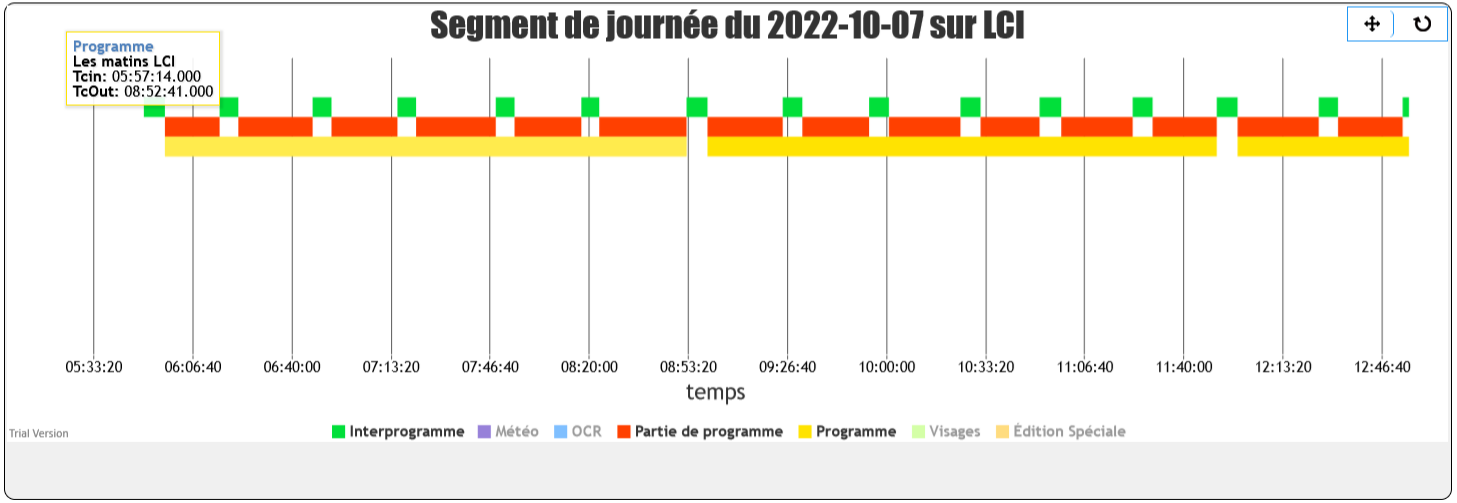


Exporter la journée

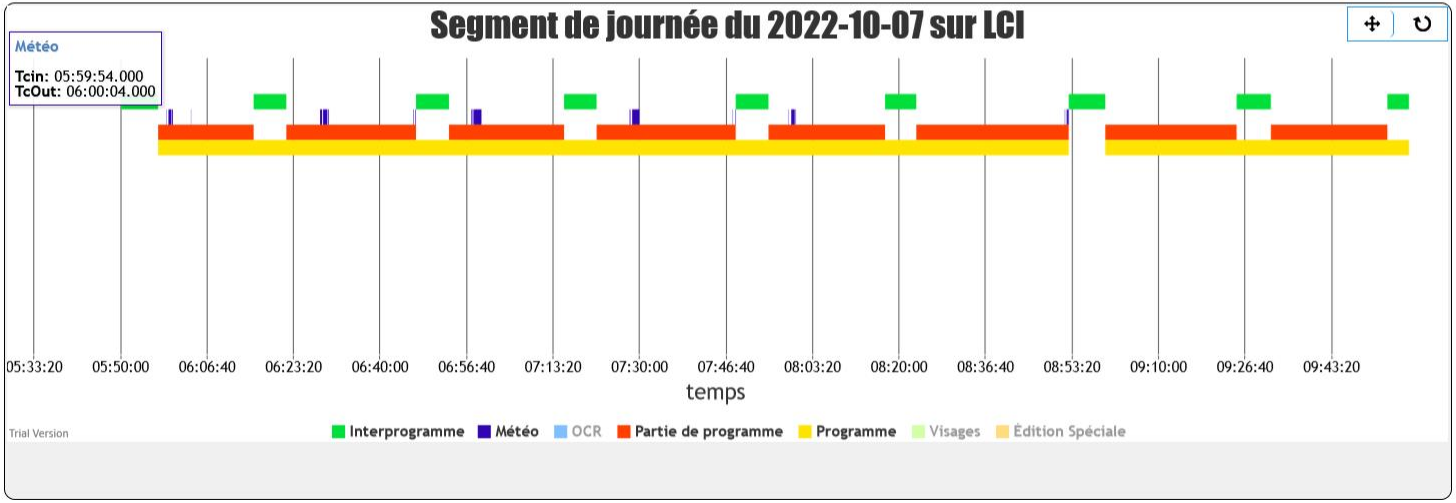




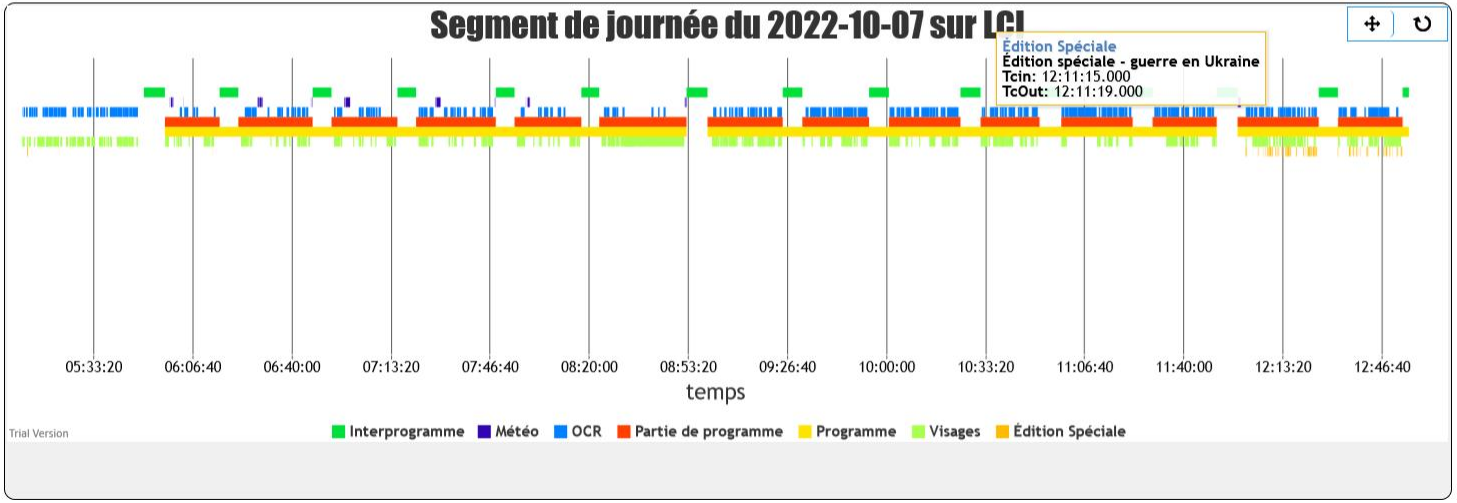
Exporter la journée



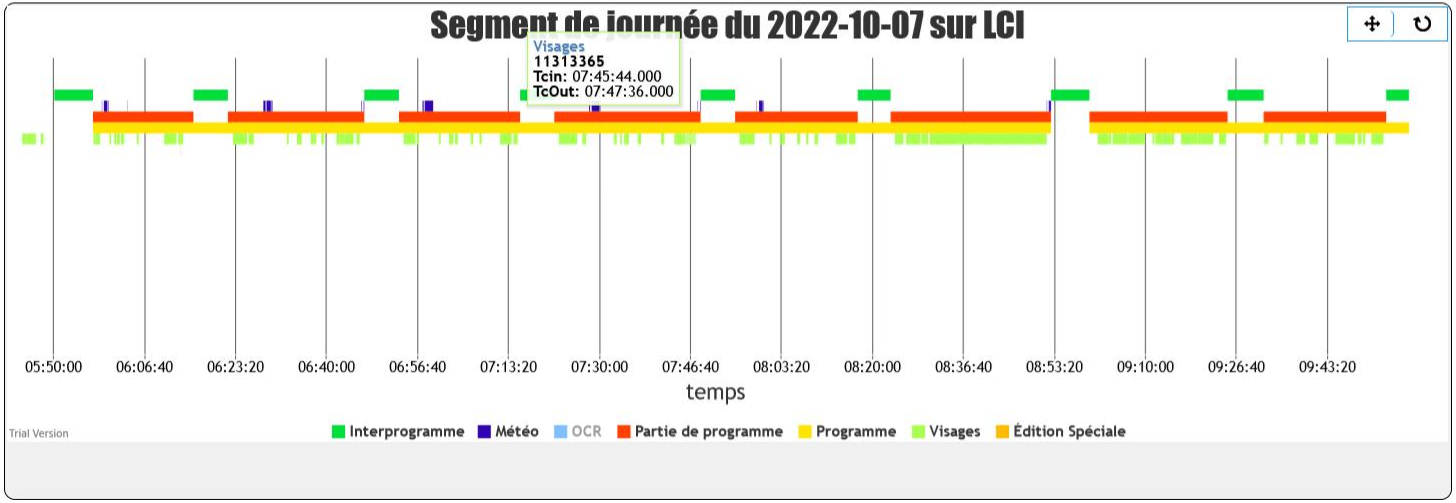
Exporter la journée



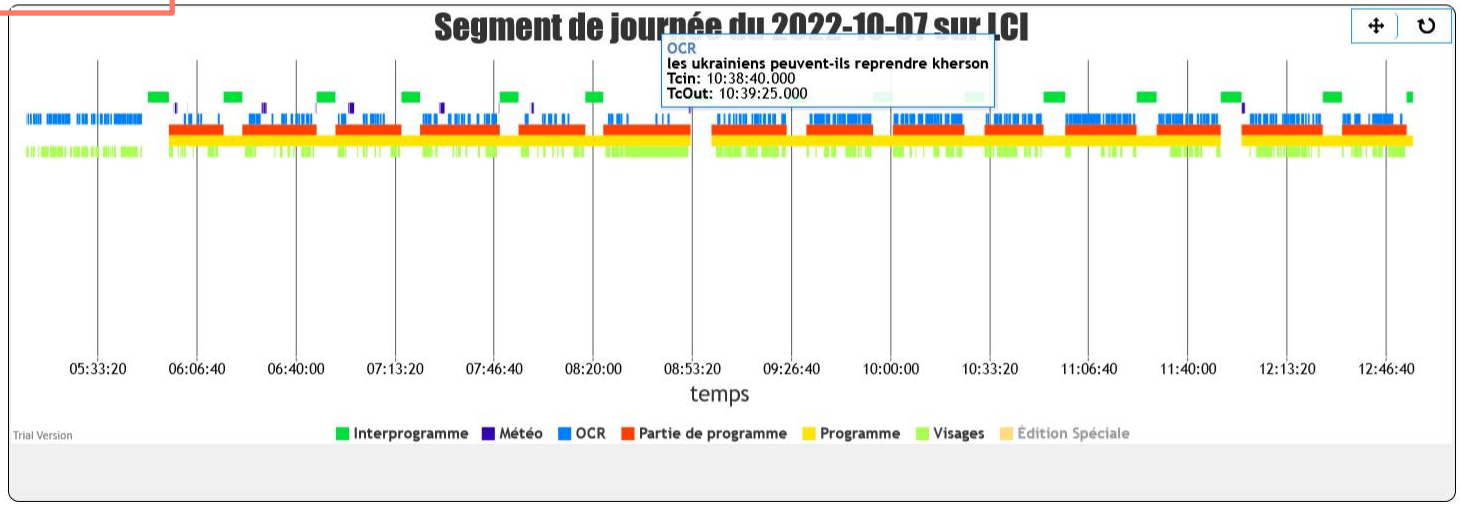
Exporter la journée



Exporter la journée



Exporter la journée



03.

Petite cartographie des IA descriptives pour l'INA

Les principaux outils d'IA descriptive

- Transcription de la parole
- Extraction d'entités nommées
- Reconnaissance des genres et locuteurs
- OCR des textes incrustés
- Découpage en programmes

- Reconnaissance des logos
- Reconnaissance des visages
- Reconnaissance d'objets - Description d'image
- Découpage en séquences

IAG — Résumés automatiques, thématisation, vectorisation…

Moteur de recherche sémantique, IHM

Les principaux outils d'IA descriptive

Des données qui ne pourraient pas être produites à cette échelle par des humains

30 millions d'heures à transcrire

X env. 6 H => 180 millions d'heures de travail

100 000 années-humains !

