



## Un projet de recherche collaboratif entre

Historiens, économistes, démographes

Archivistes

Experts en intelligence artificielle



*Et les services d'archives départementales et municipales*

Lionel Kesztenbaum

Présentation aux services de l'Ined  
14/10/2021

SOC FACE



## Un projet de recherche collaboratif entre

Historiens, économistes, démographes

Archivistes

Experts en intelligence artificielle



FranceArchives  
PORTAIL NATIONAL DES ARCHIVES

TEKLIA

*Et les services d'archives départementales et municipales*

Financement sur **3 ans ½**

ANR  
AGENCE  
NATIONALE  
DE LA  
RECHERCHE



# Pourquoi Socface?

- 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.

**Produire des données  
individuelles en masse à  
l'échelle nationale**

**Développer les technologies de  
traitement de vaste ensemble  
de sources historiques**

**Diffuser les informations  
obtenues au grand public  
comme aux chercheurs**

DÉSIGNATION		NUMÉROS, PAR QUARTIER, VILLAGE, HOMME ou rue,			NOMS	PRÉNOMS	ANNEE	LIEU	NATIONALITÉ	SITUATION	PROFESSION	
des quartiers, villages ou communes	des rues, des villages ou communes	des numéros	des numéros	des numéros	DE FAMILLE		de naissance	de naissance	LITÉ	ou chef de ménage		
1	2	3	4	5	6	7	8	9	10	11	12	13
				2	Eribout	Lucie	1886	Reims	F	épouse	ans	
				3	Eribout	Senni	1909	Claris	"	enfant	"	
				4	Eribout	Alphonse	1912	"	"	"	"	
				5	Eribout	Reni	1916	Aubervilliers	"	"	"	
				6	Eribout	Audie	1920	"	"	"	"	
				7	Pierre	Henriette	1881	Reims	"	belle sœur	Magasin	Boulonnais Valenciennes
				8	Pierre	Liliane	1912	Bagnollet	"	Niece	ans	
				9	Pierre	Lucien	1918	Woisyl-lès-Reims	"	neveu	ans	
				1	Parmentier	Justave	1877	Weschenau	"	Chef	maçon	
				2	Parmentier	Lucie	1867	Lambach	"	épouse	ans	
				3	Parmentier	Maximilien	1888	Paris	"	enfant	Confectionner	Chulman
				4	Parmentier	Emma	1900	W Reims	"	"	empl.	Minist. Pensions
				5	Parmentier	Ludovic	1902	"	"	"	plis	Alix
				1	Caliez	Georges	1878	Lille	"	Chef	Commerçant	Magist
				2	Caliez	Céline	1884	Tache	"	épouse	ans	
				1	Renault	Jean	1869	Combrigny	"	Chef	chef équipe	Magist
				2	Renault	Christine	1875	Combrigny	"	"	"	"

Une liste nominative:

