

Mezanno

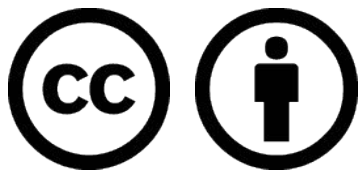
Des sources sérielles aux données structurées pour les humanités numériques

Journée “Reconnaissance automatique de caractères et extraction d’information spatiale
à partir de sources anciennes numérisées”
ANR OCEANLINKS, Campus Condorcet, 10 décembre 2025

Nathalie Abadie, Marie Carlin, Edwin Carlinet, Joseph Chazalon, Pascal Cristofoli,
Bertrand Duménieu, Jean-Philippe Moreux, Julien Perret, Jonathan Perrinet, Marie Puren

(Ré)utiliser cette présentation

Licence **Creative Common BY 4.0**



N'hésitez pas à **réutiliser** ces contenus et nous **citer**.

*Sauf citation ou mention contraire,
nous sommes les auteurs des contenus.*

Consultez le diaporama pendant et après la présentation

<https://hnurl.fr/wpgxkw>



Mezanno
est un des prolongements
de l'ANR SoDUCo

Projet ANR **SoDUCo**

(Social Dynamics in Urban Context)

Open tools, models, and data - Paris and its suburbs, 1789-1950

<https://soduco.geohistoricaldata.org/>

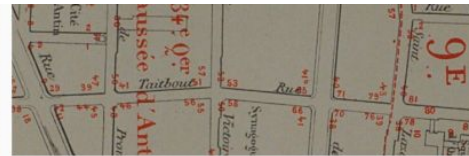
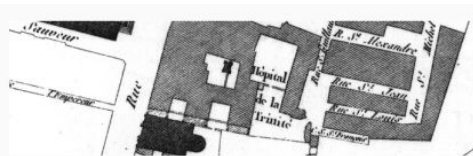
<https://anr.fr/Projet-ANR-18-CE38-0013>

2019-2023



Sources primaires

Source cartographiques



Annuaire professionnels

GRAVEURS EN GÉOGRAPHIE ET TOPOGRAPHIE.

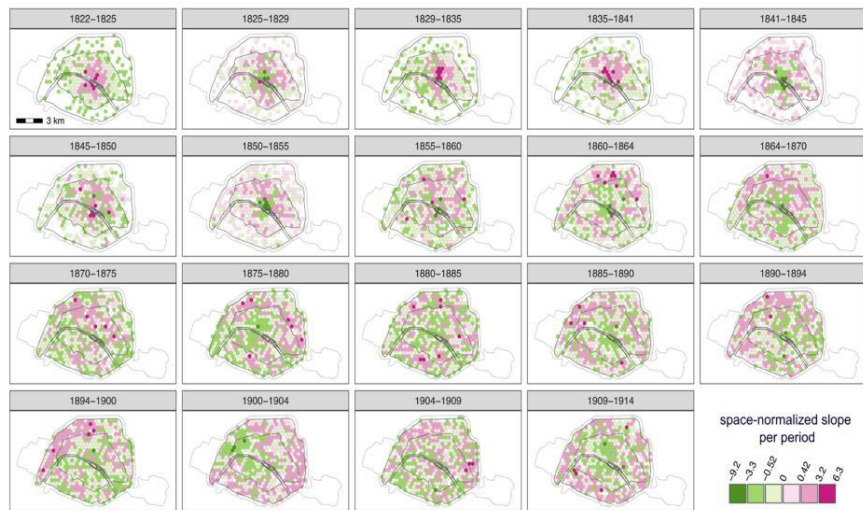
Arnault, *Serpente* 16.
Berthe, cit. 1827, *Noyers* 46.

Popincourt 150.
Dehaeck, *Charenton* 166.
Delaigle, p. *St-Germain-des-Prés* 6.
Delangle, *Clotaire* 3.
Delaporte, *Cassette* 12.

MERCERIES, SOIERIES ET NOUVEAUTÉS (MARCHANDS DE).

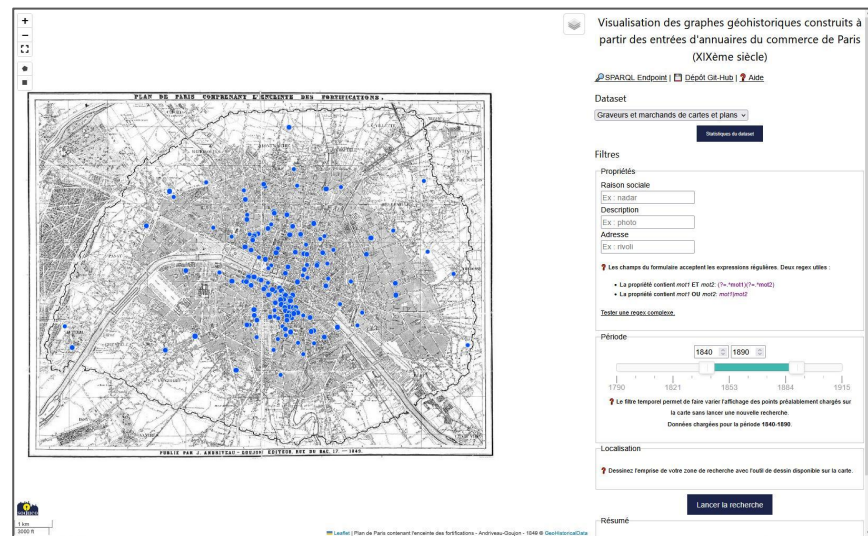
Achard, passage du Panorama, 61.	Belin, r. Saint-Martin, 125.
Alavoine, de soieries et nouveautés pour chaussure, rue de la Poterie - des- Halles, 22.	Bellard et Lhuillier (Mesd.), nouveautés, r. de la Paix, 22.

Exemples de résultats



J. Gravier, S. Baciocchi, P. Cristofoli, B. Duménieu, E. Carlinet, J. Chazalon, N. Abadie, S. Tual, J. Perret. **Evaluating and Understanding the Geocoding of City Directories of Paris (1787-1914): Data-Driven Geography of Urban Sprawl and Densification.** *Digital Humanities Quarterly* (2025). À paraître.

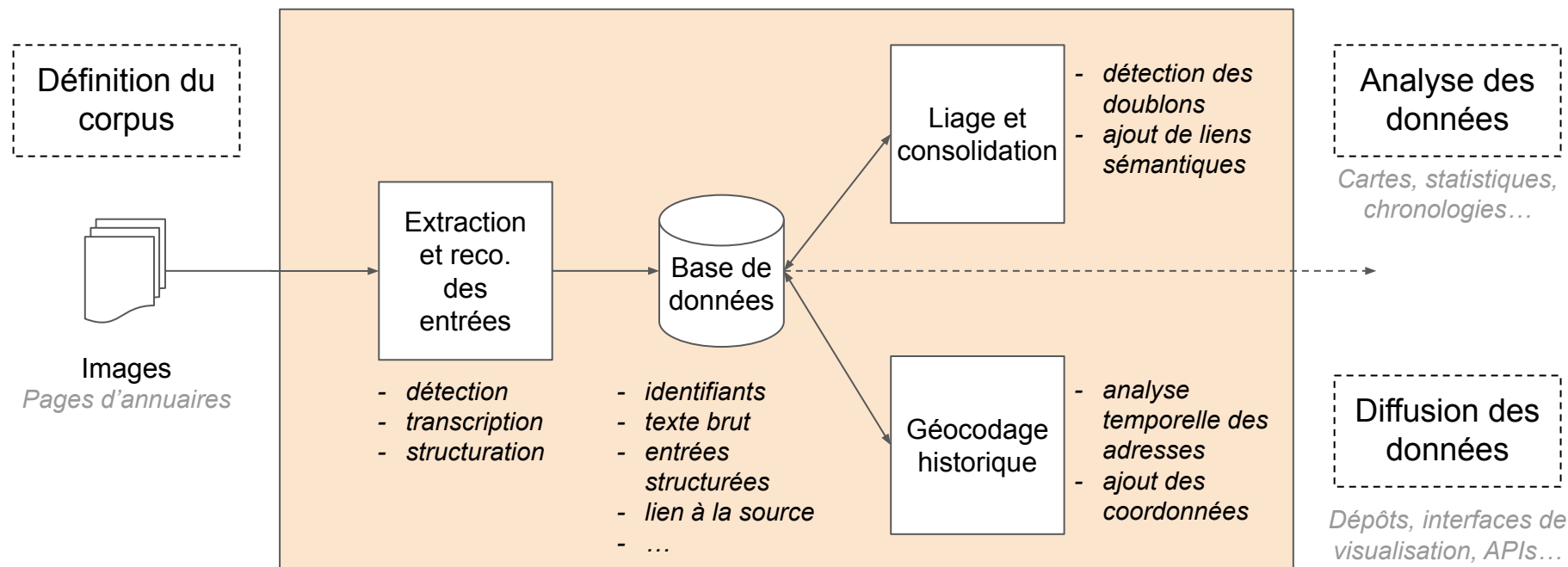
J. Gravier, M. Barthelemy. **A typology of activities over a century of urban growth.** *Nature Cities* 1, 567–575 (2024). doi: [10.1038/s44284-024-00108-7](https://doi.org/10.1038/s44284-024-00108-7)



S. Tual, N. Abadie, B. Duménieu, J. Chazalon, E. Carlinet. **Création d'un graphe de connaissances géohistorique à partir d'annuaires du commerce parisien du XXe siècle : application aux métiers de la photographie.** *Revue Ouverte d'Intelligence Artificielle, Post-actes de la conférence Ingénierie des Connaissances (IC 2021-2022-2023)*, Volume 6 (2025) no. 1-2, pp. 179-200. doi: [10.5802/roia.98](https://doi.org/10.5802/roia.98)

Vue d'ensemble de la chaîne d'extraction de données

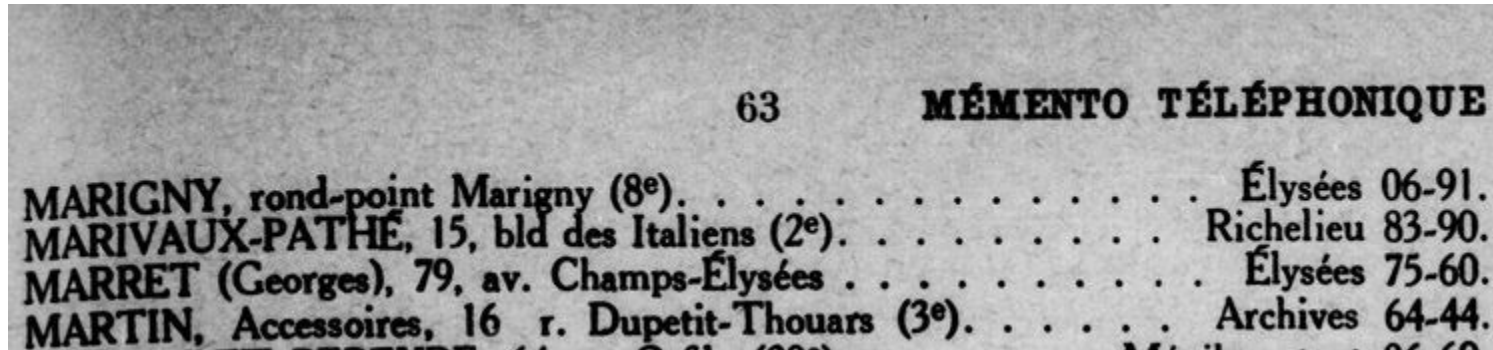
*“De l’**image** d’annuaire du commerce ancien au **graphe interrogeable** de **données structurées et géolocalisées**”*



Des images aux données structurées

Valoriser des fonds sériels

Annuaire, index, dictionnaires, registres, tables, minutes, comptes-rendus...