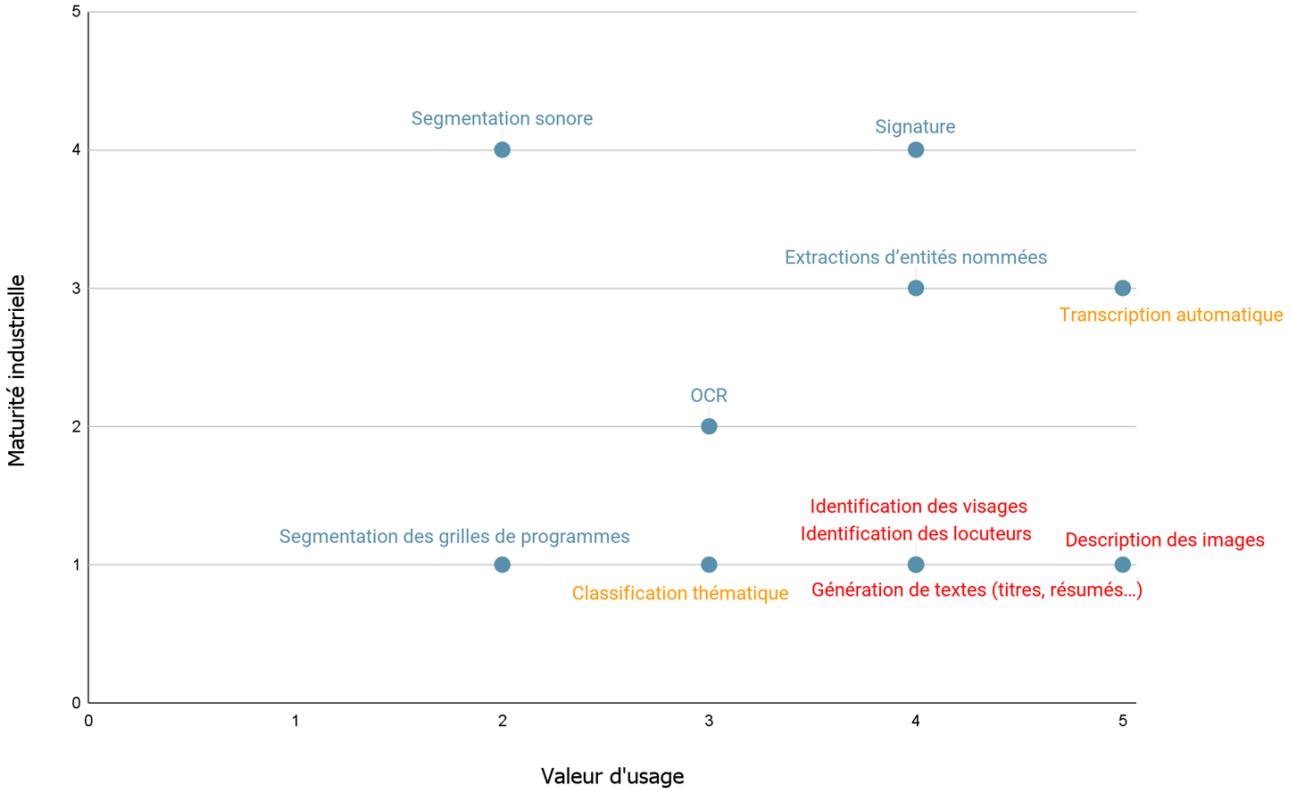
objectif : 1 heure de vidéos => env. 3 secondes de calcul IA

3 années (calcul 24/7)

Des degrés de maturité hétérogènes

R&D — Innovation — Industrialisation

Les principaux outils d'IA descriptive



Interprétation

- **Lecture**
  - Abscisse : plus la valeur d'usage d'un traitement IA est grande pour l'INA, plus le point est situé à droite
  - Ordonnée : plus la maturité d'un traitement IA est grande à l'INA, plus le point est haut

- **Enjeux et priorisation**
  - Les traitements en rouge (description des images, identification des visages et des locuteurs, génération de texte) sont ceux pour lesquels il faut prioriser la maturité industrielle au regard du bénéfice attendu.
  - Les traitements en orange doivent également gagner en maturité industrielle pour produire leur potentielle valeur d'usage.

Les principaux outils d'IA descriptive

Des défis à relever pour industrialiser

- Fiabilisation
  - vérité terrain, mesure de fiabilité, entraînement, fine-tuning
  - contrôle et supervision des traitements
- Adaptation du SI
  - à l'évolution continue des IA et des données produites (versionning)
  - à l'exigence d'interopérabilité avec les référentiels externes
  - pour passer à l'échelle (robustesse des processus, lac de données, moteurs de recherche…)

Des IHM à réinventer

- **outils de fiabilisation** (vérification, annotation, validation, mesures)
- **outils de datavisualisation** pour représenter de grandes masses de données, découvrir et suivre des tendances selon des critères multiples
- **outils d'exploration et d'annotation** (représentations temporelles interactives à différentes échelles du "macro" au "micro" pour naviguer dans les vidéos, les annoter)
- **outils de recherche sémantique** en langage naturel

04.

→ Une première application : data.ina.fr

Une première application : data.ina.fr

Ouverture du site
3 oct. 2024

Un regard inédit de l'INA sur les médias
grâce aux intelligences artificielles

Financé par
l'Union européenne
NextGenerationEU

Une première application : data.ina.fr

3 IA utilisées

Transcription de la parole

Whisper (openAI)
outil externe

Extraction d'entités nommées

TextRazor
outil externe

Reconnaissance des genres

InaSpeechSegmenter
outil INA (David Doukhan)
en collaboration avec le Laboratoire d'Informatique de l'Université du Mans (LIUM)

Une première application : data.ina.fr

PÉRIMÈTRE (Transcription, entités nommées)

De 2019 à 2024…

+ 200K heures

Puis reprise de l'antériorité, et mise à jour du flux entrant

Journaux télévisés du soir

TF1 | France 2 | France 3 | Arte | M6

Chaînes d'information en continu (6H - 24H)

BFM TV | CNews | France Info : | LCI

Tranches matinales de radio (5H-10H semaine | 7H-9H week-end)

France Culture | France Info | France Inter | RMC | RTL | Europe 1 | Sud Radio

Une première application : data.ina.fr

PÉRIMÈTRE (Segmentation femmes/hommes)

De 2019 à 2024…

+ 500K heures

Puis reprise de l'antériorité, et mise à jour du flux entrant

8 chaînes de télévision (10H-24H)

TF1 | France 2 | France 3 | France 5 | Canal + | M6 | Arte | TV5Monde

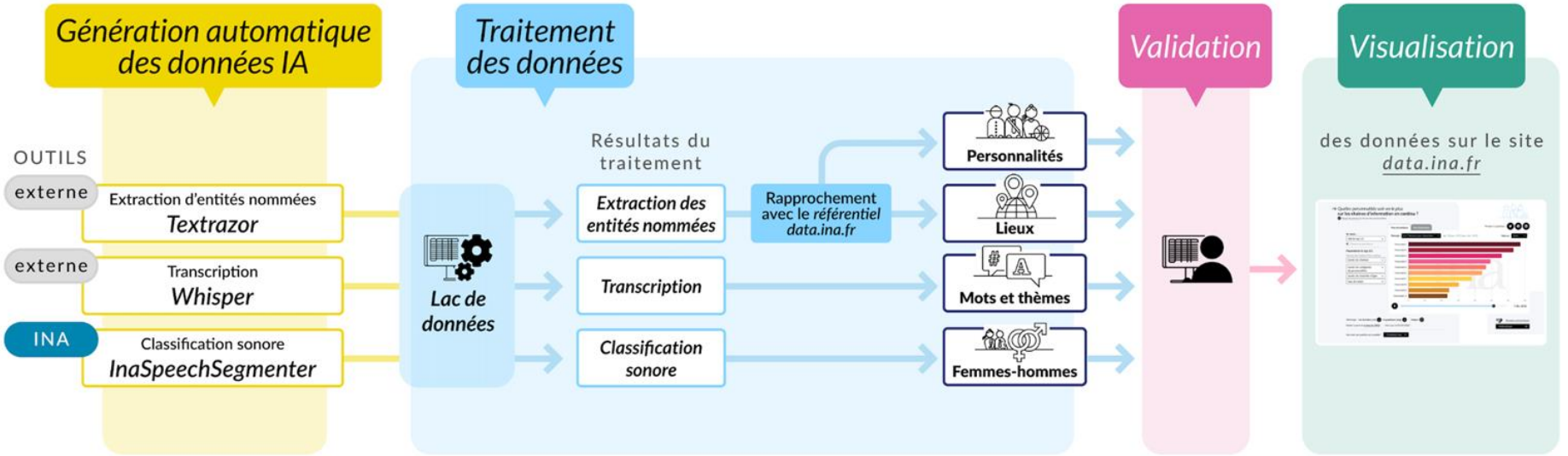
5 chaînes d'information en continu (6H - 24H)

