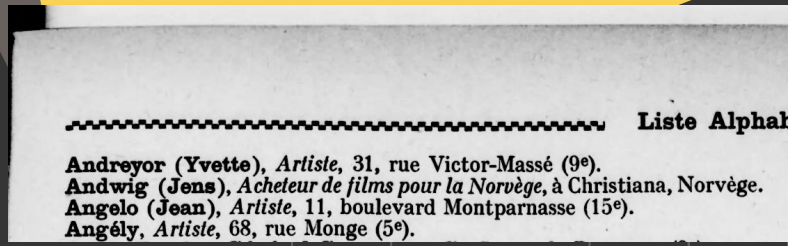
Après-midi d'étude "Archives et femmes de cinéma"
Cinémathèque française / Univ. Paris 8
12 mai 2025

Des images d'annuaires aux tables de personnes


Liste Alphab

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).

	A	B	C	D	E
1	**Nom**	**Prénom**	**Sexe/Genre**	**Activité**	**Adresse**
2	Andreyor	Yvette	F	Artiste	rue Victor-Massé (9e)
3	Andwig	Jens	?	Acheteur de films pour la Norvège	Christiana, Norvège
4	Angelo	Jean	M	Artiste	11, boulevard Montparnasse (15e)
5	Angély	Artiste	?	Artiste	68, rue Monge (5e)
6					

Laboratoire de Recherche de l'EPITA (LRE)
Joseph Chazalon

Après-midi d'étude "Archives et femmes de cinéma"
Cinémathèque française / Univ. Paris 8
12 mai 2025

Réutiliser cette présentation

Le projet Mezanno

BnF Bibliothèque nationale de France

L'ÉCOLE DES HAUTES ÉTUDES EN SCIENCES SOCIALES

EPITA SCHOOL OF ENGINEERING & COMPUTER SCIENCE

IGN INSTITUT NATIONAL DE L'INFORMATION GÉOGRAPHIQUE ET FORESTIÈRE

BnF
2 personnes

EHESS
2 personnes

EPITA
3 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**.

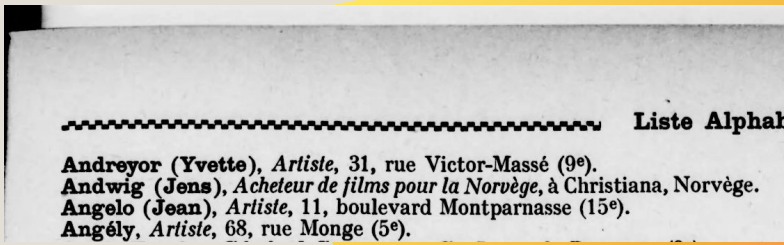
https://mezanno.xyz

Le projet "**Les femmes dans les métiers du cinéma français**" est notre **cas d'application privilégié**.

Mission confiée au LRE :

Annuaires numérisés

↳Tables de personnes



Liste Alphab

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).

	A	B	C	D	E
1	**Nom**	**Prénom**	**Sexe/Genre**	**Activité**	**Adresse**
2	Andreyor	Yvette	F	Artiste	rue Victor-Massé (9e)
3	Andwig	Jens	?	Acheteur de films pour la Norvège	Christiana, Norvège
4	Angelo	Jean	M	Artiste	11, boulevard Montparnasse (15e)
5	Angély	Artiste	?	Artiste	68, rue Monge (5e)
6					

Pourquoi est-ce nécessaire ?

- Les documents sont sous forme **image** dans Gallica.
- L'**Océrisation** n'est pas toujours de bonne qualité.
- Le **texte brut** ne permet pas de faire directement des analyses (comptages, regroupements).
- Le sexe/genre des personnes n'est pas **explicite**.
- La **masse de données** rend la saisie manuelle peu réaliste.

4 grandes questions

Notre présentation s'articule autour de **quelques-unes** des grandes questions liées à la **mise en place de ce projet**

1

Se mettre d'accord

Comment construit-on une **problématique transdisciplinaire**, avec archivistes, chercheurs en SHS, informaticiens… ?

2

Expérimenter

Quelle **chaîne de traitement de données** avons-nous mise en place, et quel rôle avons-nous donné à l'IA ?

3

Valider (1)

Au service de quelles **méthodes de travail** les outils développés se positionnent-ils ?