63	MÉMENTO TÉLÉPHONIQUE
MARIGNY, rond-point Marigny (8 ^e).	Élysées 06-91.
MARIVAUX-PATHÉ, 15, bld des Italiens (2 ^e).	Richelieu 83-90.
MARRET (Georges), 79, av. Champs-Élysées	Élysées 75-60.
MARTIN, Accessoires, 16 r. Dupetit-Thouars (3 ^e).	Archives 64-44.

Le Tout-cinéma : annuaire général illustré du monde cinématographique / dir. E. L. Fouquet ; dir. Cl. Guilhamou, 1934, source : Bibliothèque nationale de France, ark:/12148/bd6t543045578

Valoriser des fonds sériels

Annuaire, index, dictionnaires, registres, tables, minutes, comptes-rendus...

102

ALP. — ANG

ALPES (Hautes). — Emprunt de 113.000 francs à la Caisse des chemins vicinaux. V. loi 14. — Emprunt de 73.000 francs à la Caisse des chemins vicinaux. V. loi 196.

ALPES-MARITIMES. — Emprunt à la Caisse des chemins vicinaux. V. loi 137. — Emprunt de 240.000 francs. V. loi 235.

ANCEL. Seine-Inférieure. — Fait une rectification au procès-verbal [4 août] (A., t. II, p. 915; J.O., c.r., p. 915). — Membre d'une Commission (F. 45; J.O., p. 2693).

Table analytique des compte rendus des séances du Sénat, session ordinaire de 1890, source: Sénat, https://www.senat.fr/fileadmin/Histoire/Textes_debats_et_grandes_lois/Les_travaux_du_Senat_de_la_IIIe_Republique/tables/TAB_1890.pdf

Valoriser des fonds sériels

Annuaire, index, dictionnaires, registres, tables, minutes, comptes-rendus...

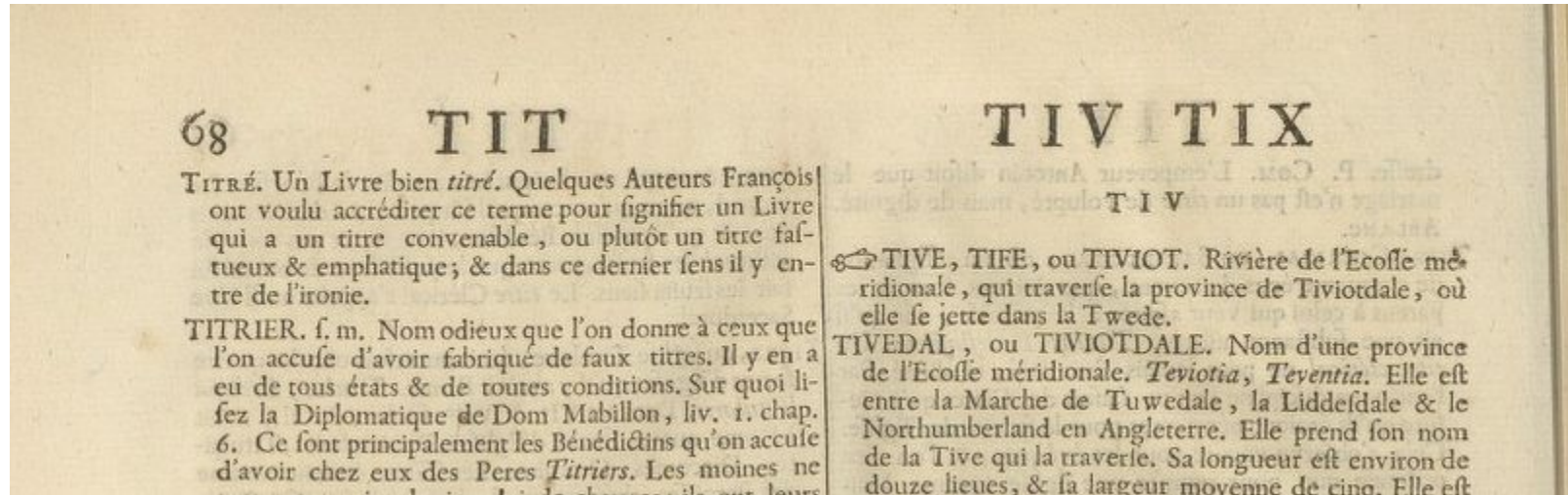
Ababsa Mohamed, Algérie, 2° cl., 14° R.T.A. 151.
Ababsa Sibti, Algérie, 2° cl., 9° Sp. 153.
Abassi Ali, Algérie, 2° cl., 11° R.T.A. 151.
Abbes Slimane, Algérie, cap., 27° R.T.T. 151.
Abbon ben Aissa, Algérie, 2° cl., 66° R.A.D. 151.
Abdallah ben Amar, Tunisie, cap., 8° R.T.T. 151.
Abd-E-Lem ben Hammed, 1915, Maroc, 1° R. Mar. St. XIII A.
Abdellah ben Hamadi, Tunisie, 2° cl., 44° R.P. 151.
Abdellah ben Salah, Tunisie, 2° cl., 44° B.P. 151.
Abdel Malek Ahmed, Algérie, 2° cl., 2° Sp. 151.
Abderrhamane Ahmed, 1915, Maroc, 2° cl., 4° R. Mar. St. VIII C.

Adam (Auguste), 2-11-02, Plouguiel, 2° cl., 50° R.A.D. St. XIII A.
Adam (François), 11-11-13, Grosblierstroff, 2° cl., 13° R.I. St. XIII A.
Adam (Lucien), 10-4-18, Le Puy, cap., 26° R.I. St. VIII C.
Adam (Raymond), 17-9-03, Mesnil-Panneville, 2° cl., 39° R.I. 135.
Adam (Théodore), 13-6-06, Ammerzwiler, 2° cl., 189° R.A. St. XVII A.
Adama Sire, Mauritanie, 2° cl., 623° R.P.S. 151.
Adamovic (Joseph), 10-2-06, Bosaca (Slov.), 2° cl., R. Tch. 151.
Adie (Victor), 9-8-03, Concoret, 2° cl., 41° R.I. St. XVII A.
Adilly (Raymond), 13-8-08, Paris, 2° cl., 29° R.A. St. XIII A.
Adjadj (Sauveur), 17-12-17, Aumale (Alg.), 2° cl., 92° R.I. St. XIII A.
Adolphe (Amédée), 14-10-18, Beaumont-s.-Lèze, 2° cl., 96° R.I.A. St. IX C.

Liste officielle [...] des prisonniers de guerre français [...], Centre national d'information sur les prisonniers de guerre, 8 mai 1941, source : Bibliothèque nationale de France, ark:/12148/bpt6k57437998

Valoriser des fonds sériels

Annuaire, index, dictionnaires, registres, tables, minutes, comptes-rendus...



Dictionnaire universel françois et latin, vulgairement appelé Dictionnaire de Trévoux [...], tome 8, 1771, source : Bibliothèque nationale de France, ark:/12148/bd6t5382849b

Valoriser des fonds sériels

Annuaire, index, dictionnaires, registres, tables, minutes, comptes-rendus...

<i>Journal officiel</i> du 15 Novembre 1889	CHAMBRE — SÉANCE DU 14 NOVEMBRE	Session extraordinaire de 1889 5
CHAMBRE DES DÉPUTÉS 5 ^e législature. — Session extraordinaire de 1889.	M. le président. Il n'y a pas d'observation sur le procès-verbal?... Le procès-verbal est adopté.	d'âge et de nationalité exigées par la loi, nous vous proposons de valider son élection. (Les conclusions du bureau sont mises aux voix et adoptées. — M. Herbet est admis.)
COMPTE RENDU IN EXTENSO. — 2^e SÉANCE Séance du jeudi 14 novembre.	EXCUSE. — DEMANDE DE CONGÉ M. le président. M. Turrel s'excuse de ne pouvoir assister à la séance de ce jour et demande un congé. La demande est renvoyée à la commission des congés.	M. Dautresme, rapporteur. — Département de l'Ain, arrondissement de Bourg, 1 ^{re} circonscription. Les élections du 22 septembre ont donné les résultats suivants : Electeurs inscrits : 17,885, dont le quart est de 4,472. Nombre des votants, 13,816. Bulletins blancs et nuls à déduire, 126. Suffrages exprimés, 13,690, dont la majorité absolue est de 6,845.
SOMMAIRE Excuse et demande de congé. Rapports d'élections, par MM. Viette, Dautresme, Guilleminot, de La Rochefoucauld duc de Boudeauville, Boudeau, Beauquier, Barodet,	VÉRIFICATION DES POUVOIRS M. le président. L'ordre du jour appelle la vérification des pouvoirs des membres de la Chambre.	

Journal officiel de la République française. Débats parlementaires. Chambre des députés : compte rendu in-extenso, Impr. du Journal officiel, 14 nov. 1889, source : Bibliothèque et Archives de l'Assemblée nationale, <ark:/12148/bpt6k6494793z>

Valoriser des fonds sériels

Annuaire, index, dictionnaires, registres, tables, minutes, comptes-rendus...