BFM TV | CNews | LCI | France Info : | France 24

5 chaînes de radio (10H-24H)

France Culture | France Info | France Inter | RMC | RTL

Les coulisses de data.ina.fr : schéma général

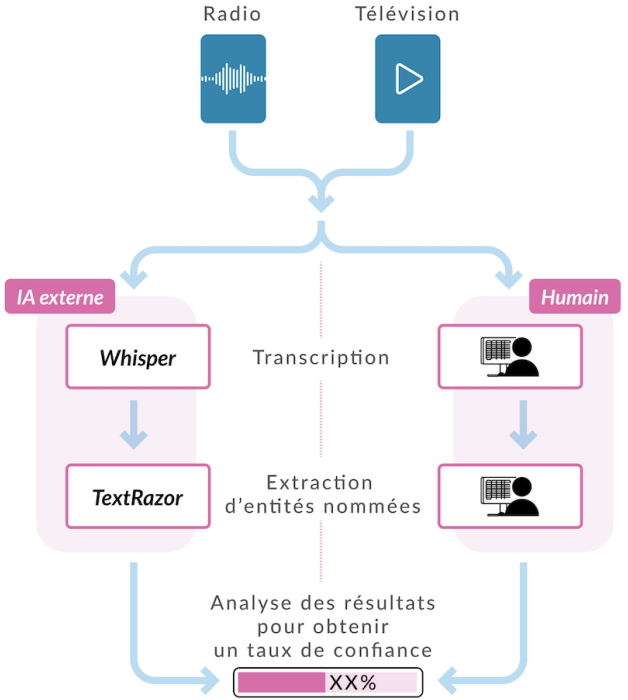


Les coulisses de data.ina.fr

3 grandes actions de contrôle

Vérité de terrain sur les IA externes

L'objectif est de comparer le traitement IA d'heures de télévision et radio (sélectionnées sur le périmètre data.ina.fr) et le travail humain de transcription et de reconnaissance des entités nommées

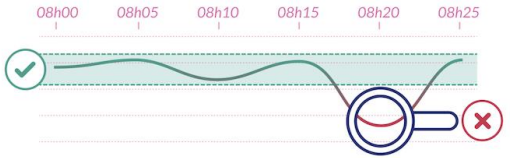


3 grandes actions de contrôle

Contrôle des traitements automatiques

S'assurer que les flux de traitements sont opérationnels

S'assurer de la complétude des segments traités



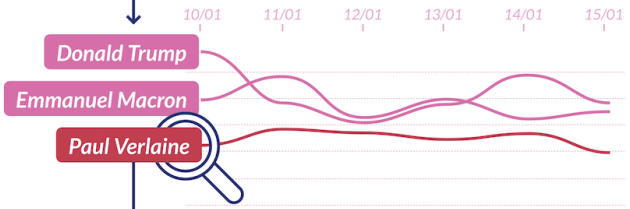
Détecter des incohérences quantitatives

Les coulisses de data.ina.fr

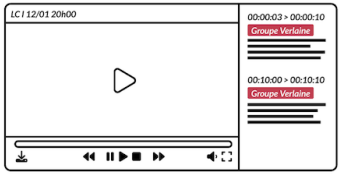
3 grandes actions de contrôle

Contrôle de la pertinence

S'assurer que les résultats des entités nommées présentées sur le site ne présentent pas de biais



Investigation à l'aide de Player Expert



Ecriture de textes d'alerte pour les internautes dans le cas de biais avérés

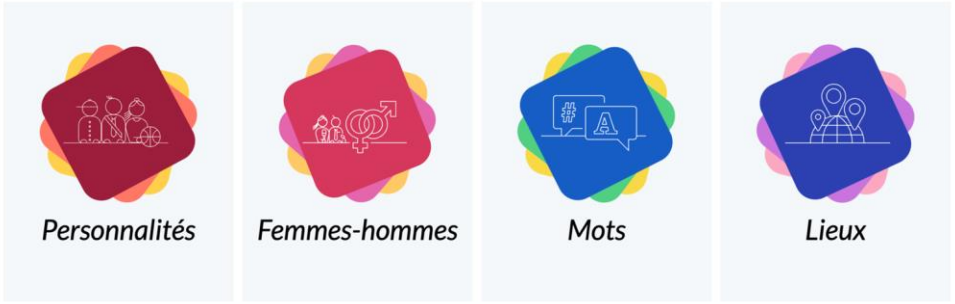
Exemple

Paul Verlaine !

À la suite de notre investigation, les résultats mentionnant Paul Verlaine tendent à indiquer que l'algorithme de reconnaissance des entités nommées ***confond le poète avec l'entreprise "Groupe Verlaine", présente dans les médias à travers une campagne de publicité.*** *Cela entraîne une sur-représentativité. L'amélioration continue des algorithmes, avec une meilleure…*

Le site data.ina.fr

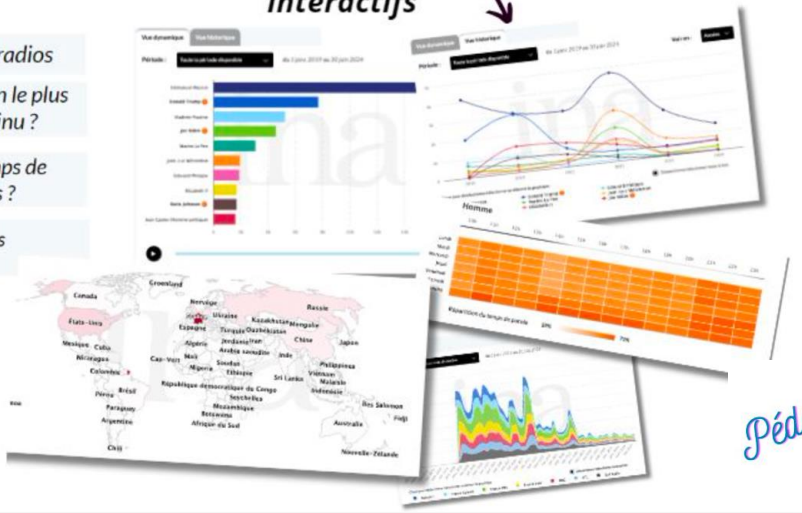
4 clés de lecture



16 questions grand public

- → Explorez les mots prononcés sur les radios
- → Quelles personnalités mentionne-t-on le plus sur les chaînes d'information en continu ?
- → Comment évolue la répartition du temps de parole entre les femmes et les hommes ?
- → Quels pays mentionne-t-on le plus dans les journaux télévisés ?

28 graphiques interactifs



→ Démo : data.ina.fr

Pédagogie TRANSPARENCE ACCESSIBILITÉ personnalisation

data.ina.fr

POUR QUI ?