Aubervilliers, 1921

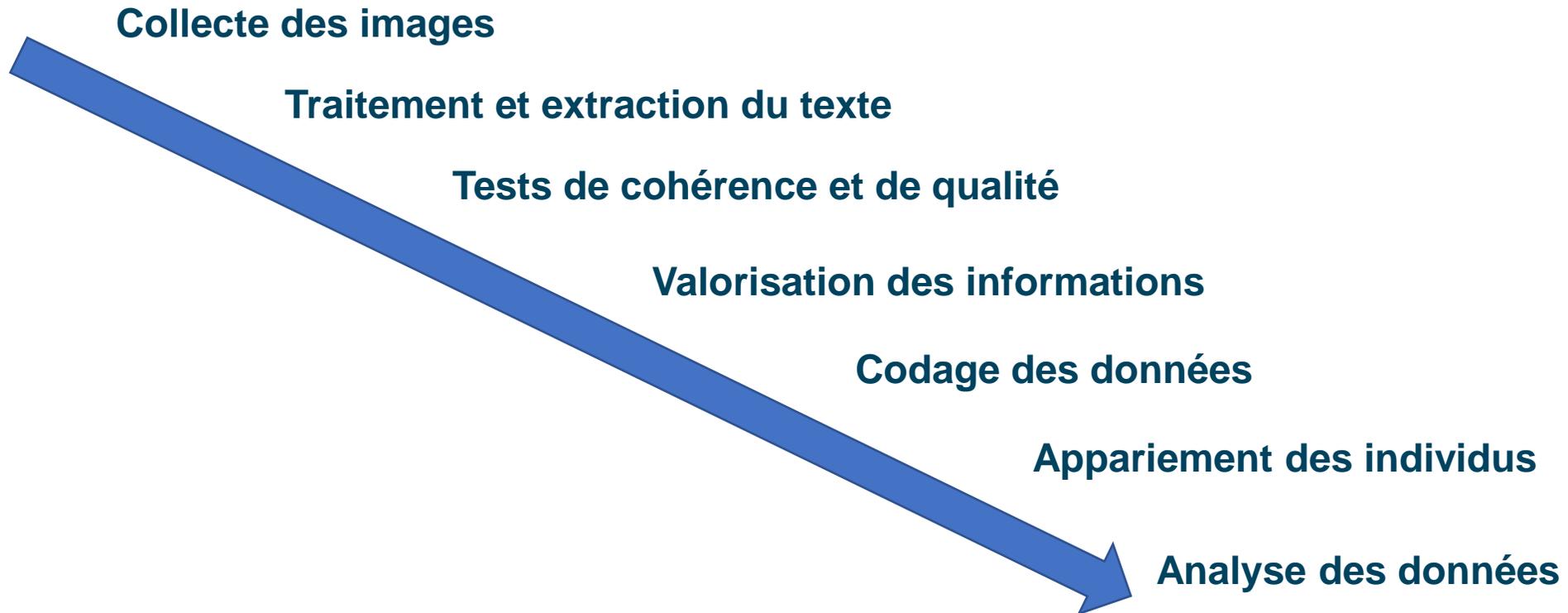
Rue  
des  
Fillettes

# Pourquoi les listes nominatives?

- Une source abondante, simple, stable et standardisée.
  - Une source régulière dans le temps.
  - Déjà photographiée par beaucoup de services d'archives.
  - Source nationale, à l'identique.
  - Permet de construire un échantillon de toute la France (ou presque...).
- Une source idéale pour le passage à l'échelle de la reconnaissance de texte.
  - Une source qui n'a de sens que étudiée à grande échelle.

The image shows a page from a historical document, likely a census or administrative list. The page is titled "Liste nominative" and contains a grid of handwritten entries. The columns are organized into several sections, with the first section containing names and addresses. The handwriting is in French and appears to be from the 19th century. The document is a source of data for linguistic research, particularly for the study of French orthography and morphology.

# Etapes et objectifs



Etude pilote: l'ensemble des étapes sur une sélection d'archives – Objectif: un an

Puis montée en charge progressive à partir d'octobre 2022

Incorporation des services d'archives peu à peu et retours au fur et à mesure

# Spécificités du projet Socface

## ➤ Traitement automatisé de vastes ensembles d'images

- ❖ Plusieurs millions d'images, de provenances variées (différents éditeurs, formats, etc.).
- ❖ Ouvre des perspectives considérables pour la recherche en intelligence artificielle.

## ➤ Une coproduction entre historiens et informaticiens

- ❖ Utiliser la structure des listes pour contrôler l'information (ordre des individus, ménages, etc.).
- ❖ Utiliser les informations agrégées (externes ou non) pour contrôler le traitement.
- ❖ Codage des informations obtenues pour les rendre accessibles aux chercheurs (lieux, professions...).

## ➤ Appariement des individus pour les suivre au cours du temps

- ❖ Première réalisation à l'échelle de la France entière, sur 100 ans.
- ❖ Ouvre des perspectives considérables pour la recherche en sciences sociales.