Quelques défis

- Des fonds qu'une **personne ne peut lire intégralement seule**.
- Sous forme **manuscrite** ou **imprimée**, de **qualité de conservation** variable.

De nombreuses opportunités

- Une **structuration "syntaxique"** souvent répétitive, contrainte et peu ambiguë.
- **Disponibilité numérique croissante**, souvent en "dormance".
- Quantité importante d'**informations exploitables** de multiples façons.
- **Progrès significatifs** des systèmes "d'IA"

⇒ *De bons candidats pour une exploitation semi-automatique !*

63	MÉMENTO TÉLÉPHONIQUE
MARIGNY, rond-point Marigny (8°)	Élysées 06-91.
MARIVAUX-PATHE, 15, bld des Italiens (2°)	Richelieu 83-90.
MARRET (Georges), 79, av. Champs-Élysées	Élysées 75-60.
MARTIN, Accessoires, 16 r. Dupetit-Thouars (3°)	Archives 64-44.

102	ALP. — ANG
ALPES (Hautes). — Emprunt de 113.000 francs à la Caisse des chemins vicinaux. V. loi 14. — Emprunt de 73.000 francs à la Caisse des chemins vicinaux. V. loi 196.	

Ababsa Mohamed, Algérie, 2° cl., 14° R.T.A. 151.
Ababsa Sibti, Algérie, 2° cl., 9° Sp. 153.
Abassi Ali, Algérie, 2° cl., 11° R.T.A. 151.
Abbes Slimane, Algérie, cap., 27° R.T.T. 151.
Abbou ben Aissa, Algérie, 2° cl., 66° R.A.D. 151.

68	TIT	TIV TIX
Titré. Un Livre bien titré. Quelques Auteurs Français ont voulu accorder ce terme pour figurer un Livre qui a un titre convenable, ou plutôt un titre fatigant & emphatique; & dans ce dernier sens il y en a de nombreux.		
TITRIER, f. m. Nom odieux que l'on donne à ceux que l'on accuse d'avoir fabriqué de faux titres. Il y en a eu de tous états & de toutes conditions. Sur quoi lisez la Diplomatique de Dom Mabillon, liv. 1, chap. 6. Ce font principalement les Bénédictins qu'on accuse d'avoir été en des <i>Papirs Titriers</i> . Les auteurs ne		
TIVE, TIFE, ou TIVAT. Rivière de l'Écosse méridionale, qui traverse la province de Tivendale, où elle se jette dans la Tweed.		
TIVÉDAL, ou TIVOTDALE. Nom d'une province de l'Écosse méridionale. <i>Tivota, Tivota</i> . Elle est entre la Marche de Tivendale, la Liddesdale & le Northumberland en Angleterre. Elle prend son nom de la Tive qui la traverse. Sa longueur est environ de quatre lieues, & la largeur moyenne de deux. Elle est		

...

Quelle(s) méthode(s) de travail ?

Nécessité (et difficulté) de connecter questions de recherche SHS et tâches “proxy” en traitement automatique de données.

Quelques approches possibles :

- **Compter**, calculer des statistiques sur des données massives et significatives
- Collecter et **croiser les sources** autour de quelques cas sélectionnés
- Détecter des **incohérences** ou anomalies dans des séries
- ...

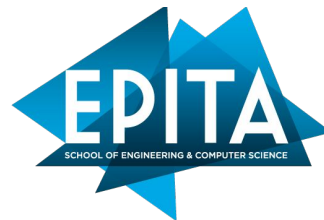
Nécessité de **produire une représentation structurée** à partir des données.

→ **Quelle(s) représentation(s) ?** *Tableaux (excel) pour le moment (et JSON)*

Le projet Mezanno

{BnF | Bibliothèque
nationale de France

L'ÉCOLE
DES HAUTES
ÉTUDES EN
SCIENCES
SOCIALES



IGN
INSTITUT NATIONAL
DE L'INFORMATION
GÉOGRAPHIQUE
ET FORESTIÈRE

BnF
2 personnes

EHESS
2 personnes

EPITA
4 personnes

IGN
2 personnes

Mezanno : Proposer des **outils numériques** aux **chercheurs·ses en SHS** leur permettant de mettre en œuvre **leurs méthodes** de travail, en **autonomie**, afin d'exploiter les données contenues dans des **corpus d'archives sérielles, répétitives, structurées et massives**.

<https://mezanno.xyz>

Mezanno (2024-2026) est un projet du **Plan quadriennal** de la recherche **BnF** (2024-2027).

En quoi **Mezanno** est-il différent ?

Plateformes existantes

Transcription collaborative (ex: *TACTEO*, *FromThePage*) : transcription à large échelle possible mais processus manuel, manque d'automatisation, tâches simples.

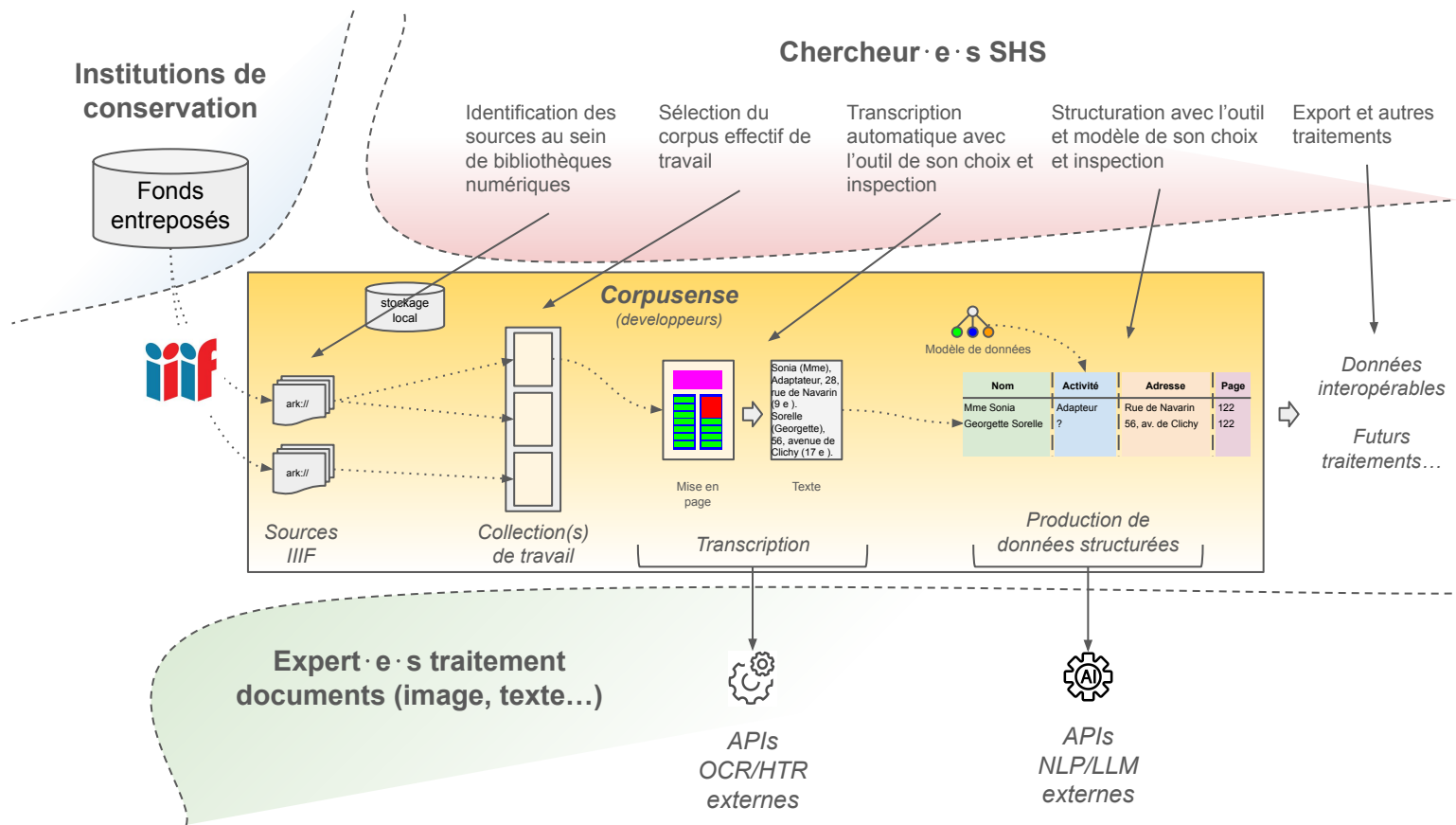
Transcription automatique (ex: *Transkribus*, *eScriptorium*) : robustesse reconnue pour la conversion image-texte mais pas d'analyse du texte produit, besoin d'expertise.

Plateformes intégrées (ex: *Arkindex*, *Glycerine*) : pipelines complexes possibles et support IIIF mais environnements fermés, besoin d'expertise.

Extraction spécialisée (ex: *Nuextract.ai*) : interfaces intuitives mais outils propriétaires, perte de contrôle sur les données, peu adaptés aux sources historiques.

Initiatives récentes interopérables (ex: *AIKON*, *Immarkus*) : contrôle par l'utilisateur (stockage local) et standards ouverts mais focalisées sur la détection d'objets dans l'image, pas sur la structuration de données textuelles.

