



CONFÉRENCE
MNESYS EXPO

|| Point d'étape du projet SOCFACE

SOC FACE



Projet de recherche collaborative 2021-2025

Historiens, économistes, démographes

Archivistes

Experts en IA



Avec la participation des services d'archives municipales et départementales

Financé par



Objectifs et challenges :

Les individus face au changement social : analyse de la structure sociale en France sur un siècle 1836-1936



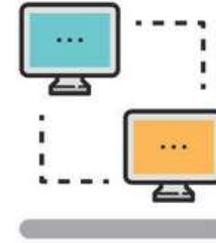
Technologie

Reconnaissance automatique d'écriture manuscrite, analyse de tableau, traitement de plusieurs million d'images, accès à de multiples sources.



Science

Etudes en économie historique et démographie : évolution du marché du travail, de l'urbanisme, de la structure sociale, des mobilités.



Valorisation

Mise à disposition des données extraites en accès libre*, versement aux propriétaires des fonds, valorisation par les portails.

Pourquoi ?

Importance croissante des données micro (individuelles) dans la recherche quantitative en sciences sociales: économie, histoire, sociologie, démographie...

Progrès considérables dans les technologies de reconnaissance d'écriture manuscrite et de traitement d'images.

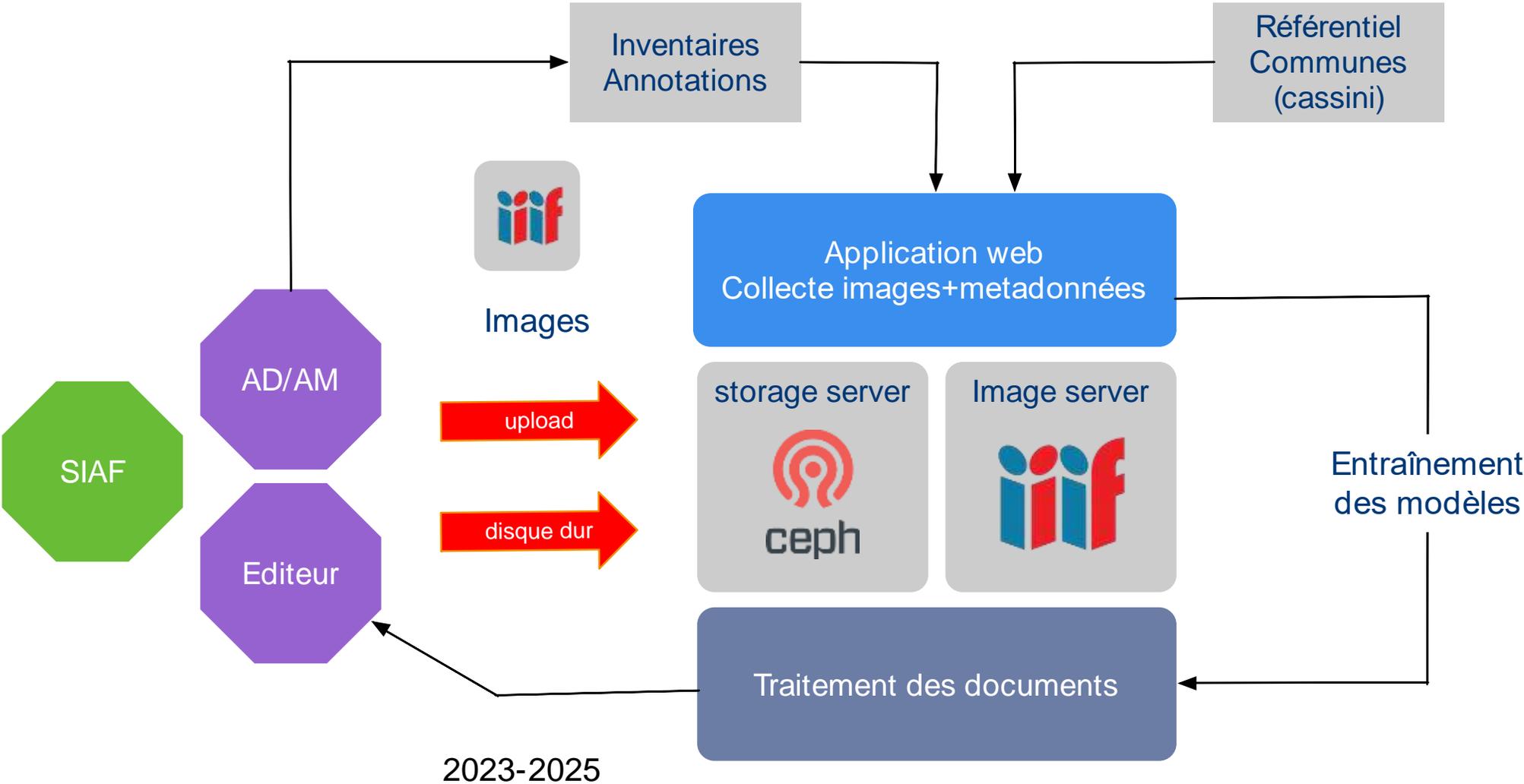
Objectifs

Collecter, traiter, retranscrire, organiser et analyser l'ensemble des listes nominatives du recensement de 1836 à 1936 (20 recensements).