4

Valider (2)

Comment garantir la **pertinence des données produites**, tant du point de vue de leur intérêt scientifique que de leur fiabilité ?

4 grandes questions

Notre présentation s'articule autour de **quelques-unes** des grandes questions liées à la **mise en place de ce projet**

1

Se mettre d'accord

Comment construit-on une **problématique transdisciplinaire**, avec archivistes, chercheurs en SHS, informaticiens… ?

2

Expérimenter

Quelle **chaîne de traitement de données** avons-nous mise en place, et quel rôle avons-nous donné à l'IA ?

3

Valider (1)

Au service de quelles **méthodes de travail** les outils développés se positionnent-ils ?

4

Valider (2)

Comment garantir la **pertinence des données produites**, tant du point de vue de leur intérêt scientifique que de leur fiabilité ?

Construire une problématique transdisciplinaire

Nous sommes dans le cadre d'un projet **d'analyse de données.**

Les GLAM* déterminent les **corpus disponibles**.

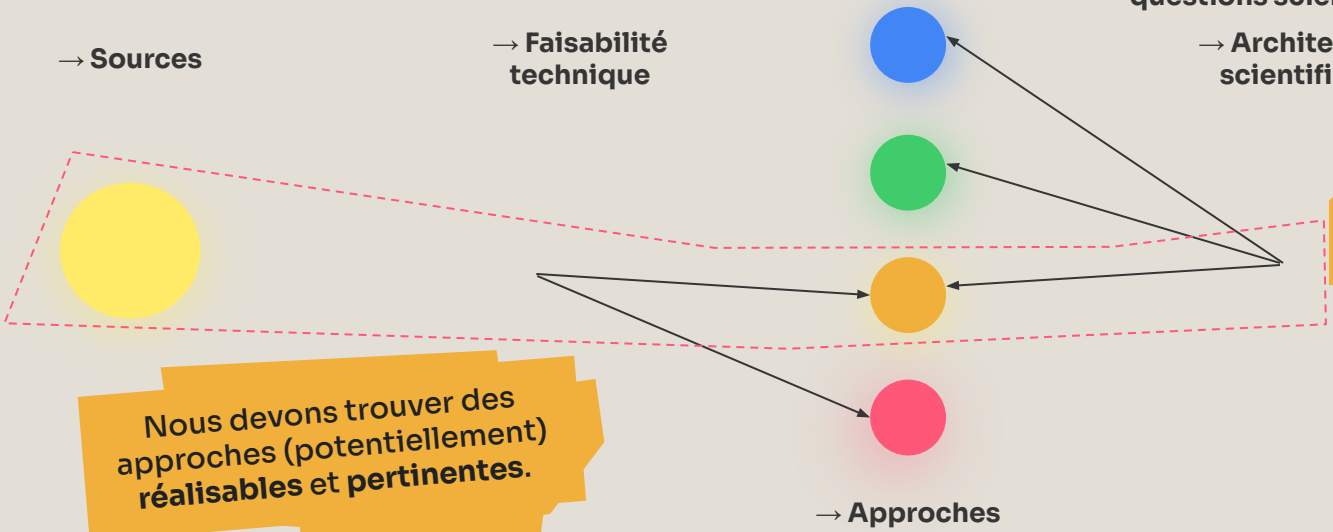
→ **Sources**

Les "traiteurs de données" proposent des **méthodes automatiques.**

→ **Faisabilité technique**

Les chercheurs en SHS (histoire, socio, politique, économie...) posent des **questions scientifiques**.

→ **Architecture scientifique**



Nous cherchons une **chaîne de production** de valeur / données.

Nous devons trouver des approches (potentiellement) **réalisables** et **pertinentes.**

→ **Approches**

Gestion du projet

Disponibilité des expertises

Moyens techniques et financiers

*: *Galleries, Libraries, Archives, Museums*

4 grandes questions

Notre présentation s'articule autour de **quelques-unes** des grandes questions liées à la **mise en place de ce projet**

1

Se mettre d'accord

Comment construit-on une **problématique transdisciplinaire**, avec archivistes, chercheurs en SHS, informaticiens… ?

2

Expérimenter

Quelle **chaîne de traitement de données** avons-nous mise en place, et quel rôle avons-nous donné à l'IA ?