L'écosystème en cours de développement



Caractéristiques principales

Faciliter le “**transfert technologique**”

- Permettre aux collègues en SHS de travailler en **autonomie**
- Faciliter la mise à disposition d'outils sous forme **d'API externes**

Limiter les **problèmes et coûts de maintenance** au minimum

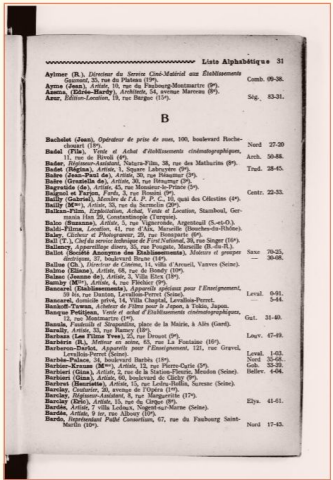
- **Pas d'installation** ou d'hébergement
- **Stockage local** des données
- Tirer profit des **entrepôts de données existants compatibles IIIF**
- **Code ouvert** et libre

Expérimenter une **méthode simple**

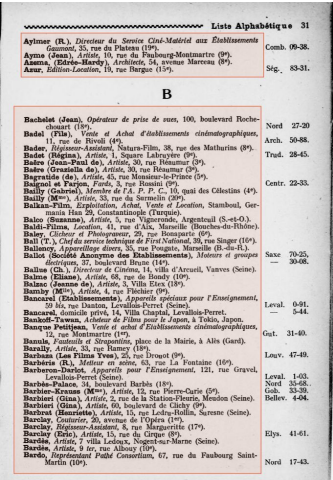
- **Focalisation sur un cas d'utilisation**
- **Petites volumétries**
- Un enchaînement de **traitements élémentaires**
- **Faciliter l'usage** tout en menant une **réflexion critique** en parallèle

Une chaîne de traitement *ad hoc*

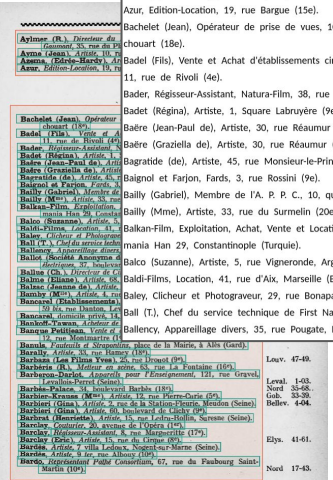
Chaîne de traitement



Détection de la mise en page



Correction manuelle (si nécessaire)



OCR PERO OCR

Besoin de définir un modèle de données et des instructions.

	A	B	
1	nom	prénom	activité
2	Aylmer	R.	Directeur du Service Ciné
3	Ayme	Jean	Artiste
4	Azema	Edrée-Hardy	Architecte
5	Azur		Edition-Location
6	Bachelet	Jean	Opérateur de prise de vue
7	Badel	Fils	Vente et Achat d'établissements
8	Bader		Régisseur-Assistant

Génération de données structurées Mistral Large

Extraire des données structurées

Approches possibles

1

Expressions régulières,
recherche de motifs dans le
texte

2

3

Atouts

Technologie simple.
Calculs très rapide.

Limites

Expressions régulières

Donnée en **entrée**

```
Liste Alphabétique 27  
Trud. 60-73.  
Andreyor (Yvette), Artiste, 31, rue Victor-Massé (9e).  
Andwig (Jens), Acheteur de films pour la Norvège, à Christiana, Norvège.  
Angelo (Jean), Artiste, 11, boulevard Montparnasse (15e).  
Angély, Artiste, 68, rue Monge (5e).  
Anglo Foreing Général Commerce Co, 5, rue de Provence (9e).  
Annales (Les), Hebdomadaire, 5, rue Labruyère (9e). Réd. cinégr.  
Mlle Brézillon.  
Argenteuil (Cinéma Moderne d').  
Argenteuil (Lutétia d').  
  
[...]
```

Donnée en **sortie**

```
Liste Alphabétique 27  
Trud. 60-73.  
Andreyor (Yvette), Artiste, 31, rue Victor-Massé (9e).  
Andwig (Jens), Acheteur de films pour la Norvège, à Christiana, Norvège.  
Angelo (Jean), Artiste, 11, boulevard Montparnasse (15e).  
Angély, Artiste, 68, rue Monge (5e).  
Anglo Foreing Général Commerce Co, 5, rue de Provence (9e).  
Annales (Les), Hebdomadaire, 5, rue Labruyère (9e). Réd. cinégr.  
Mlle Brézillon.  
Argenteuil (Cinéma Moderne d').  
Argenteuil (Lutétia d').  
  
[...]
```

Idée de l'approche

Décrire le format d'une entrée, sur la base de sa "syntaxe" dont certaines parties peuvent être [optionnelles], comme par exemple :

nom [(prénom)], [activité], [numéro, rue (arr)].

Chercher les entrées **correspondant à ce format** (ignorer celles qui ne correspondent pas) et extraire les **fragments** correspondant à chaque partie.

Problèmes : Comment décrire toutes les situations possibles ? De surcroît en présence d'erreurs de transcription ?

Extraire des données structurées

Approches possibles

1

Expressions régulières,
recherche de motifs dans le
texte

2

Utilisation de modèles IA
extractifs (BERT-like)

3

Atouts

Technologie simple.
Calculs très rapide.

Meilleure performance en théorie.
Calculs assez rapides.

Limites

Aucune robustesse au bruit OCR ni à la
variabilité des syntaxes. Très fastidieux à
mettre en œuvre. À éviter.

Modèles extractifs

Donnée en entrée

Liste Alphabétique 27
Trud. 60-73.
Andreyor (Yvette), Artiste, 31, rue Victor-Massé (9e).
Andwig (Jens), Acheteur de films pour la Norvège, à Christiana, Norvège.
Angelo (Jean), Artiste, 11, boulevard Montparnasse (15e).
Angély, Artiste, 68, rue Monge (5e).
Anglo Foreign Général Commerce Co, 5, rue de Provence (9e).
Annales (Les), Hebdomadaire, 5, rue Labruyère (9e). Réd. cinégr.
Mlle Brézillon.
Argenteuil (Cinéma Moderne d').
Argenteuil (Lutétia d').

[...]

Donnée en sortie (exemple possible)

Liste Alphabétique 27
Trud. 60-73.
Andreyor (Yvette), Artiste, 31, rue Victor-Massé (9e).
Andwig (Jens), Acheteur de films pour la Norvège, à Christiana, Norvège.
Angelo (Jean), Artiste, 11, boulevard Montparnasse (15e).
Angély, Artiste, 68, rue Monge (5e).
Anglo Foreign Général Commerce Co, 5, rue de Provence (9e).
Annales (Les), Hebdomadaire, 5, rue Labruyère (9e). Réd. cinégr.
Mlle Brézillon.
Argenteuil (Cinéma Moderne d').
Argenteuil (Lutétia d').

[...]

Idée de l'approche

Apprendre automatiquement comment “peindre” les données selon leur catégorie, grâce à un **grand nombre d'exemples**, sur la base de motifs probable de début/fin de champ.

Catégories possibles : nom, prénom, activité, numéro, rue, art

Données d'apprentissage possibles :

Andreyor (Yvette), Artiste, 31, rue Victor-Massé (9e).
Angelo (Jean), Artiste, 11, boulevard Montparnasse (15e).

Appliquer le modèle appris sur de nouvelles données, en espérant qu'il s'adapte bien aux situations imprévues.

Problèmes : Besoin d'un nombre conséquent d'exemples en pratiques, et surtout de compétences en IA pour entraîner un nouveau modèle.

Extraire des données structurées

Approches possibles

1

Expressions régulières,
recherche de motifs dans le
texte

2

Utilisation de modèles IA
extractifs (BERT-like)

3

Utilisation de modèles IA
génératifs (ChatGPT, Mistral...)
avec **prédictions structurées**.

Atouts

Technologie simple.
Calculs très rapide.

Meilleure performance en théorie.
Calculs assez rapides.

Pas besoin de données d'entraînement.
Utilisation de modèles peu onéreux en
ligne (facile à prototyper).

Limites

Aucune robustesse au bruit OCR ni à la
variabilité des syntaxes. Très fastidieux à
mettre en œuvre. À éviter.

Besoin d'un entraînement spécifique,
donc de données annotées en quantité
suffisante.

Modèles génératifs

Donnée en **entrée**

Liste Alphabétique 27
Trud. 60-73.
Andreyor (Yvette), Artiste, 31, rue Victor-Massé (9e).
Andwig (Jens), Acheteur de films pour la Norvège, à Christiana, Norvège.
Angelo (Jean), Artiste, 11, boulevard Montparnasse (15e).
Angély, Artiste, 68, rue Monge (5e).
Anglo Foreing Général Commerce Co, 5, rue de Provence (9e).
Annales (Les), Hebdomadaire, 5, rue Labruyère (9e). Réd. cinégr.
Mlle Brézillon.
Argenteuil (Cinéma Moderne d').
Argenteuil (Lutétia d').
[...]

Donnée en **sortie** (exemple possible)

```
[
  { "nom": "Andreyor", "prenom": "Yvette", "genre": "F", ... },
  { "nom": "Andwig", "prenom": "Jens", "genre": "X", ... },
  { "nom": "Angelo", "prenom": "Jean", "genre": "H", ... },
  { "nom": "Angély", "prenom": "", "genre": "F", ... },
  { "nom": "Anglo", "prenom": "Foreing", "genre": "X", ... },
  { "nom": "Hebdomadaire Les Annales", "prenom": "", "genre": "X", ... },
  ...]
```

Idée de l'approche

Utiliser un **modèle de langue large** (ChatGPT, Mistral), et lui demander de “**traduire**” les données dans un **nouveau format**.

Il est possible (usage avancé) de contraindre la réponse à respecter un **format précis, lisible par une machine**.

Exemple de prompt : “Voici un extrait d'annuaire du cinéma. Extrais la liste des personnes en indiquant leur nom, prénom et genre. S'il s'agit d'une entreprise, indique “X” pour le genre. Respecte le format suivant ...”

+ Schéma : “**nom**”: texte, “**prenom**”: texte...

Problèmes : Difficulté à définir le bon prompt, risque d'hallucination du système...

Extraire des données structurées

	Approches possibles	Atouts	Limites
1	Expressions régulières, recherche de motifs dans le texte	Technologie simple. Calculs très rapide.	Aucune robustesse au bruit OCR ni à la variabilité des syntaxes. Très fastidieux à mettre en œuvre. À éviter.
2	Utilisation de modèles IA <u>extractifs</u> (BERT-like)	Meilleure performance en théorie. Calculs assez rapides.	Besoin d'un entraînement spécifique, donc de données annotées en quantité suffisante.
3	Utilisation de modèles IA <u>génératifs</u> (ChatGPT, Mistral...) avec prédictions structurées .	Pas besoin de données d'entraînement. Utilisation de modèles peu onéreux en ligne (facile à prototyper).	Difficulté de prompting. Besoin d'une évaluation rigoureuse, en particulier au niveau des biais et (idéalement) d'une détection des risques d'hallucination. Besoin de machines puissantes.

Note: Les Large Vision-Language Models sont une 4e voie possible.