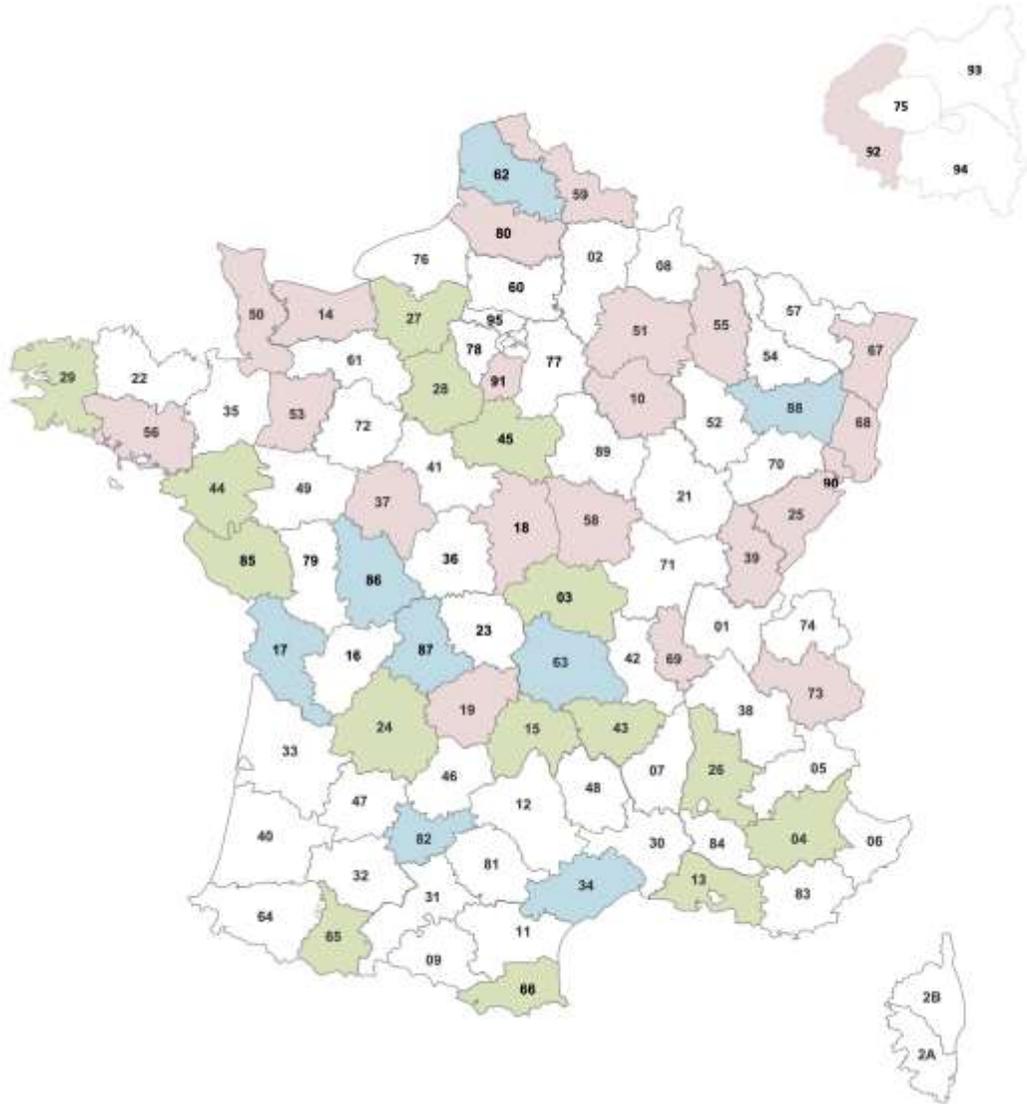
Livrables

SocFace produira une base de données complète des individus ayant vécu en France entre 1836 et 1936 et l'utilisera pour analyser le changement social dans la longue durée