- Spécialistes des médias, journalistes, experts, décideurs ou citoyens curieux : **l'ergonomie du site a été pensé pour le plus grand nombre.**
- Il permet de **représenter visuellement de grandes masses de données**, de découvrir, objectiver, suivre des **tendances** : vocation heuristique.
- Pour les chercheurs désireux d'approfondir des hypothèses, **il ne remplace pas l'enquête et l'analyse poussée** (avec accès aux sources audiovisuelles et aux outils d'analyse) à l'INA thèque et au Lab.

data.ina.fr

L'ÉDITORIALISATION

Un travail éditorial régulier de data journalisme : des enquêtes et analyses sur des angles précis sont publiées sur le site de l'INA **La Revue des médias**

Une première application : data.ina.fr

BILAN INTERNE

Un projet accélérateur de l'intégration de l'IA à l'INA

- Première exploitation de l'IA à grande échelle qui concrétise les apports et opportunités de l'IA pour accroître nos capacités de description des fonds et leur exploitation
- Prise de conscience
  - des biais, des enjeux de fiabilisation
  - des défis techniques et organisationnels de l'industrialisation

Annexes

→ Captures d’écran du site data.ina.fr

Le site data.ina.fr

→ *Quelles personnalités mentionne-t-on le plus dans les journaux télévisés ?*

Nombre de mentions verbales des personnalités (hors journalistes) détectées par l'IA dans les journaux télévisés du soir de 5 chaînes de télévision.

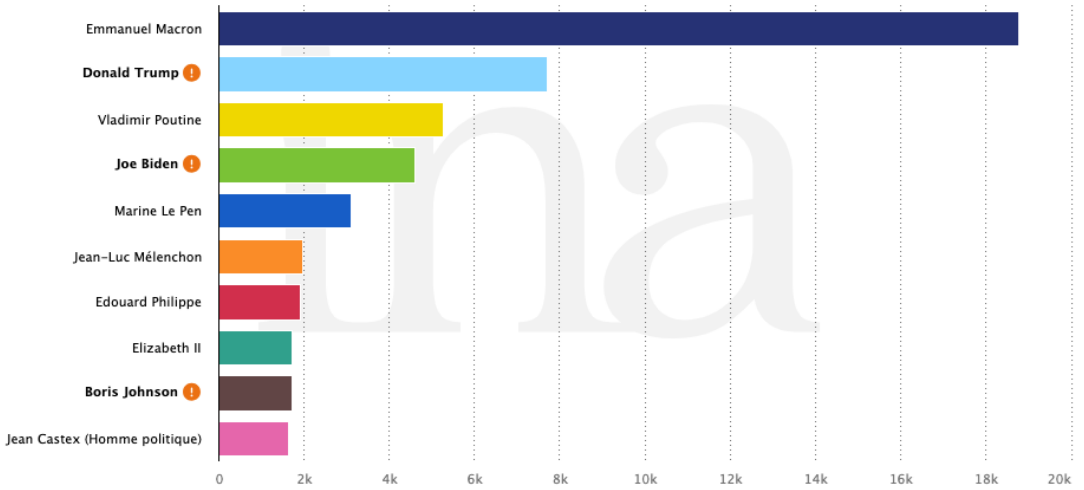
Vue dynamique | Vue historique

Je veux :

- *Voir le top 10*
- *Toutes les chaînes*
- *Tous les genres*

Période : Toute la période disponible — du *1 janv. 2019* au *30 juin 2024*

Voir en : Mois



juin 2024

Le site data.ina.fr

→ *Quelles personnalités mentionne-t-on le plus dans les journaux télévisés ?*

Nombre de mentions verbales des personnalités (hors journalistes) détectées par l'IA dans les journaux télévisés du soir de 5 chaînes de télévision.

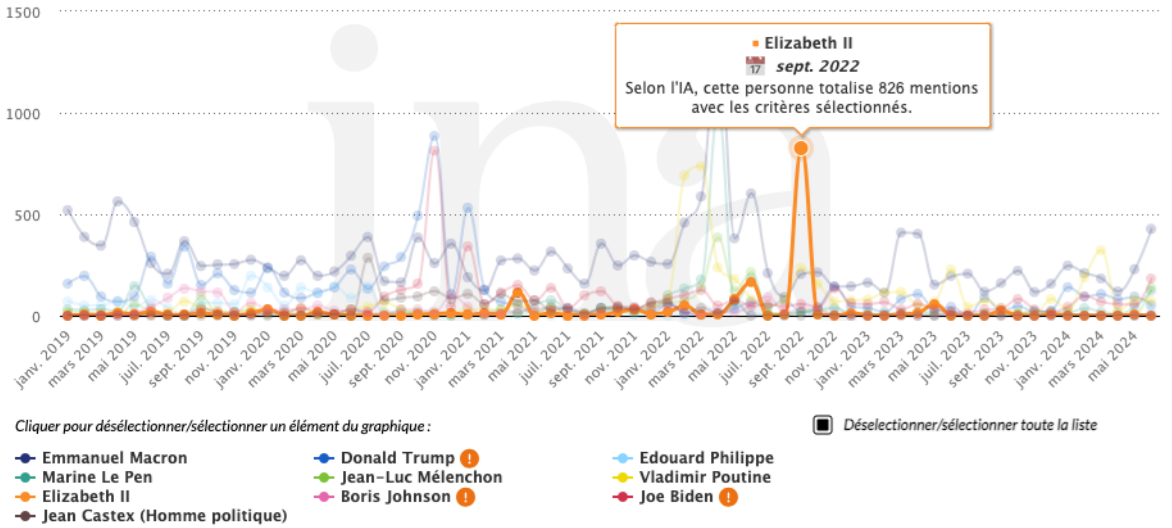
Vue dynamique | Vue historique

Je veux :

- *Voir le top 10*
- *Toutes les chaînes*
- *Tous les genres*

Période : Toute la période disponible — du *1 janv. 2019* au *30 juin 2024*

Voir en : Mois



Le site data.ina.fr

→ *Quelles personnalités mentionne-t-on le plus dans les journaux télévisés ?*

Nombre de mentions verbales des personnalités (hors journalistes) détectées par l'IA dans les journaux télévisés du soir de 5 chaînes de télévision.

Vue dynamique | Vue historique

Je veux :

Faire ma recherche

Rechercher et ajouter des noms

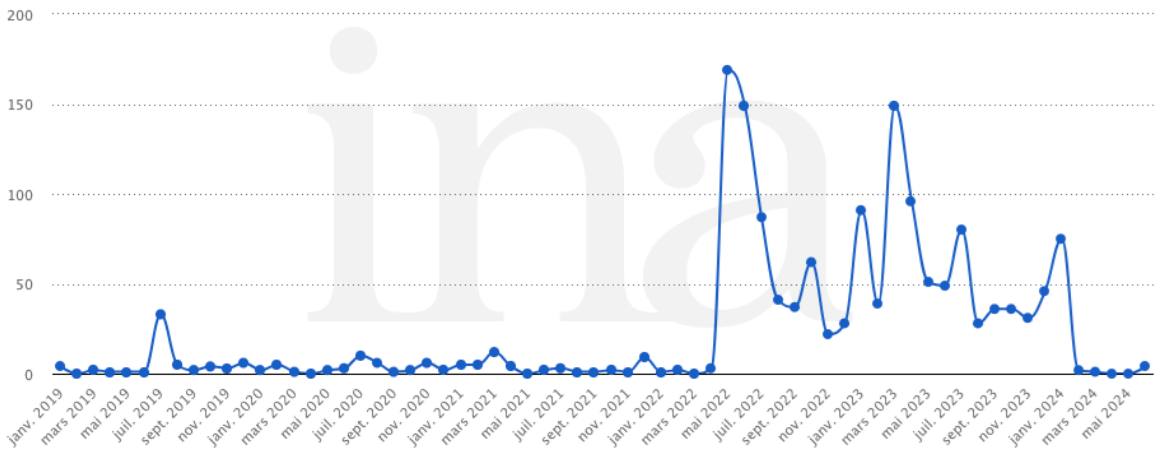
Elisabeth Borne ×

Toutes les chaînes

Tous les genres

Période : Toute la période disponible — du *1 janv. 2019* au *30 juin 2024*

Voir en : Mois



Cliquer pour désélectionner/sélectionner un élément du graphique :

Déselectionner/sélectionner toute la liste

Elisabeth Borne

Le site data.ina.fr

→ *Quelles personnalités mentionne-t-on le plus dans les journaux télévisés ?*

Nombre de mentions verbales des personnalités (hors journalistes) détectées par l'IA dans les journaux télévisés du soir de 5 chaînes de télévision.