Validation **pratique**

Premiers résultats et démarche

Accélération sensible par rapport à l'approche SoDUCo pour les **petites séries**

→ *de 2 semaines de travail à quelques dizaines de minutes*

Accueil positif de la part des **testeurs**

→ 4+ sessions de formations organisées depuis cet été

→ documentation et tutoriels vidéo presque terminés

Validation de la capacité à **produire des données pertinentes** sur un cas d'annuaires du cinéma (*diapos suivantes*)

Projet “Les femmes dans les métiers du cinéma”

Porté par la **BnF**, Soutenu par l'**EUR ArTeC** sur 2 ans (2025-2026)

Objectif : Sortir les professionnelles du cinéma français de leur invisibilité les compter, les nommer, raconter leur histoire

Forte dimension **collective** (50 participants à ce jour)

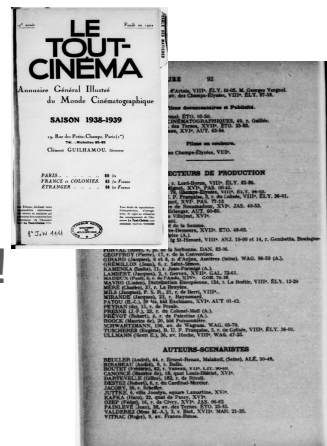
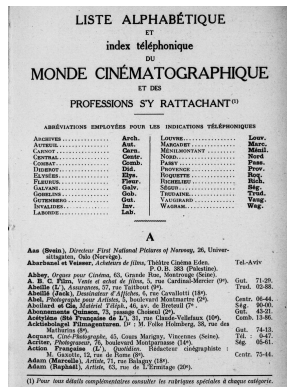
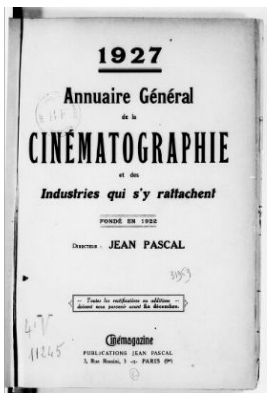
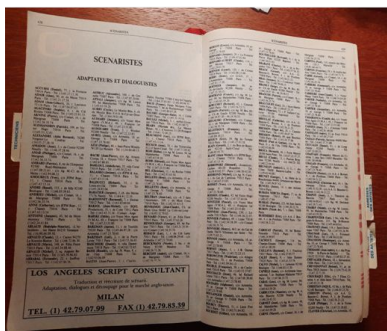
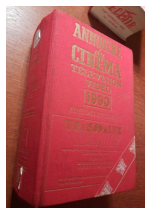


Corpus : Annuaires professionnels du cinéma 1920-1960

Une dizaine d'annuaires identifiés pour l'instant

Traitement manuel trop **fastidieux** (1000+ pages / annuaire)

Structuration et informations qui peuvent **varier** selon les listes !



Annotation de la liste alphabétique du Cinémagazine 1927

Démarche générale :

- Génération automatique des données structurées pour la liste alphabétique principale du Cinémagazine 1927 (96 pages)
- Correction par une équipe de volontaires en comparant pages et “entrées” extraites
→ *Pas de problème d'alignement*

Liste Alphabétique 29

Astruc, Artiste, 76, rue Rébéval (19e).
Ata et Cie, Acheteurs de films pour le Japon, Teramachi-Kurumagachi, Sagaru, à Kioto, Japon
Ateliers de la Clicherie spéciale (Dubray), 23, boulevard des Italiens (2e). Louv. 16-41.
Ateliers Fantasia (S. A.). Tout usinage du film, 10, 13 et 15, rue Piat (20e). Roq. 22-68.
Attilio Buccianti, Achat de Films pour l'Egypte, la Syrie et la Palestine, 14, rue Nabi-Daniel, Alexandrie, Egypte.
Attraction du Cinéma (L'), Agence théâtrale et cinématographique 38, rue Vignon (9e).
Aubert (Louis), O. *, Président de la Chambre Syndicale de la Cinématographie française, 124, avenue de la République (11e). Roq. 73-31.
Aubert (Etablissements Louis), Edition, Location, Appareillage, 124, avenue de la République (11e). Roq. 73-31, 73-32.

Tr nom	prénom/titre/init.	Tr activité	Tr adresse	type	genre
Astruc		Artiste	76, rue Rébéval (19e).	physique	I
Ata et Cie		Acheteurs de films pour le Japon	Teramachi-Kurumagachi, Sagaru, à Kioto, Japon	morale	
Ateliers de la Clicherie spéciale (Dubray)			23, boulevard des Italiens (2e).	morale	
Ateliers Fantasia (S. A.)		Tout usinage du film	10, 13 et 15, rue Piat (20e).	morale	
Attilio Buccianti		Achat de Films pour l'Egypte, la Syrie	14, rue Nabi-Daniel, Alexandrie, Egypte.	morale	X
Attraction du Cinéma (L')		Agence théâtrale et cinématographi	38, rue Vignon (9e).	morale	
Aubert	Louis	Président de la Chambre Syndicale c	124, avenue de la République (11e).	physique	H
Etablissements Louis Aubert		Edition, Location, Appareillage	124, avenue de la République (11e).	morale	

Annotation de la liste alphabétique du Cinémagazine 1927

Résultats :

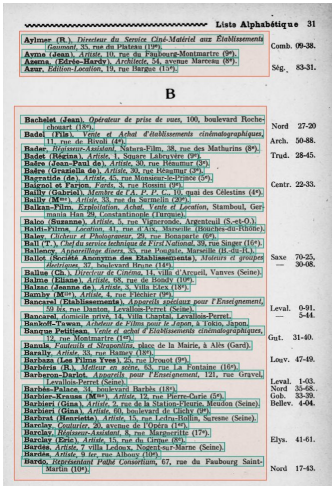
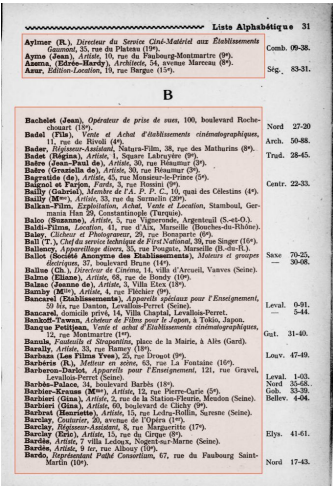
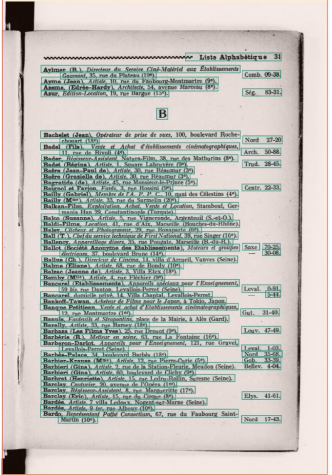
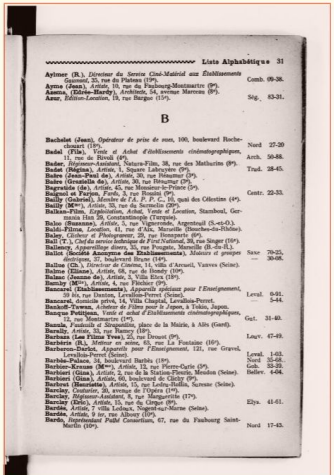
- 4309 “entrées”, dont 1489 personnes morales, 2820 personnes physiques
 - Personnes physiques : **661 femmes**, 1159 hommes, 1000 indéterminés
- Un matériau essentiel pour une première estimation de la pertinence de l’approche

Dépôt Nakala : DOI: [10.34847/nkl.e7bcd9nn](https://nkl.e7bcd9nn) 

- Données produites automatiquement
- Données corrigées
- Code évaluation

Petit bémol : détection de mise en page semi-automatique

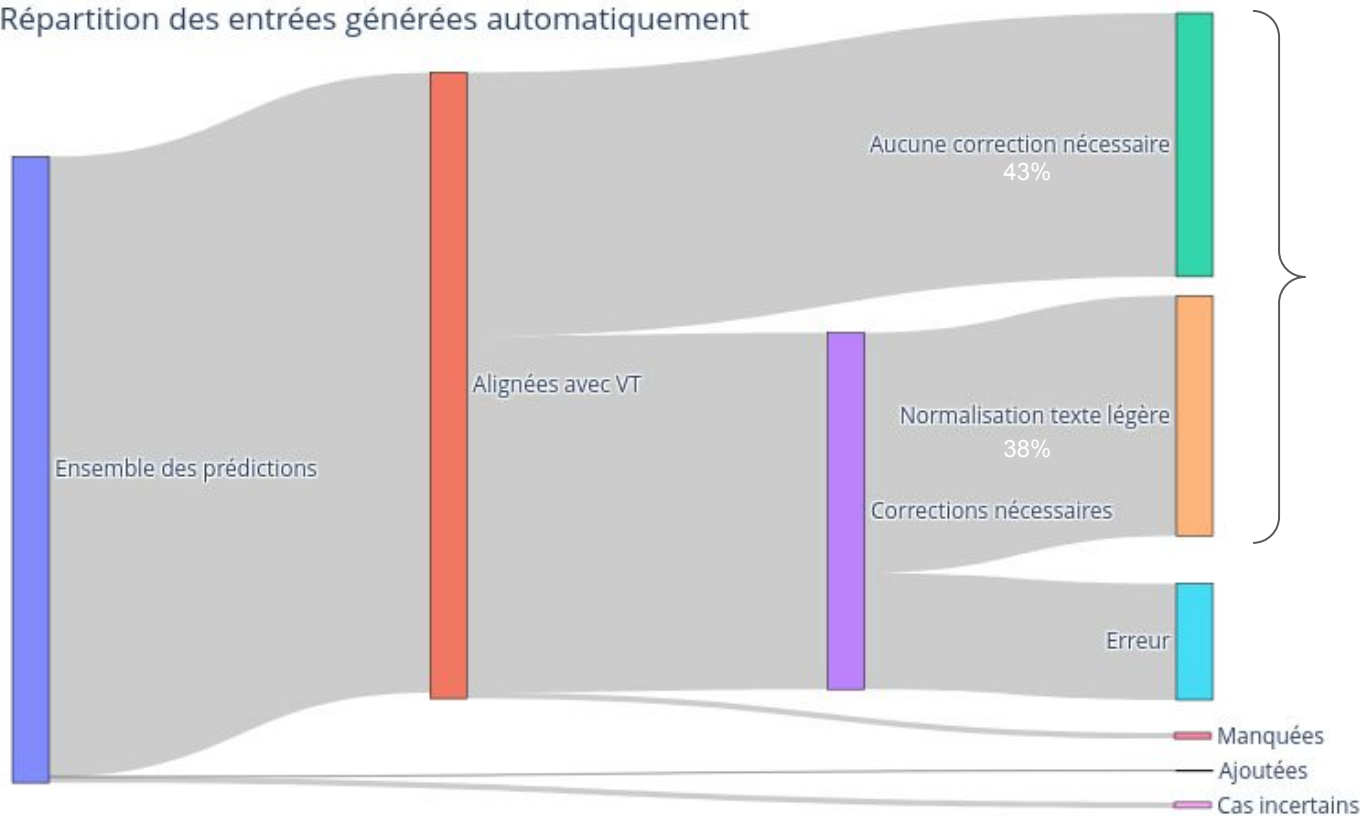
En l'absence de système permettant une détection robuste de l'ordre de lecture, nous avons corrigé manuellement la détection de mise en page.