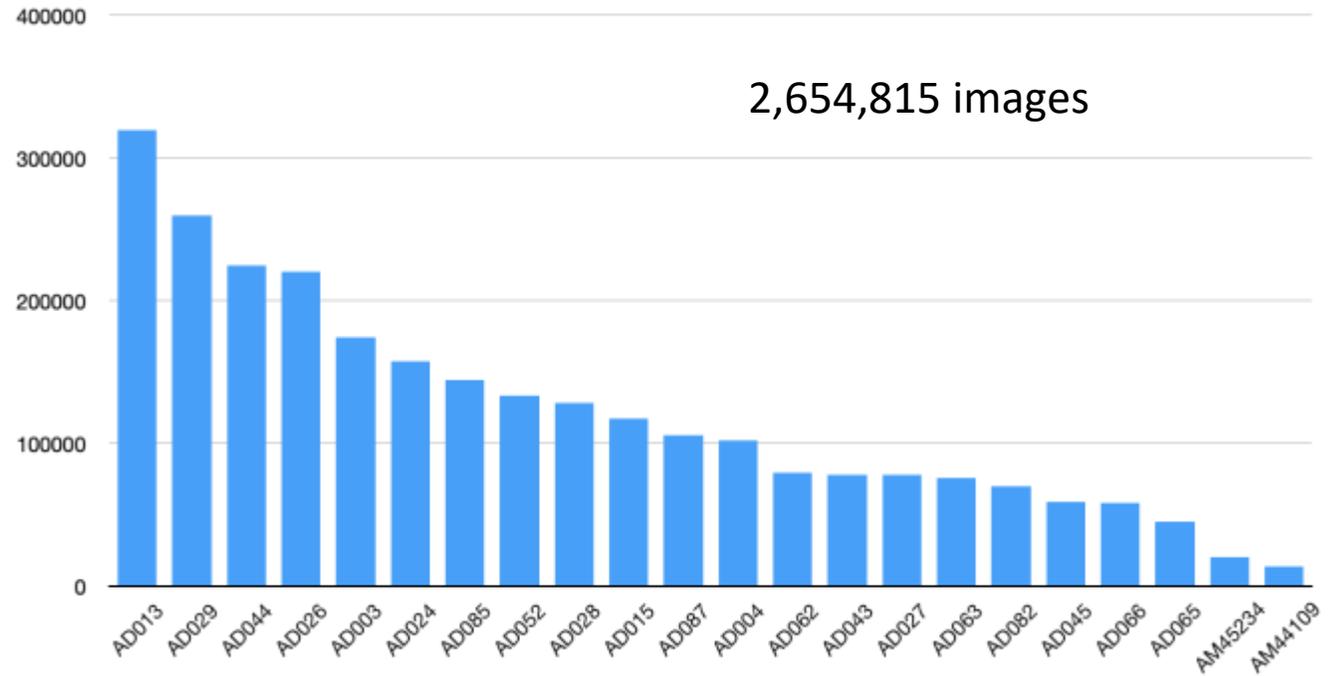
Collecte des images+meta-données



Progression de la collecte



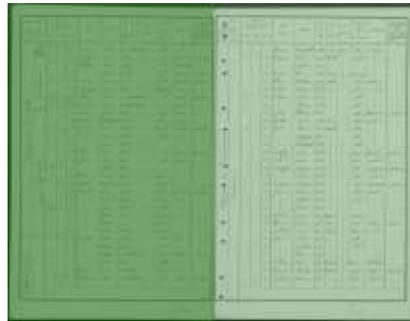
- 15 AD ont transmis leurs fichiers (images+méta-données)
- 9 AD sont en cours de transfert
- 24 AD sont en phase de préparation des données



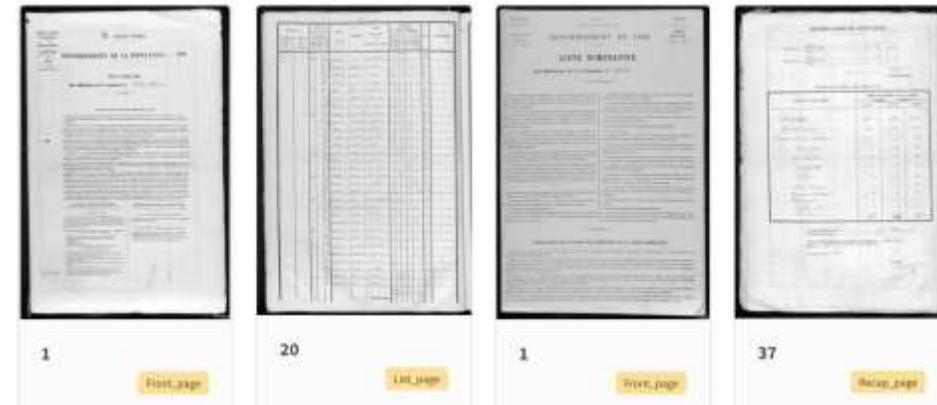
Projection France: 15M d'images, 30To

Traitement automatique

1. Découpage des doubles pages



2. Typages des pages



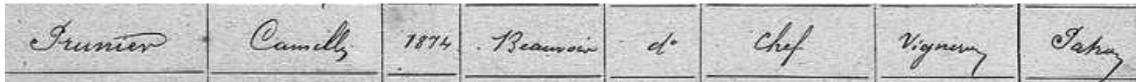
3. Découpage des zones: header, designation, number, footer, subject_zone

gouffon	je	de	1	1
Douglas	de	1	1	2
Trévisan	de	1	1	3
Alain	de	1	1	4
Alain	de	1	1	5
Alain	de	1	1	6
Alain	de	1	1	7
Alain	de	1	1	8
Alain	de	1	1	9
Alain	de	1	1	10
Alain	de	1	1	11