Vue dynamique | Vue historique

Je veux :

Faire ma recherche

Rechercher et ajouter des noms

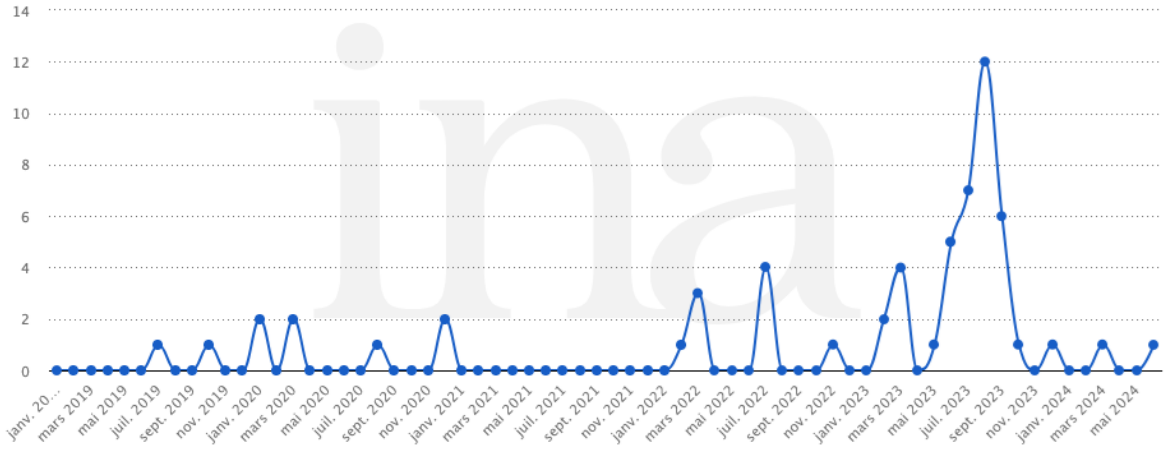
Richard Wagner (co...

Toutes les chaînes

Tous les genres

Période : Toute la période disponible — du *1 janv. 2019* au *30 juin 2024*

Voir en : Mois



Cliquer pour désélectionner/sélectionner un élément du graphique :

Richard Wagner (compositeur)

Déselectionner/sélectionner toute la liste

Le site data.ina.fr

→ *Quelles personnalités mentionne-t-on le plus dans les journaux télévisés ?*

Nombre de mentions verbales des personnalités (hors journalistes) détectées par l'IA dans les journaux télévisés du soir de 5 chaînes de télévision.

Vue dynamique | Vue historique

Je veux :

Faire ma recherche

Rechercher et ajouter des noms

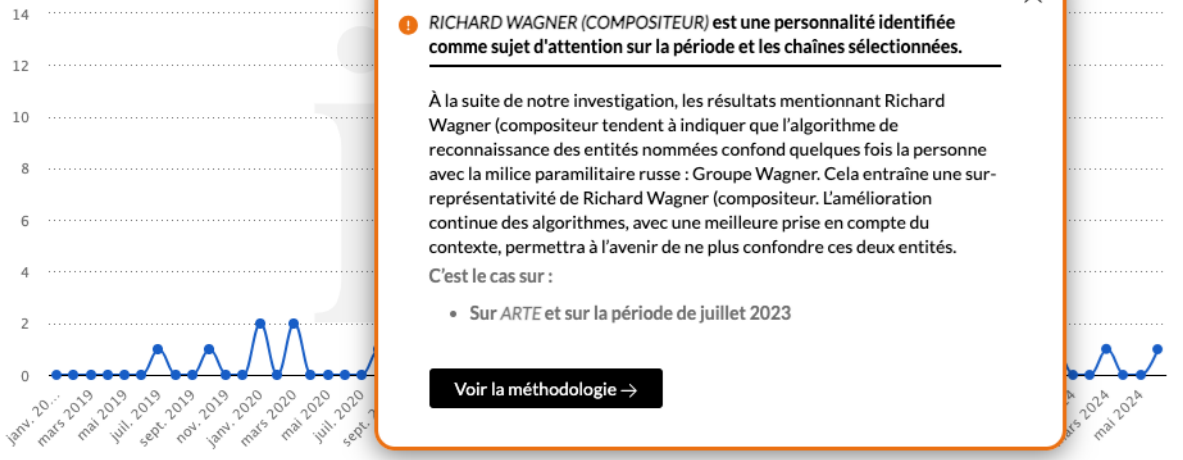
Richard Wagner (co...

Toutes les chaînes

Tous les genres

Période : Toute la période disponible — du *1 janv. 2019* au *30 juin 2024*

Voir en : Mois



RICHARD WAGNER (COMPOSITEUR) **est une personnalité identifiée comme sujet d'attention sur la période et les chaînes sélectionnées.**

À la suite de notre investigation, les résultats mentionnant Richard Wagner (compositeur tendent à indiquer que l'algorithme de reconnaissance des entités nommées confond quelques fois la personne avec la milice paramilitaire russe : Groupe Wagner. Cela entraîne une sur-représentativité de Richard Wagner (compositeur. L'amélioration continue des algorithmes, avec une meilleure prise en compte du contexte, permettra à l'avenir de ne plus confondre ces deux entités.

C'est le cas sur :

- Sur *ARTE* et sur la période de juillet 2023

Voir la méthodologie →

Cliquer pour désélectionner/sélectionner un élément du graphique :

Richard Wagner (compositeur)

Déselectionner/sélectionner toute la liste

Le site data.ina.fr

DONNÉES PROVISOIRES

→ *Comment évolue la répartition du temps de parole entre les femmes et les hommes ?*

Cinq catégories de sons (voix de femmes, voix d'hommes, musique, bruit et silence) détectées par l'IA dans le flux sonore de 18 chaînes de télévision et de radio.