Sans correction manuelle Avec correction manuelle

Analyse rapide des résultats

Répartition des entrées générées automatiquement



81%
de données
exploitables

Cas positifs

43% des entrées générées n'ont nécessité **aucune correction** des annotateurs.

38% des entrées générées sont identiques à une normalisation près.

Minuscules/majuscules, espaces, ponctuation...

86% des entrées générées ont des champs “Nom”, “Genre” et “Type” exactement identiques aux valeurs attendues

Le **genre** inféré est-il fiable ? → Exactitude de **94%**

Le **type de personne** inféré (morale/physique) est-il fiable ? → Exactitude **97%**

Erreurs : quelques exemples choisis au hasard

	id	nom_gt	nom_pred	prénom/titre/init.	prénom	adresse_gt	adresse_pred	activité_gt	activité_pred	type_gt	type_pred	genre_gt	genre_pred
4101	4140	Verrerie Scientifique (La)	Verrerie Scientifique	nan	La	12, av. du Maine (14e)	12, av. du Maine (14e)	Appareillage électr.	Appareillage électr.	morale	morale	X	X
1435	1454	Dix	Dix	Mad	Mad	31, avenue Mac-Mahon (17e).	31, avenue Mac-Mahon (17e).	Artiste	Artiste	physique	physique	I	F
3430	3458	Rateau (Société)	Rateau	nan	nan	40, rue du Colisée (8e)	40, rue du Colisée (8e)	Aspirateur de fumée	Aspirateur de fumée	morale	morale	X	X
2262	2282	Homme Libre (L')	Homme Libre	nan	L'	13, rue de Marivaux (2e)	13, rue de Marivaux (2e)	Quotidien	Quotidien	morale	morale	X	X
3141	3168	Taponier	Taponier	Em.	Em.	47, avenue Trudaine (9e).	nan	Directeur (Office Cinématographique Suisse)	Directeur	physique	physique	I	I
961	967	Cinéma Pathé-Secrétan	Pathé-Secrétan	nan	nan	1, rue Secrétan (19e)	1, rue Secrétan (19e)	nan	nan	morale	morale	X	X
3001	3027	Montrouge	Montrouge	Mlle	nan	8, rue Pierre Ginier (8e).	8, rue Pierre Ginier (8e).	Représentante des Films Paramount	Représentante des Films Paramount	physique	physique	F	I
926	932	Cinéma Luna	Luna	nan	nan	9, cours de Vincennes (20e)	9, cours de Vincennes (20e)	nan	nan	morale	morale	X	X
1311	1328	Delavenne	Delavenne	nan	nan	80, rue Taitbout (9e)	80, rue Taitbout (9e)	Administrateur d'Edifo	Administrateur d'Edifo	physique	physique	I	H
970	976	Cinéma Rochechouart	Rochechouart	nan	nan	66, rue Rochechouart (18e)	66, rue Rochechouart (18e)	nan	nan	morale	morale	X	X
1471	1490	Dracourd	Dracourd	Mlle	Mile	28, place Saint-Georges, Paris (9e).	28, place Saint-Georges, Paris (9e).	Représentante Mappemonde-Film	Représentante Mappemonde-Film	physique	physique	F	F
915	921	Cinéma Grenelle-Aubert-Palace	Grenelle-Aubert-Palace	nan	nan	142, avenue Emile-Zola (15e)	142, avenue Emile-Zola (15e)	nan	nan	morale	morale	X	X
2534	2555	Lapierre (Etablissements)	Lapierre	nan	nan	68, rue des Haies (20e).	68, rue des Haies (20e).	Appareils de projection	Appareils de projection	morale	morale	X	X
509	513	Boulogne-Billancourt (Kursaal-Cinéma de)	Boulogne-Billancourt	nan	nan	nan	nan	nan	Kursaal-Cinéma de	morale	morale	X	X
3297	3325	Picard	Picard	nan	nan	126, rue de Javel (15e).	126, rue de Javel (15e).	Moteurs et Groupes électrogènes	Moteurs et Groupes électrogènes	morale	physique	X	I
1951	1971	Gaulois (Le)	Gaulois	nan	nan	2, rue Drouot (9e).	2, rue Drouot (9e).	Quotidien	Quotidien	morale	morale	X	X

Erreurs : quelques exemples choisis au hasard

Beaucoup de **micro-différences** (présence de déterminants ou non par exemple)

Et les **erreurs de classification** de genre et de type de personne.

Une **méthode de liage robuste** devrait réussir à regrouper une partie de ces cas

Éléments ajoutés

Détail des 3 “hallucinations” potentielles

#	Id	Tr nom	prénom/titre/init.	Tr activité	Tr adresse	type	genre
737		Manutention			19, rue Oberkampf (11e)	morale	
2287		Georges-Presberg		Instruments de Musique	32, rue des Archives (4e)	physique	I
3249		Pathé				morale	

Centrale Cinématographique (La). (MM. H. Etievant, M. Nalpas et J. Natanson). Siège Social : 74, avenue Kléber (16^e).
— Manutention, 19, rue Oberkampf (11^e).

Hervilleur (H.) et Georges-Presberg, Instruments de Musique, 32, rue des Archives (4^e).

Pathé (Charles), O. ✱, 38, avenue du Président-Wilson (16^e).
Pathé (Jacques), Administrateur de la Société « Pathé Baby », rue Louis-Besquel, Vincennes (Seine).
Pathé (Matériel Cinématographique), 20 bis, rue Lafayette (9^e).
Pathé-Baby, Caméra et Appareil de projections, 20 bis, r. Lafayette (9^e).
— même adresse.
Pathé-Cinéma (Anciens Etab. Pathé Frères, Société anon. au capital de 45 millions de francs). Siège social : 30, boul. des Italiens (9^e).
Pathé-Cinéma. Administration centrale : 117, boul. Haussmann (8^e).
—
—
Pathé-Cinéma. Usines : 30, rue des Vignerons, Vincennes (Seine) et quai Victor-Brisson, Joinville-le-Pont (Seine).
Pathé-Consortium-Cinéma Directeur, Administrateur-délégué : M. Jean Sapène, C. ✱, 5 et 7, rue du Faubourg-Poissonnière (10^e).
Pathé Frères, Phonos, Radio, 30, boulevard des Italiens (9^e).
Pathé-Journal, 24, rue des Vignerons, Vincennes (Seine).
Pathé News of América, 28, rue Tronchet (8^e).
Pathé-Nord (Société Française), 11, boulevard de la Madeleine (8^e).

Pas de problème ?

Éléments manqués

Il s'agit très majoritairement d'éléments en début / fin de page pour lesquels **les zones de texte n'ont pas été détectées**.

Pour rappel, la chaîne de traitement fonctionne de la façon suivante :

Pour chaque image

→ *détection mise en page*

→ *transcription*

→ *production de données structurées avec LLM*

Ces erreurs sont plutôt dues aux étages de détection de mise en page et d'OCR.

À retenir de l'évaluation

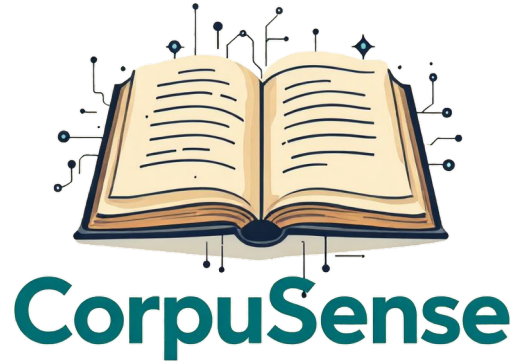
La méthode et les outils qui la mettent en œuvre peuvent **produire des données pertinentes**, lorsque les sources sont exploitables, le modèle de données bien établi, le *prompt* bien choisi, le LLM adapté, les risques faibles...

Les résultats sont **très dépendants des instructions** fournies au modèle.

Nous sommes encore au **début de l'utilisation de grands modèles de langue** pour ces problématiques, et la **modélisation** est encore largement perfectible.

D'autres corpus sont à l'étude, ouvrez Corpusense pour analyser le vôtre !

Démonstration de **Corpusense**



<https://mezanno.xyz/corpusense>

https://youtu.be/q8bAvNmv0_4



CorpuSense

En résumé

Points forts

APIs **génériques** + interface **spécialisée** = pari payant

- Utilisabilité
- Facilité d'ajout de nouveaux modèles

Méthode pertinente : la qualité des données produites peut être bonne

Terrain d'expérimentation et d'identification des verrous restants

Défis restants

Sécuriser la **maintenance** du projet

Consolider le **déploiement** des outils (APIs)

Enrichir l'offre de **modèles**

Poursuivre les travaux sur la **confiance** qu'on peut accorder aux données

- Évaluations avec des données de référence
- Retour à la source

Toucher davantage d'**utilisateurs**

Pour en savoir plus / suivre nos avancées

Abonnez-vous à corpusense-news@groupe.renater.fr

via <https://groupe.renater.fr/sympa/info/corpusense-news>

Consultez notre **site web** : <https://mezanno.xyz>

Regardez nos **tutos YouTube** : <https://www.youtube.com/@Mezanno>

Récupérer cette **présentation** → <https://hnurl.fr/wpqgkw>



Réserve

Chaîne de traitements SoDUCo

Chaîne de traitement pour l'extract. et la reco. des entrées

Pré-traitement image
(correction géométrique,
réduction de bruit...)

Segmentation du canvas
Classification des blocs

Lemonnyer, plomberie, r. de Bondy, 86, et
r. Bouchardon, 1.

Transcription

Lemonnyer, plomberie, r. de Bondy,
86, et r. Bouchardon, 1.

Extraction des
informations clés

PERSONNE	ACTIVITÉ	RUE
Lemonnyer,	plomberie,	r. de Bondy,
NUMÉRO	RUE	NUMÉRO
86,	et	r. Bouchardon, 1.