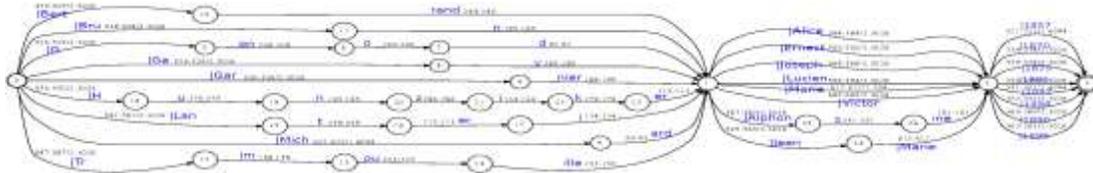
4. Découpage des lignes

Entraînement des modèles IA

Reconnaissance de texte
avec contrainte syntaxique

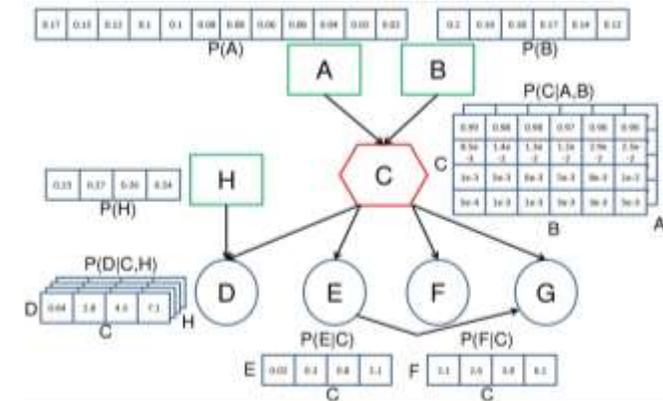


Modèle optique



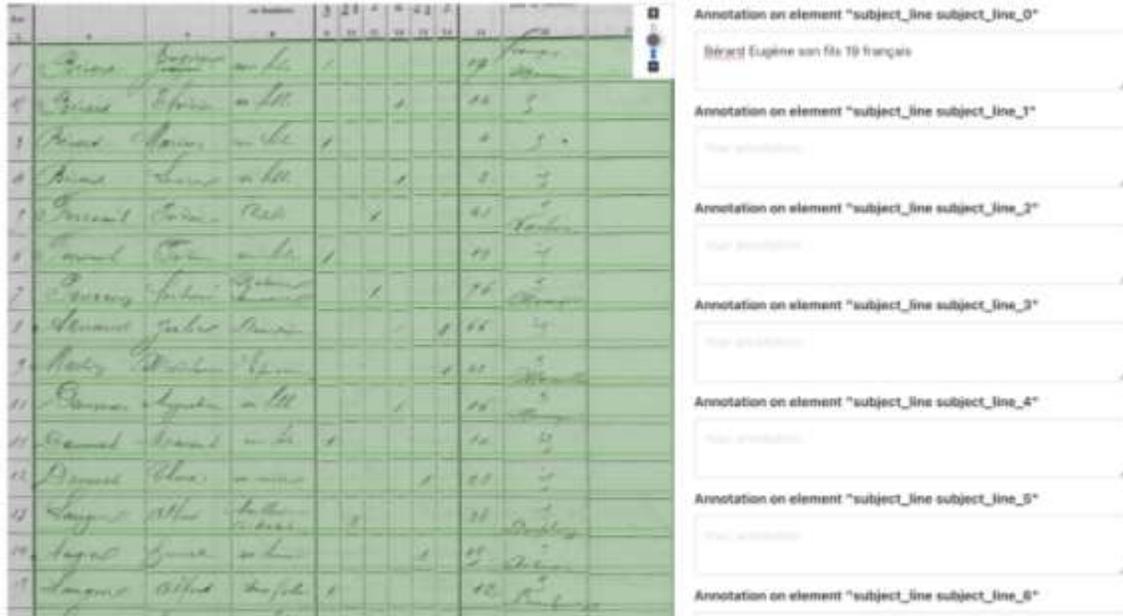
Modèle linguistique

Validation de cohérence
Avec contraintes



Données d'entraînement

Annotation dédiée



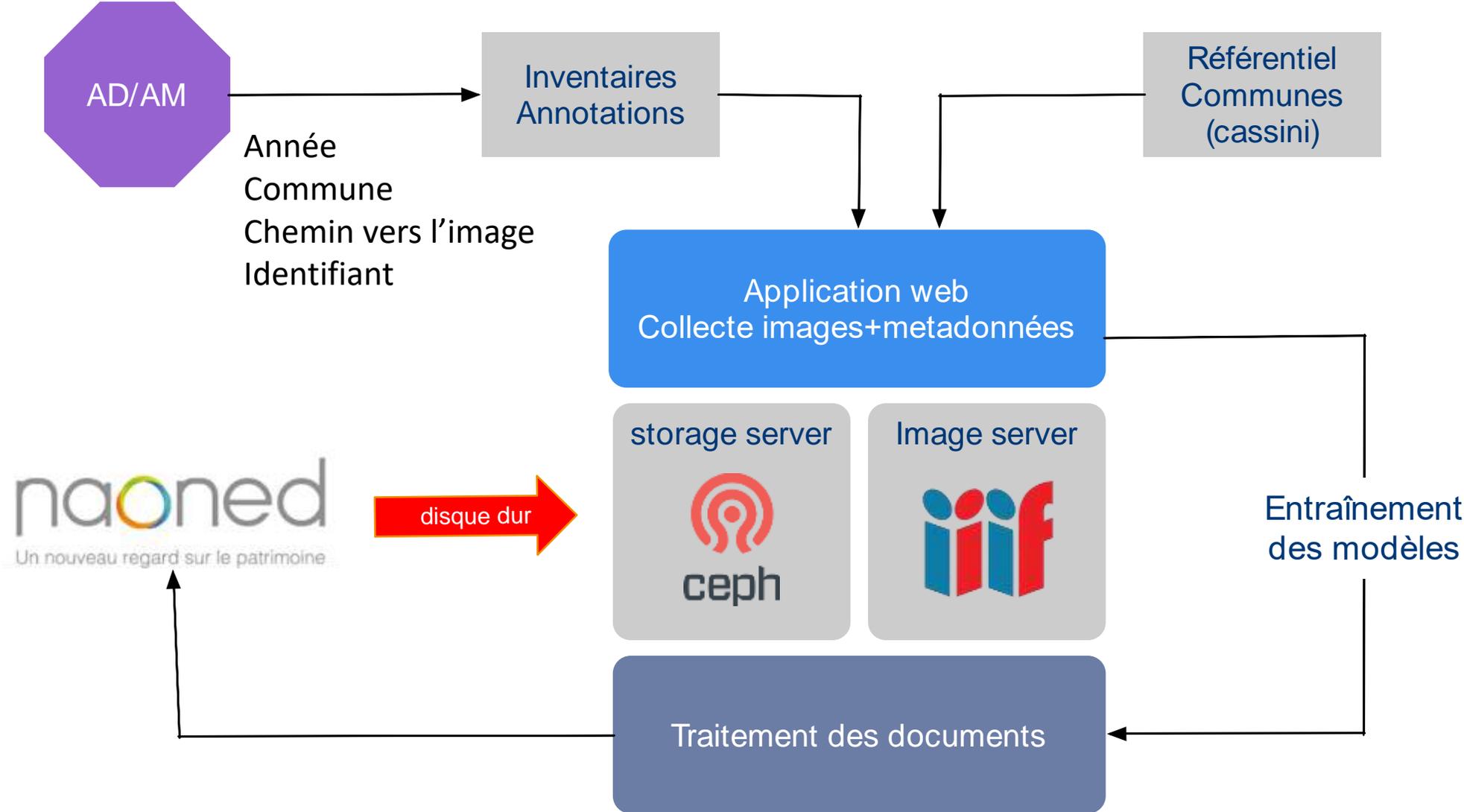
- Immédiatement exploitable par les algorithmes
- Complète
- Validable (double saisie)
- Transcription sur mesure
- Prend du temps
- Consomme du budget

Indexation collaborative



- Doit être transformée pour être exploitée
- Partielle
- De qualité variable
- Règles de transcription pas toujours adaptées
- Existe déjà
- Gratuit

Comment participer ?



Diffusion/utilisation



Développement d'une base de donnée nominative sur le portail France Archive

Service participant

The image shows a screenshot of a data entry interface. On the left, there is a sidebar with a search bar and several icons. The main area displays a table with handwritten entries. The table has columns for 'Date T.T.', 'Nom', 'Prénom', and 'Adresse'. The entries are grouped by date and name. The following table represents the data shown in the screenshot:

Date T.T.	Nom	Prénom	Adresse
-	Podet	Maria-Françoise	
-	Podet	Maria	
1	Podet	Paule	
1	Jouin	Claudine	
1	Lebrun	Claudine	
2	Grand	Maria	
2	Lebrun	Henri	
2	Lebrun	Joseph	
3	Baccus	Geneviève	
3	Baccus	Antoine	
3	Baccus	Lucienne	

Livraison des données brutes et integration dans le portail web du service