3

Valider (1)

Au service de quelles **méthodes de travail** les outils développés se positionnent-ils ?

4

Valider (2)

Comment garantir la **pertinence des données produites**, tant du point de vue de leur intérêt scientifique que de leur fiabilité ?

Chaîne de traitement “idéale”

Étape 1

Recenser les sources.

Étape 2

Numériser les sources et organiser les images.

Étape 3

Extraction de données structurées à partir des images.

Étape 4

Injection des données dans un modèle métier, pour identifier les individus, dédoublonner les entrées...

Étape 5

Analyser les données produites pour répondre aux questions scientifiques originales.

But final

Publier la méthode et les résultats, archiver les outils et les données produites, former à leur réutilisation...

Nous en sommes là...

Chaîne de traitement provisoire

images → tables

(étape 3)

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.

Andreyor (Yvette), *Artiste*, 31, rue Victor-Massé (9e). Trud. 60-73.
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). Centr. 48-32.
Annales (Les), *Hebdomadaire*, 5, rue Labruyère (9e). Réd. cinégr. Mlle 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 (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. Nord 09-76
Antoine (André), O. ✻, *Critique dramatique*, 28, place Dauphine (1er). Gob. 03-59.
Apers, *Photographe pour artistes*, 23, rue Boissy-d'Anglas (8e). Elys. 62-58.
Appareillages électriques Grivolas, 16, rue Montgolfier (3e). Arch. 30-55.
Appareillages électriques Supra, 15, boulevard Saint-Germain (5e). Gob. 49-64.
Appareils contrôleurs *(Sté Univers. des)*, Construct. mécan. 44, rue de Chanzy (11e). 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 (4e). Gob. 19-29.
Arango et Salvador, *Acheteurs de films pour la République de Colombie*, 51, rue de Paradis (10e). Gut. 27-75. — 27-26.
Arbid (David), *Acheteur de films* 11, rue Mariette-Pacha, Alexandrie, Egypte.
Arcos, 1, rue Taitbout (9e).
Arditi, *Vente et achat de films*, 5, rue Bouchardon (10e). 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 (9e). Gut. 18-07. Centr. 55-84.
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 (14e).
Aristocratia del film (La), *Acheteurs de films*, 45, rue Laborde (8e). Lab. 26-20.
Arlix (W. d'), *Artiste*, 27, rue du Château, Fontainebleau (S.-et-M.).
Arly (Jacqueline), *Artiste*, 30, rue du Général-Foy (8e).
Armand (Aug.), *Administrateur du Gaumont-Palace*, 3, rue Caulaincourt (18e). Marc. 00-46.
Armand, *Ameublement*, 109, rue Lamarck (18e). Marc. 29-92.
Armand (Georgette), *Artiste*, 35, rue Jouffroy (17e). Wag. 95-21.
Armand (Paul Loys), *Décorateur*, 20, rue Ravignan (18e).
Armel (Yette), *Artiste*, 160, boulevard Malesherbes (8e).
Armont (Paul), *Artiste*, 82, boulevard Flandrin. Pass. 27-40.
Armor (Les Films), *Location*. Directeur : Marcel Sprecher, 12, rue Gaillon (2e). Centr. 84-37.
Arnal, *Directeur Cinéma*, 2, avenue de Taillebourg, Paris (11e).
Arnna (Jacques), *Artiste*, 76, rue des Petits-Champs (2e).
Arnold, *Artiste*, 48, rue Condorcet (9e).
Arnold (Agence Roger), *Attractions*, 8, Boulev. Bonne-Nouvelle (10e) 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 (12e).
Arnoux (Jeanne), *Artiste*, 4, rue Clairant (17e).