Reco. d'entités nommées (*Named Entity Recognition*)

Identifier et classer des expressions d'un ou plusieurs mots d'un texte en catégories prédéfinies : personne, lieu, organisation, date, prix...

Entités spécifiques dans le cas des annuaires.

Exemple:

PERSON ACTIVITY
Ravrio et comp., fabr. de bronzes et curiosités,
r. Richelieu, 93; la fabrique rue Montmartre, 161.
LOCATION CARDINAL FEATURE TYPE LOCATION CARDINAL

Chaîne de traitements Mezanno

Chaîne de traitement actuelle

images → *tables*

Nous utilisons une chaîne de traitement en 3 étapes

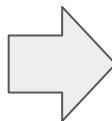
Étape 1

Détection de la **mise en page** des documents

Module dédié ou intégré à l'OCR.

Liste Alphabétique 27

Andreyor (Yvette), Artiste, 31, rue Victor-Massé (9 ^e).	Trud. 60-73.
Andwig (Jens), Acheteur de films pour la Norvège, à Christiania, Norvège.	
Angelo (Jean), Artiste, 11, boulevard Montparnasse (15 ^e).	
Angély, Artiste, 68, rue Monge (5 ^e).	
Anglo Foreign Général Commerce C ^o , 5, rue de Provence (9 ^e).	Centr. 48-32.
Annales (Les), Hebdomadaire, 5, rue Labryère (9 ^e). Réd. cinégr. M ^{lle} Brézillon.	Trud. 00-60
Argenteuil (Cinéma Moderne d').	Argent. 1-60.
Argenteuil (Lutétia d').	Argent. 3-73.
Argenteuil (Casino-Cinéma d').	Argent. 0-34.
Anonges, Artiste, 128, rue d'Aboukir (2 ^e).	
Ansald, Sièges et Fauteuils, 115, 117, rue Sainte, Marseille (B.-du-R.).	
Antillan-Film C ^o , Achat de films pour les Antilles, Aguila, La Havane, Cuba.	
Antiseptique Olfior, Assainissement, 86, rue du Président-Wilson, La Plaine Saint-Denis, Seine.	Nord 09-76
Antoine (André), O. s. Critique dramatique, 28, place Dauphine (1 ^{er}).	Gob. 03-59.
Apers, Photographe pour artistes, 23, rue Boissy-d'Anglas (8 ^e).	Elys. 62-58.
Appareillages électriques Grivolais, 16, rue Montgolfier (3 ^e).	Arch. 30-55.
Appareillages électriques Supra, 15, boulevard Saint-Germain (5 ^e).	Gob. 49-64.
Appareils contrôleurs (Sté Univers. des), Construct. mécan. 44, rue de Chanzy (11 ^e).	Roq. 02-57.
Appignan et Pinotti, Acheteurs de films pour l'Italie, 210, via Tritone, Rome, Italie.	
Aps (Félix d'), Artiste, 1, rue des Ursins (4 ^e).	Gob. 19-29.
Arango et Salvador, Acheteurs de films pour la République de Colombie, 51, rue de Paradis (10 ^e).	Gut. 27-75.
Arbid (David), Acheteur de films 11, rue Mariette-Pacha, Alexandrie, Egypte.	— 27-26.
Arcos, 1, rue Taïlbout (9 ^e).	
Arditi, Vente et achat de films, 5, rue Bouchardon (10 ^e).	Nord 84-02.
Ardouin (Léon), Directeur technique de la Société « Ciné Documentaire », 7, rue Beaurepaire.	
Argus Films Productions, S. A., Production, Denis Ricaud, Administrateur, 39, boulevard Haussmann (9 ^e).	Gut. 18-07.
Arias Film, Acheteur de films pour l'Italie, 336, via Balangero, Turin.	Centr. 55-84.
Ariel (Raymonde), Artiste, 1, rue Bellanger, Neuilly-sur-Seine, Seine.	
Ariscan, Artiste, 37, rue du Départ (14 ^e).	
Aristocratie del film (La), Acheteurs de films, 45, rue Laborde (8 ^e).	Lab. 26-20.
Arlix (W. d'), Artiste, 27, rue du Château, Fontainebleau (S.-et-M.).	
Arly (Jacqueline), Artiste, 30, rue du Général-Foy (8 ^e).	
Armand (Aug.), Administrateur du Gaumont-Palace, 3, rue Caulaincourt (18 ^e).	Marc. 00-46.
Armand, Ameublement, 109, rue Lamarck (18 ^e).	Marc. 29-92.
Armand (Georgette), Artiste, 35, rue Joffroy (17 ^e).	Wag. 95-21.
Armand (Paul Loys), Décorateur, 20, rue Ravignan (18 ^e).	
Armel (Yette), Artiste, 160, boulevard Malesherbes (8 ^e).	
Armort (Paul), Artiste, 82, boulevard Flandrin.	Pass. 27-40.
Armor (Les Films), Location. Directeur : Marcel Sprecher, 12, rue Gaillon (2 ^e).	Centr. 84-37.
Arnal, Directeur Cinéma, 2, avenue de Taillebourg, Paris (11 ^e).	
Arna (Jacques), Artiste, 76, rue des Petits-Champs (2 ^e).	
Arnold, Artiste, 48, rue Condorcet (9 ^e).	
Arnold (Agence Roger), Attractions, 8, Boul. Bonne-Nouvelle (10 ^e).	Prov. 36-35.
Arnou (André), Opérateur de prise de vues, 19, avenue Bellevue, Bry-sur-Marne (Seine).	
Arnou (Maurice), Opérateur de prise de vues, 11, rue Louis-Braille (12 ^e).	
Arnoux (Jeanne), Artiste, 4, rue Clairant (17 ^e).	



Liste Alphabétique 27

Andreyor (Yvette), Artiste, 31, rue Victor-Massé (9 ^e).	Trud. 60-73.
Andwig (Jens), Acheteur de films pour la Norvège, à Christiania, Norvège.	
Angelo (Jean), Artiste, 11, boulevard Montparnasse (15 ^e).	
Angély, Artiste, 68, rue Monge (5 ^e).	
Anglo Foreign Général Commerce C ^o , 5, rue de Provence (9 ^e).	Centr. 48-32.
Annales (Les), Hebdomadaire, 5, rue Labryère (9 ^e). Réd. cinégr. M ^{lle} Brézillon.	Trud. 00-60
Argenteuil (Cinéma Moderne d').	Argent. 1-60.
Argenteuil (Lutétia d').	Argent. 3-73.
Argenteuil (Casino-Cinéma d').	Argent. 0-34.
Anonges, Artiste, 128, rue d'Aboukir (2 ^e).	
Ansald, Sièges et Fauteuils, 115, 117, rue Sainte, Marseille (B.-du-R.).	
Antillan-Film C ^o , Achat de films pour les Antilles, Aguila, La Havane, Cuba.	
Antiseptique Olfior, Assainissement, 86, rue du Président-Wilson, La Plaine Saint-Denis, Seine.	Nord 09-76
Antoine (André), O. s. Critique dramatique, 28, place Dauphine (1 ^{er}).	Gob. 03-59.
Apers, Photographe pour artistes, 23, rue Boissy-d'Anglas (8 ^e).	Elys. 62-58.
Appareillages électriques Grivolais, 16, rue Montgolfier (3 ^e).	Arch. 30-55.
Appareillages électriques Supra, 15, boulevard Saint-Germain (5 ^e).	Gob. 49-64.
Appareils contrôleurs (Sté Univers. des), Construct. mécan. 44, rue de Chanzy (11 ^e).	Roq. 02-57.
Appignan et Pinotti, Acheteurs de films pour l'Italie, 210, via Tritone, Rome, Italie.	
Aps (Félix d'), Artiste, 1, rue des Ursins (4 ^e).	Gob. 19-29.
Arango et Salvador, Acheteurs de films pour la République de Colombie, 51, rue de Paradis (10 ^e).	Gut. 27-75.
Arbid (David), Acheteur de films 11, rue Mariette-Pacha, Alexandrie, Egypte.	— 27-26.
Arcos, 1, rue Taïlbout (9 ^e).	
Arditi, Vente et achat de films, 5, rue Bouchardon (10 ^e).	Nord 84-02.
Ardouin (Léon), Directeur technique de la Société « Ciné Documentaire », 7, rue Beaurepaire.	
Argus Films Productions, S. A., Production, Denis Ricaud, Administrateur, 39, boulevard Haussmann (9 ^e).	Gut. 18-07.
Arias Film, Acheteur de films pour l'Italie, 336, via Balangero, Turin.	Centr. 55-84.
Ariel (Raymonde), Artiste, 1, rue Bellanger, Neuilly-sur-Seine, Seine.	
Ariscan, Artiste, 37, rue du Départ (14 ^e).	
Aristocratie del film (La), Acheteurs de films, 45, rue Laborde (8 ^e).	Lab. 26-20.
Arlix (W. d'), Artiste, 27, rue du Château, Fontainebleau (S.-et-M.).	
Arly (Jacqueline), Artiste, 30, rue du Général-Foy (8 ^e).	
Armand (Aug.), Administrateur du Gaumont-Palace, 3, rue Caulaincourt (18 ^e).	Marc. 00-46.
Armand, Ameublement, 109, rue Lamarck (18 ^e).	Marc. 29-92.
Armand (Georgette), Artiste, 35, rue Joffroy (17 ^e).	Wag. 95-21.
Armand (Paul Loys), Décorateur, 20, rue Ravignan (18 ^e).	
Armel (Yette), Artiste, 160, boulevard Malesherbes (8 ^e).	
Armort (Paul), Artiste, 82, boulevard Flandrin.	Pass. 27-40.
Armor (Les Films), Location. Directeur : Marcel Sprecher, 12, rue Gaillon (2 ^e).	Centr. 84-37.
Arnal, Directeur Cinéma, 2, avenue de Taillebourg, Paris (11 ^e).	
Arna (Jacques), Artiste, 76, rue des Petits-Champs (2 ^e).	
Arnold, Artiste, 48, rue Condorcet (9 ^e).	
Arnold (Agence Roger), Attractions, 8, Boul. Bonne-Nouvelle (10 ^e).	Prov. 36-35.
Arnou (André), Opérateur de prise de vues, 19, avenue Bellevue, Bry-sur-Marne (Seine).	
Arnou (Maurice), Opérateur de prise de vues, 11, rue Louis-Braille (12 ^e).	
Arnoux (Jeanne), Artiste, 4, rue Clairant (17 ^e).	