# De multiples objectifs et enjeux



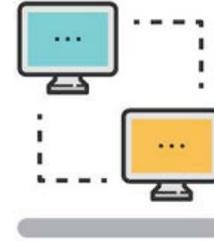
## Science

**Etudes en économie, en histoire et en démographie :** évolution du marché du travail, des inégalités, de la structure sociale, des mobilités.



## Technologie

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



## Valorisation

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

# De multiples objectifs et enjeux



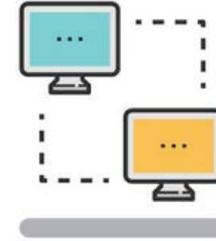
## Science

**Etudes en économie, en histoire et en démographie :** évolution du marché du travail, des inégalités, de la structure sociale, des mobilités.



## Technologie

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



## Valorisation

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

# Reconnaissance d'écriture: un des plus vieux défis de l'IA



RAND corporation, 1960



The MNIST database

“The drosophila of machine learning”

Geoffrey Hinton

Mais les machines ne rivalisent toujours pas avec l'humain pour la lecture de documents manuscrits

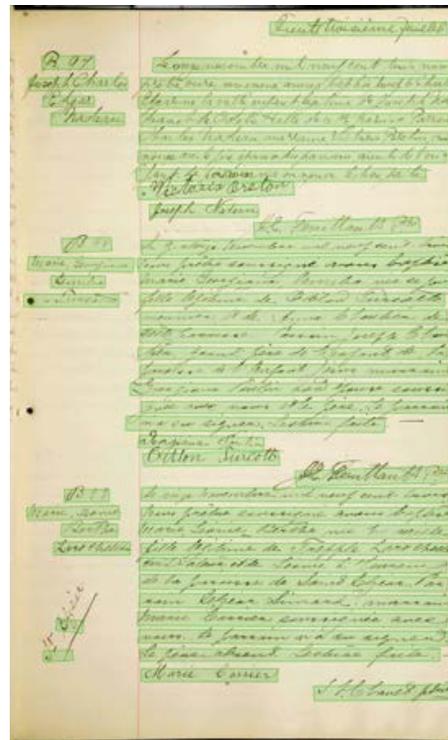
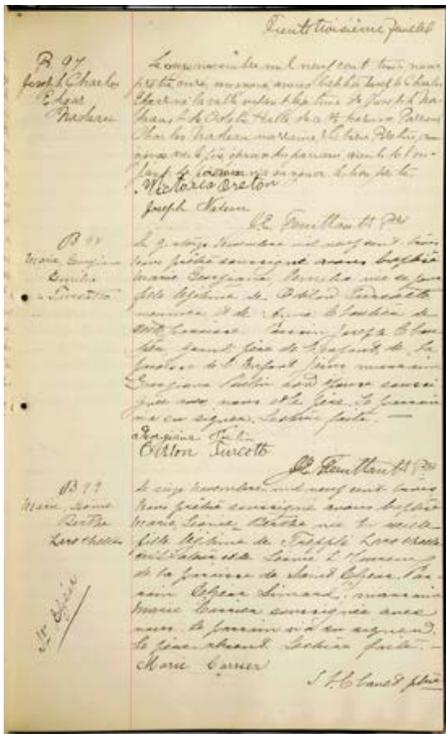
# Chaîne de traitement automatique