Liste Alphabétique 27

Andreyor (Yvette), Artiste, 31, rue Victor-Massé (9e) Trud. 60-73.
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) Centr. 48-32.
Annales (Les), Hebdomadaire, 5, rue Labruyère (9e). Réd. cinégr. Mlle 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 (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. Nord 09-76
Antoine (André), O. ✻, Critique dramatique, 28, place Dauphine (1er). Gob. 03-59.
Apers, Photographe pour artistes, 23, rue Boissy-d'Anglas (8e). Elys. 62-58.
Appareillages électriques Grivolas, 16, rue Montgolfier (3e). Arch. 30-55.
Appareillages électriques Supra, 15, boulevard Saint-Germain (5e). Gob. 49-64.
Appareils contrôleurs (Sté Univers. des), Construct. mécan. 44, rue de Chanzy (11e). 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 (4e). Gob. 19-29.
Arango et Salvador, Acheteurs de films pour la République de Colombie, 51, rue de Paradis (10e). Gut. 27-75. — 27-26.
Arbid (David), Acheteur de films 11, rue Mariette-Pacha, Alexandrie, Egypte.
Arcos, 1, rue Taitbout (9e).
Arditi, Vente et achat de films, 5, rue Bouchardon (10e). 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 (9e). Gut. 18-07. Centr. 55-84.
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 (14e).
Aristocratia del film (La), Acheteurs de films, 45, rue Laborde (8e). Lab. 26-20.
Arlix (W. d'), Artiste, 27, rue du Château, Fontainebleau (S.-et-M.).
Arly (Jacqueline), Artiste, 30, rue du Général-Foy (8e).
Armand (Aug.), Administrateur du Gaumont-Palace, 3, rue Caulaincourt (18e). Marc. 00-46.
Armand, Ameublement, 109, rue Lamarck (18e). Marc. 29-92.
Armand (Georgette), Artiste, 35, rue Jouffroy (17e). Wag. 95-21.
Armand (Paul Loys), Décorateur, 20, rue Ravignan (18e).
Armel (Yette), Artiste, 160, boulevard Malesherbes (8e).
Armont (Paul), Artiste, 82, boulevard Flandrin. Pass. 27-40.
Armor (Les Films), Location. Directeur : Marcel Sprecher, 12, rue Gaillon (2e). Centr. 84-37.
Arnal, Directeur Cinéma, 2, avenue de Taillebourg, Paris (11e).
Arnna (Jacques), Artiste, 76, rue des Petits-Champs (2e).
Arnold, Artiste, 48, rue Condorcet (9e).
Arnold (Agence Roger), Attractions, 8, Boulev. Bonne-Nouvelle (10e) 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 (12e).
Arnoux (Jeanne), Artiste, 4, rue Clairant (17e).

Chaîne de traitement provisoire

images → tables

(étape 3)

Nous utilisons une chaîne de traitement en 3 étapes

Étape 1

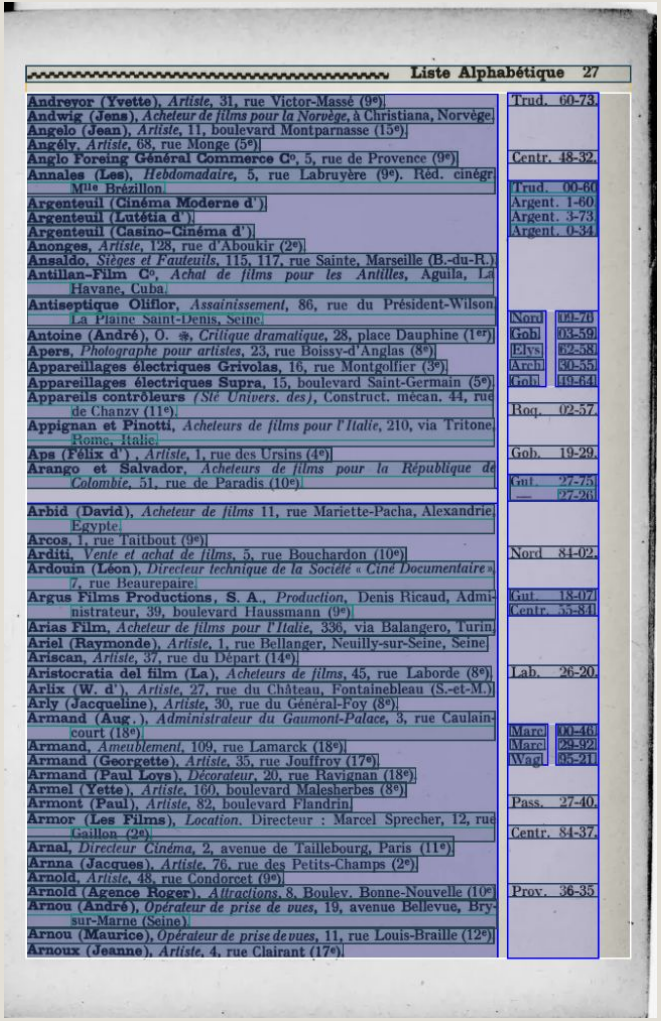
Détection de la **mise en page** des documents