Chaîne de traitement actuelle

images → *tables*

Nous utilisons une chaîne de traitement en 3 étapes

Étape 1

Détection de la **mise en page** des documents

Module dédié ou intégré à l'OCR.

Étape 2

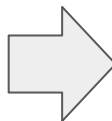
Transcription du texte (OCR).

Nous utilisons **PERO OCR**.

Liste Alphabétique 27

Andreyor (Yvette), Artiste, 31, rue Victor-Massé (9^e)
 Andwig (Jens), Acheteur de films pour la Norvège, à Christiana, Norvège.
 Angelo (Jean), Artiste, 1, boulevard Montparnasse (15^e)
 Angély, Artiste, 68, rue Monge (5^e)
 Anglo Foreign Général Commerce Co, 5, rue de Provence (9^e)
 Annales (Les), Hebdomadaire, 5, rue Labruyère (9^e). Réd. cinégr.
 Mlle Brézillon.
 Argenteuil (Cinéma Moderne d')
 Argenteuil (Lutétia d')
 Argenteuil (Casino-Cinéma d')
 Anonges, Artiste, 128, rue d'Aboukir (2^e)
 Ansaldo, Sièges et Fauteuils, 115, 117, rue Sainte, Marseille (B.-du-R.)
 Antillan-Film Co, Achat de films pour les Antilles, Aguila, La
 Havane, Cuba.
 Antiseptique Oliflor, Assainissement, 86, rue du Président-Wilson,
 La Plaine Saint-Denis, Seine.
 Antoine (André), O. ♦, Critique dramatique, 28, place Dauphine (1^{er})
 Apers, Photographe pour artistes, 23, rue Boissy-d'Anglas (8^e)
 Appareillages électriques Grivolos, 16, rue Montgolfier (3^e)
 Appareillages électriques Supra, 15, boulevard Saint-Germain (5^e)
 Appareils contrôleurs (Sté Univers. des), Construct. mecan. 44, rue
 de Chanzy (11^e)
 Appignan et Pinotti, Acheteurs de films pour l'Italie, 210, via Tritone,
 Rome, Italie.
 Aps (Félix d'), Artiste, 1, rue des Ursins (4^e)
 Arango et Salvador, Acheteurs de films pour la République de
 Colombie, 51, rue de Paradis (10^e)
 Arbid (David), Acheteur de films 11, rue Mariette-Pacha, Alexandrie,
 Egypte.
 Arcos, 1, rue Taillout (9^e)
 Arditi, Vente et achat de films, 3, rue Bouchardon (10^e)
 Ardouin (Leon), Directeur technique de la Société « Cine Documentaire »,
 7, rue Beaurepaire.
 Argus Films Productions, S. A., Production, Denis Ricaud, Admi-
 nistrateur, 39, boulevard Haussmann (9^e)
 Arias Film, Acheteur de films pour l'Italie, 336, via Balangero, Turin.
 Ariel (Raymonde), Artiste, 1, rue Bellanger, Neuilly-sur-Seine, Seine.
 Ariscan, Artiste, 37, rue du Départ (14^e)
 Aristocratie del film (La), Acheteurs de films, 45, rue Laborde (8^e)
 Arlix (W. d'), Artiste, 27, rue du Château, Fontainebleau (S.-et-M.)
 Arly (Jacqueline), Artiste, 30, rue du Général-Foy (8^e)
 Armand (Ang.), Administrateur du Gaumont-Palace, 3, rue Caulain-
 court (18^e)
 Armand, Ameublement, 109, rue Lamarck (18^e)
 Armand (Georgette), Artiste, 35, rue Jouffroy (17^e)
 Armand (Paul Lova), Décorateur, 20, rue Ravignan (18^e)
 Armei (Yette), Artiste, 160, boulevard Malesherbes (8^e)
 Armont (Paul), Artiste, 82, boulevard Flandrin.
 Armor (Les Films), Location. Directeur : Marcel Sprecher, 12, rue
 Gaillon (2^e)
 Arnal, Directeur Cinéma, 2, avenue de Taillebourg, Paris (11^e)
 Arna (Jacques), Artiste, 76, rue des Petits-Champs (2^e)
 Arnold, Artiste, 48, rue Condorcet (9^e)
 Arnold (Agence Roger), Attractions, 8, Boulev. Bonne-Nouvelle (10^e)
 Arnou (André), Opérateur de prise de vues, 19, avenue Bellevue, Bry-
 sur-Marne (Seine).
 Arnou (Maurice), Opérateur de prise de vues, 11, rue Louis-Braille (12^e)
 Arnoux (Jeanne), Artiste, 4, rue Clairant (17^e)

Trud. 60-73
 Centr. 48-32
 Trud. 00-60
 Argent. 1-60
 Argent. 3-73
 Argent. 0-34
 Nord 12-70
 Gob. 43-59
 Elys. 62-58
 Arch. 30-55
 Gob. 49-84
 Roq. 02-57
 Gob. 19-29
 Gut. 27-75
 Gut. 27-20
 Nord 84-02
 Gut. 18-07
 Centr. 35-84
 Lab. 26-20
 Marc. 10-46
 Marc. 29-32
 Wagl. 35-21
 Pass. 27-40
 Centr. 84-37
 Prov. 36-35



Liste Alphabétique 27

Trud. 60-73.

Andreyor (Yvette), Artiste, 31, rue Victor-Massé (9e).

Andwig (Jens), Acheteur de films pour la Norvège, à Christiana, Norvège.

Angelo (Jean), Artiste, 11, boulevard Montparnasse (15e).

Angély, Artiste, 68, rue Monge (5e).

Anglo Foreign Général Commerce Co, 5, rue de Provence (9e).

Annales (Les), Hebdomadaire, 5, rue Labruyère (9e). Réd. cinégr.

Mlle Brézillon.

Argenteuil (Cinéma Moderne d').

Argenteuil (Lutétia d').

Argenteuil (Casino-Cinéma d').

Anonges, Artiste, 128, rue d'Aboukir (2e).

Ansaldo, Sièges et Fauteuils, 115, 117, rue Sainte, Marseille (B.-du-R.).

Antillan-Film Co, Achat de films pour les Antilles, Aguila, La

Havane, Cuba.

Antiseptique Oliflor, Assainissement, 86, rue du Président-Wilson,

La Plaine Saint-Denis, Seine.

Antoine (André), O. ♦, Critique dramatique, 28, place Dauphine (1er).

Apers, Photographe pour artistes, 23, rue Boissy-d'Anglas (8e).

Appareillages électriques Grivolos, 16, rue Montgolfier (3e).

Appareillages électriques Supra, 15, boulevard Saint-Germain (5e).

Appareils contrôleurs (Sté Univers. des), Construct. mécan. 44, rue

de Chanzy (11e).

Appignan et Pinotti, Acheteurs de films pour l'Italie, 210, via Tritone,

Rome, Italie.

Aps (Félix d'), Artiste, 1, rue des Ursins (4e).

Arango et Salvador, Acheteurs de films pour la République de

Colombie, 51, rue de Paradis (10e).

09-76

Nord

Gob.

03-59.

62-58.

Elys.

30-55.

Arch.

Gob.

49-64.

Roq. 02-57.

Gob. 19-29.

Gut. 27-75.

[...]

Chaîne de traitement actuelle

images → tables

Nous utilisons une chaîne de traitement en 3 étapes

Étape 1

Détection de la **mise en page** des documents

Module dédié ou intégré à l'OCR.

Étape 2

Transcription du texte (OCR).

Nous utilisons PERO OCR.

Étape 3

Extraction de données structurées.

Plusieurs solutions possibles, domaine en évolution très rapide.

→ *plus de détails...*

État actuel

Méthode

Basée sur l'expérience collective et interdisciplinaire SoDUCo

Tournée vers les petits corpus (pour l'instant)

Permet la production de données compatibles avec une approche quantitative en premier lieu : calculs de statistiques, de tendances...

Mais une approche plus qualitative est possible également : identification de sous-populations, détection d'éléments inattendus

Outils

Images : compatibilité IIIF large, Gallica++

Détection de mise en page : outil SoDUCo encore, offre à compléter

Image → texte : PERO OCR, offre à compléter

Texte → données structurées : API Mistral, offre à compléter

Interface utilisateur (Corpusense) : en stabilisation, APIs à formaliser

Projet “Les femmes dans les métiers du cinéma français”

Programme de recherche « Les femmes dans les métiers du cinéma français 1920-1960 »



{BnF

ARCHIVES
NATIONALES



Objectif : Sortir les professionnelles du cinéma français de leur invisibilité les compter, les nommer, raconter leur histoire

- **Les métiers de l'ombre**, de l'écriture du scénario à la critique et la conservation des films, en passant par les métiers techniques, la distribution, l'exploitation.
- **Un programme de recherche utilisant l'IA** pour la fouille de sources imprimées numérisées (annuaires professionnels et presse)
- **Porté par la BnF, Soutenu par l'EUR ArTeC** sur 2 ans (2025-2026)
- **Amorce d'un projet ANR**

Forte dimension collective (50 participants à ce jour)

- **4 institutions patrimoniales** / 24 pers
- **2 laboratoires en IA**, cellule IA et DataLab de la BnF / 10 pers
- **16 référents universitaires en études cinématographiques** répartis par métiers du cinéma / 6 universités françaises, Ecole des chartes, 2 universités américaines (Women's Film Pioneer Project)

Calendrier

- **2025 : traitement par IA**, croisement avec des dossiers IDHEC/CNC et listes
- **2026 : recherche dans les archives** à partir des noms extraits des traitements
- **Fin 2026 : colloque et exposition itinérante sur panneaux**



Programme de recherche « Les femmes dans les métiers du cinéma français 1920-1960 »

Corpus des annuaires professionnels du cinéma

- Sélection de 10 annuaires / 3 titres conservés série / numérisation par la BnF
- Sources trop volumineuses pour dépouillement manuel (1 300 à 1500 pages/annuaire)
- Structurées sous forme de listes associant noms, adresses et métiers (liste générale et listes par chapitres de métier)
- Besoin : extraire les noms des femmes par métier pour alimenter la création d'une base de données + compter la part des femmes
- Premiers tests sur l'Annuaire général de la cinématographie et les industries qui s'y rattachent, 1927
- Sélection des chapitres à traiter : métiers du cinéma, hors fournisseurs non spécialisés