Vue dynamique | Vue historique agrégée

Je veux :

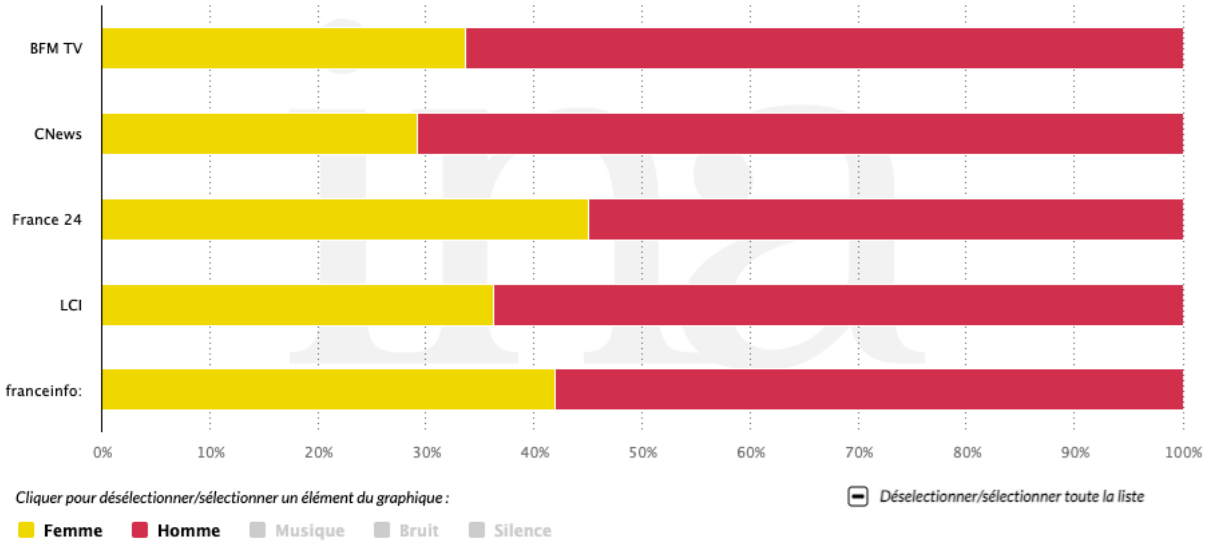
Télévision d'information en continu

Toutes les chaînes

Sur une base 100%

Période : Toute la période disponible — du *1 janv. 2019* au *30 juin 2024*

Voir en : Années



Cliquer pour désélectionner/sélectionner un élément du graphique :

Déselectionner/sélectionner toute la liste

Femme **Homme** Musique Bruit Silence

Le site data.ina.fr

→ *Explorez les mots prononcés dans les journaux télévisés*

Nombre de tours de parole dans lesquels le mot recherché a été détecté au moins une fois par l'IA dans les journaux télévisés du soir de 5 chaînes de télévision.

Je veux :

Choisissez un mot

covid

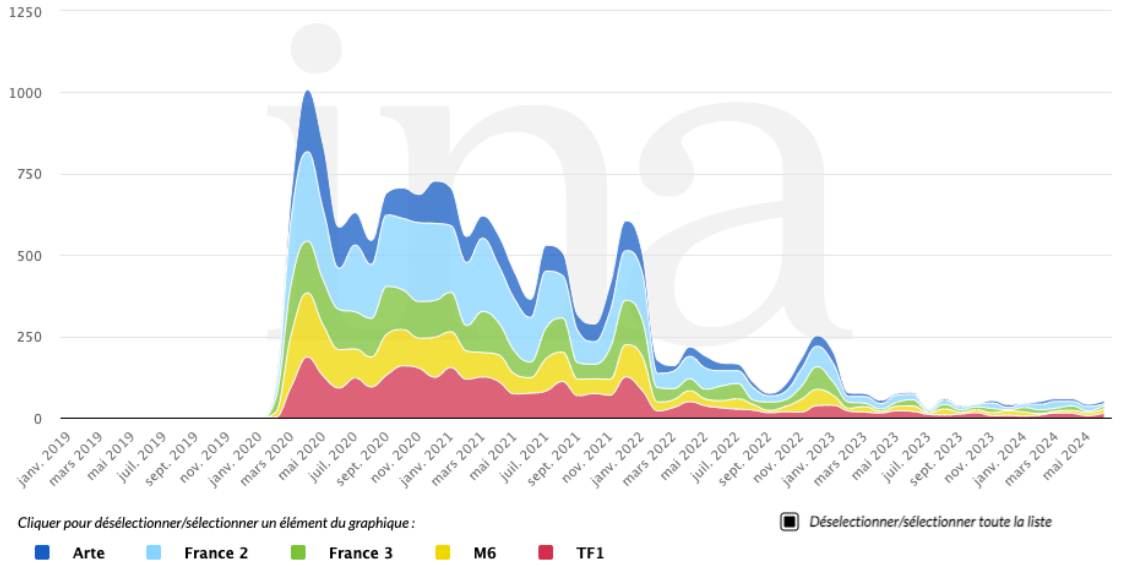
Lorsque vous indiquez 2 mots, notre moteur va additionner les tours de parole avec le mot 1, avec le mot 2, et avec les mots 1 et 2 ensemble. Exemple : pour République française, il additionne les tours de parole avec République, avec française et ceux avec République française.

Lorsque vous cherchez une expression particulière, mettez des guillemets. Ainsi, le moteur comptabilisera uniquement les tours de parole avec les deux mots associés. Par exemple : "réchauffement climatique", "gilets jaunes", "République française"…

Vue historique

Période : Toute la période disponible — du *1 janv. 2019* au *30 juin 2024*

Voir en : Mois



Le site data.ina.fr

→ *Explorez les mots prononcés dans les journaux télévisés*

Nombre de tours de parole dans lesquels le mot recherché a été détecté au moins une fois par l'IA dans les journaux télévisés du soir de 5 chaînes de télévision.

Je veux :

Choisissez un mot

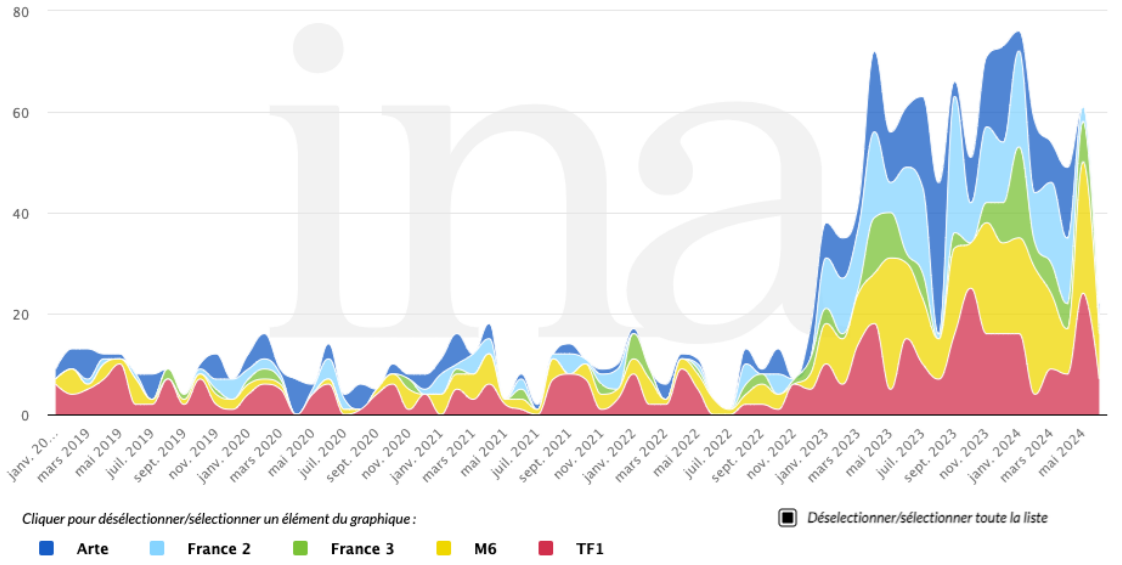
Lorsque vous indiquez 2 mots, notre moteur va additionner les tours de parole avec le mot 1, avec le mot 2, et avec les mots 1 et 2 ensemble. Exemple : pour République française, il additionne les tours de parole avec République, avec française et ceux avec République française.