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 (9e)| Trud. 60-73.
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)| Centr. 48-32.
Annales (Les), *Hebdomadaire*, 5, rue Labruyère (9e). Réd. cinégr. Mlle 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 (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| Nord 09-76
Antoine (André), O. ✻, *Critique dramatique*, 28, place Dauphine (1er)| Gobl 03-59
Apers, *Photographe pour artistes*, 23, rue Boissy-d'Anglas (8e)| Elys 62-58
Appareillages électriques Grivolas, 16, rue Montgolfier (3e)| Arch 30-55
Appareillages électriques Supra, 15, boulevard Saint-Germain (5e)| Gobl 49-64
Appareils contrôleurs *(Sté Univers. des)*, Construct. mécan. 44, rue de Chanzy (11e)| Roq. 02-57.
Appignan et Pinotti, *Acheteurs de films pour l'Italie*, 210, via Tritone, Rome, Italie| Gob. 19-29.
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)| Gut. 27-75.
— 27-26
Arbid (David), *Acheteur de films* 11, rue Mariette-Pacha, Alexandrie, Egypte|
Arcos, 1, rue Taitbout (9e)|
Arditi, *Vente et achat de films*, 5, rue Bouchardon (10e)| 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 (9e)| Gut. 18-07
Centr. 55-81
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 (14e)|
Aristocratia del film (La), *Acheteurs de films*, 45, rue Laborde (8e)| Lab. 26-20.
Arlix (W. d'), *Artiste*, 27, rue du Château, Fontainebleau (S.-et-M.)|
Arly (Jacqueline), *Artiste*, 30, rue du Général-Foy (8e)|
Armand (Aug.), *Administrateur du Gaumont-Palace*, 3, rue Caulaincourt (18e)| Marc 00-46
Armand, *Ameublement*, 109, rue Lamarck (18e)| Marc 29-92
Armand (Georgette), *Artiste*, 35, rue Jouffroy (17e)| Wagl 95-21
Armand (Paul Loys), *Décorateur*, 20, rue Ravignan (18e)|
Armel (Yette), *Artiste*, 160, boulevard Malesherbes (8e)|
Armont (Paul), *Artiste*, 82, boulevard Flandrin| Pass. 27-40.
Armor (Les Films), *Location*. Directeur : Marcel Sprecher, 12, rue Gaillon (2e)|
Arnal, *Directeur Cinéma*, 2, avenue de Taillebourg, Paris (11e)| Centr. 84-37.
Arnna (Jacques), *Artiste*, 76, rue des Petits-Champs (2e)|
Arnold, *Artiste*, 48, rue Condorcet (9e)|
Arnold (Agence Roger), *Attractions*, 8, Boulev. Bonne-Nouvelle (10e)| 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 (12e)|
Arnoux (Jeanne), *Artiste*, 4, rue Clairant (17e)|

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").
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 C°, 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.✻4, Critique dramatique, 28, place Dauphine (1er).
Apers, Photographe pour artistes, 23, rue Boissy-d'Anglas (8e).
Appareillages électriques Grivolas, 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 provisoire

images → tables

(étape 3)

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...

Extraire des données structurées

	Approches possibles	Atouts	Limites
1	Expressions régulières, recherche de motifs dans le texte	Technologie simple. Calculs très rapide.	
2			
3			

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	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 extractifs (BERT-like)	Meilleure performance en théorie. Calculs assez rapides.	
3			