Image

Détection des lignes

Reconnaissance

Extraction d'entités



Le onze novembre mil neuf cent trois , nous prêtre , curé soussigné , avons baptisé Joseph Charles Edgar né la veille , enfant légitime de Joseph na deau et de Odibs Hallé , de cette paroisse . Parrain Charles Nadeau , marraine Victoria Bleton , vu n ' a signé avec le père époux du parrain , Etaient de l ' en - Vant . Le parrain n ' a su signer . Lechore faite . J . Victoria Arston Joseph Nadeau J . E . Feuilteault Ptre Le quatorze Novembre mil neuf cent trois nous prêtre soussigné avons baptisé Marie Georgiana , Emilia née ce jour fille légitime de Odilon Turcotte meunier , et de Anna Cloutier de cette paroisse . Parrain Joseph Clau tier , grand père de l ' enfant de la paroisse de l ' Enfant Jésus ; marraine Georgianna Pinaulin son épouse , soussi gné avec nous . et le père . Le parrain n ' a su signer . Lecture faite



Le [onze novembre mil neuf cent trois](#) , nous prêtre , curé soussigné , avons baptisé [Joseph Charles Edgar](#) né [la veille](#) , enfant légitime de [Joseph na](#) deau et de [Odibs Hallé](#) , de cette paroisse . Parrain [Charles Nadeau](#) , marraine [Victoria Bleton](#) , vu n ' a signé avec le père époux du parrain , Etaient de l ' en - Vant . Le parrain n ' a su signer . Lechore faite . J . Victoria Arston Joseph Nadeau J . E . Feuilteault Ptre Le [quatorze Novembre mil neuf cent trois](#) nous prêtre soussigné avons baptisé [Marie Georgiana](#) , Emilia née [ce jour](#) fille légitime de [Odilon Turcotte meunier](#) , et de [Anna Cloutier](#) de cette paroisse . Parrain [Joseph Clau tier](#) , grand père de l ' enfant de la paroisse de l ' Enfant Jésus ; marraine [Georgianna Pinaulin](#) son épouse , soussi gné avec nous . et le père . Le parrain n ' a su signer . Lecture faite

Tous les modèles sont entraînés à partir d'exemples

# Est-ce que ça fonctionne ?

The image shows a digital interface for document analysis. On the left, a handwritten document is displayed with green highlights and labels. The text is in French and appears to be a church record. On the right, a structured transcription of the document is shown, with entities like names, dates, and professions highlighted in blue and labeled with terms like 'PERSONNE', 'DATE', and 'PROFESSION'. The interface includes a sidebar with navigation icons and a top bar with the text 'Paragraphe 3'.

Paragraphe 3

CLASSIFICATIONS

TRANSCRIPTIONS

Created by Kaldi Balsac

Filter entities by worker version

Le **DATE** *un* janvier dix-neuf cent un ✓ x nous prêtre curé soussigné, avons bap-  
tisé **PRENOM** Marie Juliette **ARMOA** ✓ x née **DATE** ce jour ✓ x , fille légitime de  
**PERSONNE** Joseph Gravel ✓ x  
**PROFESSION** Cultivateur ✓ x , et de **PERSONNE** Ananda Broutard ✓ x  
de cette paroisse, le parrain a été **PERSONNE** Armand Gravel ✓ x ,  
**PROFESSION** Cultivateur ✓ x de cette paroisse, et  
la marraine a été **PERSONNE** Alphonsine Lessard ✓ x  
épouse du parrain, oncle et tante de l'en-  
fant, lesquels ont signé avec nous.  
ainsi que le père, après lecture faite  
Alphonsine Lessard  
Omer Frant le  
Joseph Gravel  
S. Turcotte Ptre

METADATA

ALL ENTITIES

9 entities

Any worker version

- 2.7M d'images de registres paroissiaux de 1850 à 1920
- HTR, segmentation en acte, détection et typage des entités
- Taux d'erreur caractère moyen : 6.4%

# Listes nominatives de recensement

## Défis du projet:

- Documents décentralisés
- Volume inconnu mais important (entre 5 et 15 M d'images ?)
- Conditions de conservation/numérisation variables
- Lecture en 2 dimensions
- Evaluer l'erreur
- Structurer, consolider, croiser l'information