Lorsque vous cherchez une expression particulière, mettez des guillemets. Ainsi, le moteur comptabilisera uniquement les tours de parole avec les deux mots associés. Par exemple : "réchauffement climatique", "gilets jaunes", "République française"…

Vue historique

Période : Toute la période disponible — du *1 janv. 2019* au *30 juin 2024*

Voir en : Mois



Le site data.ina.fr

→ *Explorez les mots prononcés dans les journaux télévisés*

Nombre de tours de parole dans lesquels le mot recherché a été détecté au moins une fois par l'IA dans les journaux télévisés du soir de 5 chaînes de télévision.

Je veux :

Choisissez un mot

"intelligence artificielle"

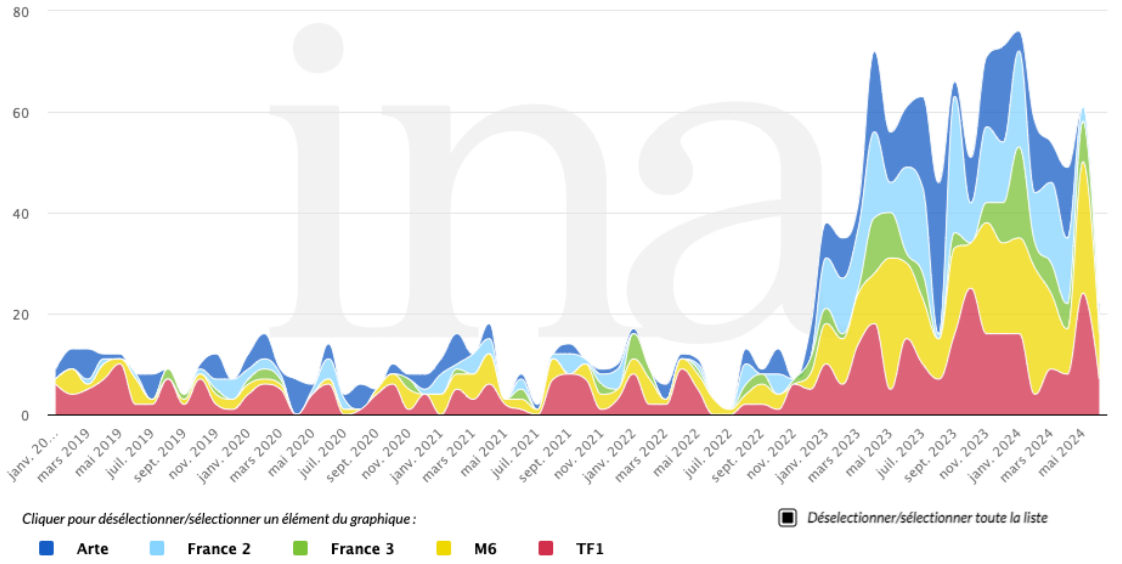
Lorsque vous indiquez 2 mots, notre moteur va additionner les tours de parole avec le mot 1, avec le mot 2, et avec les mots 1 et 2 ensemble. Exemple : pour République française, il additionne les tours de parole avec République, avec française et ceux avec République française.

Lorsque vous cherchez une expression particulière, mettez des guillemets. Ainsi, le moteur comptabilisera uniquement les tours de parole avec les deux mots associés. Par exemple : "réchauffement climatique", "gilets jaunes", "République française"…

Vue historique

Période : Toute la période disponible — du *1 janv. 2019* au *30 juin 2024*

Voir en : Mois



Le site data.ina.fr

→ *Quels lieux mentionne-t-on le plus dans les journaux télévisés ?*

Nombre de mentions verbales des lieux (géographie administrative) détectées par l'IA dans les journaux télévisés du midi et du soir de 4 chaînes de télévision.

Vue cartographique | Vue dynamique | Vue historique

Toutes les chaînes

Exclure la France du classement

Période : Toute la période disponible — du *1 janv. 2022* au *31 déc. 2023*

Intensité de la présence - +

Le site data.ina.fr

→ *Quels lieux mentionne-t-on le plus dans les journaux télévisés ?*

Nombre de mentions verbales des lieux (géographie administrative) détectées par l'IA dans les journaux télévisés du midi et du soir de 4 chaînes de télévision.

Vue cartographique | Vue dynamique | **Vue historique**

Je veux :

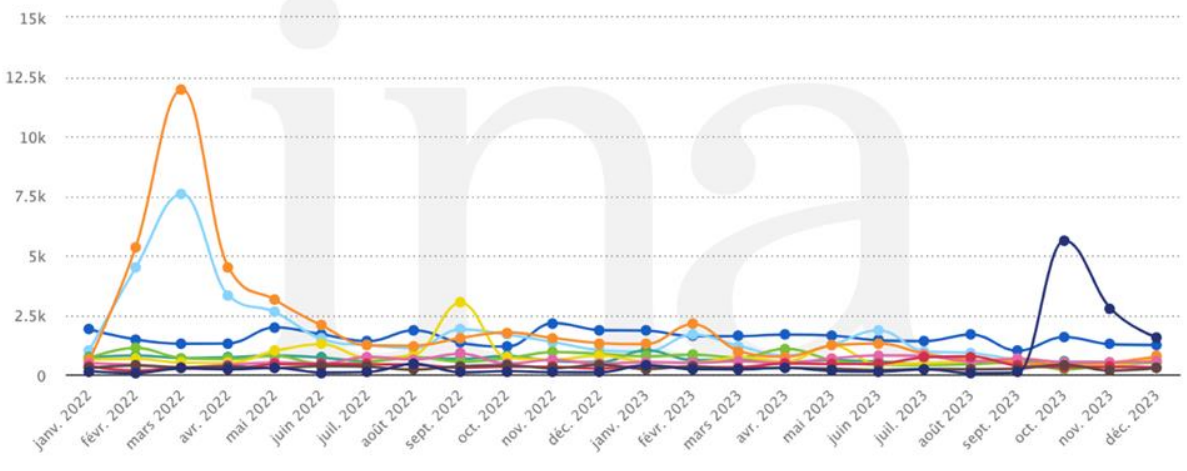
Voir le top 10

Toutes les chaînes

Exclure la France du classement

Période : Toute la période disponible — du *1 janv. 2022* au *31 déc. 2023*

Voir en : Mois



Cliquer pour désélectionner/sélectionner un élément du graphique :

Déselectionner/sélectionner toute la liste

Etats-Unis | Russie | Allemagne | Chine | Royaume-Uni | Ukraine
Italie | Espagne | Belgique | Israël

Merci
de votre *attention*