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 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)*

```
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**

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, arr

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	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 extractifs (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 génératifs (ChatGPT, Mistral…) avec **prédictions structurées** (nouveauté).	Pas besoin de données d'entraînement. Utilisation de modèles peu onéreux en ligne (facile à prototyper).	

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 ..."*

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 extractifs (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 génératifs (ChatGPT, Mistral…) avec **prédictions structurées** (nouveauté).	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.*

4 grandes questions

Notre présentation s'articule autour de **quelques-unes** des grandes questions liées à la **mise en place de ce projet**

1

Se mettre d'accord

Comment construit-on une **problématique transdisciplinaire**, avec archivistes, chercheurs en SHS, informaticiens… ?

2

Expérimenter

Quelle **chaîne de traitement de données** avons-nous mise en place, et quel rôle avons-nous donné à l'IA ?

3

Valider (1)

Au service de quelles **méthodes de travail** les outils développés se positionnent-ils ?

4

Valider (2)

Comment garantir la **pertinence des données produites**, tant du point de vue de leur intérêt scientifique que de leur fiabilité ?

Méthodes de travail possibles

Cas 1 : Approche statistique

Collecter un grand nombre de données, et utiliser des modèles statistiques pour dégager de grande tendances en neutralisant les erreurs. Les approches numériques permettent de changer d' échelle.

Cas 2 : Usage ciblé

Accélérer la transcription précise de données "précieuses" en petite quantité, qui seront ensuite considérées comme fiables et utilisées comme partie dans une étude plus large. Gain modéré.

Et bien d'autres à élaborer :

Découverte d'anomalies, "Discussion" avec sa collection...

4 grandes questions

Notre présentation s'articule autour de **quelques-unes** des grandes questions liées à la **mise en place de ce projet**

1

Se mettre d'accord

Comment construit-on une **problématique transdisciplinaire**, avec archivistes, chercheurs en SHS, informaticiens… ?

2

Expérimenter

Quelle **chaîne de traitement de données** avons-nous mise en place, et quel rôle avons-nous donné à l'IA ?

3

Valider (1)

Au service de quelles **méthodes de travail** les outils développés se positionnent-ils ?

4

Valider (2)

Comment garantir la **pertinence des données produites**, tant du point de vue de leur intérêt scientifique que de leur fiabilité ?

Evaluer

Il faut mettre en place un **protocole d'évaluation** à plusieurs niveaux

Utilité : Chaque étape de la chaîne de traitement produit-elle les **données attendue** par l'étape suivante ?

Utilisabilité : La **qualité "factuelle"** des données produites est-elle suffisante ? Y a-t-il des biais problématiques ?

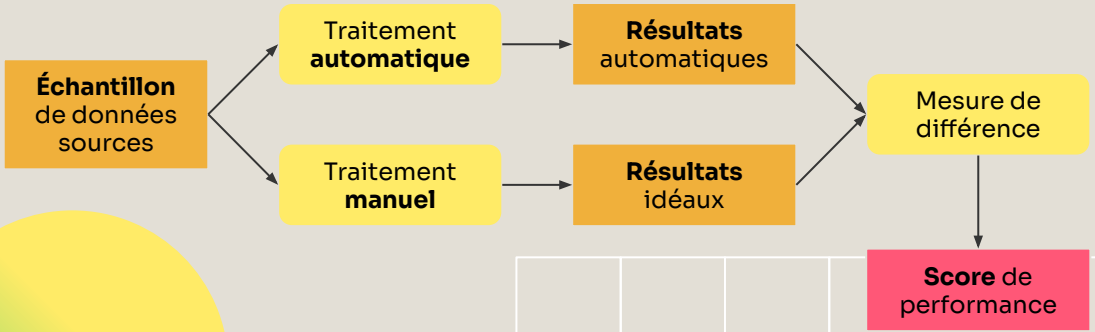
Acceptabilité : La **démarche** scientifique est-elle correcte ? Permet-elle de dépasser des lieux communs ? L'usage de l'IA est-il réfléchi et maîtrisé ?

Qualité des données

(focus sur l'utilisabilité)

Comment mesurer la performance d'un système automatique ?

Le "schéma à 7 boîtes". *Bords droits : données ; arrondis : processus.*



Les partenaires doivent s'accorder sur la **performance plancher.**

En conclusion

Un exemple de projet en construction

Notre objectif est de **valider la faisabilité** des étapes préliminaires pour monter un projet plus ambitieux.

Les étapes les plus avancées ne sont pas intégrées à cette expérimentation.

Un cas de projet d'analyse de données

Un exemple de projet pluridisciplinaire courant, avec un défi d'**alignement des objectifs** des partenaires.

Il faut **connecter les maillons** de la chaîne de création de valeur au plus vite et valider les résultats produits.

Une chaîne de traitement avec IA

L'IA est un outil qui **automatise une tâche de traitement de données.**

Il existe **plusieurs variantes** de ces outils, à combiner idéalement.