# Lecture en 2 dimensions

312 — 4 —

DÉSIGNATION		NUMÉROS. PAR QUARTIER, VILLAGES, HOMMES EN TOUT.			NOMS	ANNÉE	LIEU	NATIONAL.	SITUATION	PROFESSION.		
des	des	des	des		PRÉNOMS.	de	de		PAR RAPPORT			
quartiers,	quartiers,	maisons,	maisons,	DE FAMILLE.		naissance.	naissance.	LITÉ.	ou			
1	2	3	4	5	6	7	8	9	10	11	12	
Rue des Éillettes	15	11	2	Coribout	Lucie	1886	Reims	F	épouse	ans		
			3	Coribout	Senni	1909	Paris	"	enfant	"		
			4	Coribout	Alphonse	1912	"	"	"	"		
			5	Coribout	Reni	1916	subisillies	"	"	"		
			6	Coribout	Audie	1920	"	"	"	"		
			7	Pierre	Henriette	1881	Reims	"	bell. soeur	Magasinier	Parabonville	
			8	Pierre	Liliane	1912	Bagnollet	"	filie	ans	Valenciennes	
			9	Pierre	Lucien	1918	Winghote	"	neveu	ans		
			10	Parmenier	Justave	1877	Wesphalen	"	chef	maçon		
			11	Parmenier	Lucie	1877	Lambach	"	époux	ans		
			12	Parmenier	Maximilien	1885	Paris	"	enfant	Confectionnier	Thulenau	
			13	Parmenier	Emma	1900	W. Denis	"	"	empl.	Minist. Paris	
			14	Parmenier	Ludovic	1902	"	"	"	fil.	Calix	
			15	Caliez	Jorge	1878	Lill.	"	chef	Confectionn.	Mag.	
			16	Caliez	Celine	1884	Stade	"	époux	ans		
17	Renault	Jean	1869	Langreny	"	chef	chef équipe	Minist.				

1. Rue
2. Maison
3. Ménage
4. Individu



# De multiples objectifs et enjeux



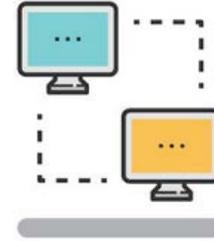
## Science

**Etudes en économie, en histoire et en démographie :** évolution du marché du travail, des inégalités, de la structure sociale, des mobilités.



## Technologie

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



## Valorisation

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

# De multiples objectifs et enjeux



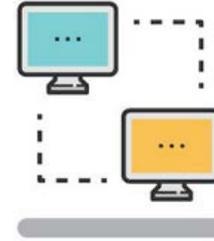
## Science

**Etudes en économie, en histoire et en démographie :** évolution du marché du travail, des inégalités, de la structure sociale, des mobilités.



## Technologie

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



## Valorisation

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

# Des données individuelles à l'échelle nationale

- Une histoire économique et sociale à la fois individuelle et nationale
  - ❖ Combler un vide entre enquêtes nationales à partir de statistiques agrégées et monographie plus locales.
  - ❖ Une période décisive dans l'histoire de la France: industrialisation, urbanisation, transformations sociales profondes.
- Différentes pistes suivies: sur une année, plusieurs années combinées, avec appariement des individus ou non.
- Et après?
  - ❖ Base d'étude ou d'analyse pour de futures enquêtes historiques: matrice de la recherche en histoire quantitative contemporaine.
  - ❖ Lien avec la période contemporaine.

# Différents angles d'analyse

## ➤ Changement de structure dans la durée

- ❖ Evolution du marché du travail: situation selon l'état marital, le genre, etc.
- ❖ Hétérogénéité spatiale des phénomènes.

## ➤ Chocs et modifications ponctuelles

- ❖ Conséquences à court, moyen et long terme.
- ❖ Par exemple: crise du phylloxera; Première guerre mondiale.

## ➤ Structuration de l'espace et des activités économiques

- ❖ Croisement des informations individuelles avec des données spatiales: structuration de l'espace.
- ❖ Effet de la transition entre agriculture et industrie à l'échelle locale.

# L'apport des appariements: étudier la mobilité

- Trajectoire des individus au cours de leur vie
  - ❖ Dans l'enfance et dans l'âge adulte.
  - ❖ Transition entre les deux.
- Etudier la mobilité géographique et sociale.
- Et son évolution au cours du temps.

# L'apport des appariements: étudier la mobilité

1886

La Place.

31	36	92	Baillandier	Edouard	44	Notaire	chef de ménage
		93	Houssard	Octavie	30	propriétaire	sa femme
		94	Baillandier	Edouard	5	"	leur fils
		95	Baillandier	Marie	3	"	leur fille
		96	Besnard	Welfphine	72	propriétaire	meie du chef
		97	Bourcelot	Emili	20	cuisinière	domestique
32	36	98	Nelaunay	Aimable	40	ordonnier	chef de ménage
		99	Beaudet	Eugénie	32	M <sup>rs</sup> a nouveauté	sa femme
35	37	100	Chapin	Julien	73	rentier	chef de ménage
		101	Féat	Marie Anne	60	rentière	sa femme
		102	Guille	Joseph	42	curé	chef de ménage
34	38	103	Grandin	Michel	27	Maçon	Maçon
35	39	104	Buchard	Jeann			
		105	Briguel	Louise			
35	40	106	Bourcelot	Rein			

1896

Le Bourg - La Place

1	1	1	Guille	Joseph	52	id	desservant	chef
		2	Renard	Henri	43	id	receveur	"
		3	Buchard	Jeann	47	id	domestique	E
		4	Stebatun	Sœur Charles	40	id	percepteur	chef
3	5	1	Nelaunay	Aimable	50	id	propriétaire	époux
		2	Beaudet	Eugène	43	id	id	épouse
		3	Baillandier	Edouard	53	id	notaire	époux
		4	Houssard	Octavie	40	id	propriétaire	épouse
		5	Robert	Reine	19	id	domestique	E
		6	Seroy	Victor	30	id	maître hotel	époux
		7	Néle	Rosette	36	id	id	épouse
		8	Seroy	Victor	11	id	"	fils
		9	Seroy	Robert	7	id	"	id
		10	Néle	Eugène	27	id	domestique	beau-père

1906

Le Bourg - La Place

3	3	1	Baillandier	Edouard	1882	Notaire	id	chef	maître hotel	patron
		2	Houssard	Octavie	1886	Propriétaire	id	épouse	"	"
		3	Baillandier	Edouard	1880	Notaire	id	fils	"	"
		4	Baillandier	Marie	1883	Notaire	id	filie	"	"
		5	Seroy	Victor	1881	Notaire	id	domestique	domestique	Baillandier
		6	Beaudet	Eugène	1880	Propriétaire	id	id	id	id
4	4	1	Seroy	Julie	1870	Propriétaire	id	chef	maître hotel	patron
		2	Beaudet	Justine	1871	Propriétaire	id	épouse	maître hotel	"
		3	Beaudet	Henri	1830	Propriétaire	id	chef	"	patron
5	5	1	Stellen	Esther	1881	Propriétaire	id	patron	maître hotel	maître hotel
		2	Seroy	Victor	1882	Propriétaire	id	chef	maître hotel	patron
6	6	1	Charraud	Victe	1861	Propriétaire	id	épouse	maître hotel	"
		2	Seroy	Victor	1887	Propriétaire	id	id	"	"
		3	Coussard	Thyasth	1851	Propriétaire	id	chef	ordonnaire	patron
		4	Lebrun	Thérèse	1851	Propriétaire	id	épouse	"	"
7	7	1	Coussard	Edouard	1873	Propriétaire	id	fils	ordonnaire	Coussard
		2	Stebatun	Thomas	1882	Propriétaire	id	accordeur	id	id

# De multiples objectifs et enjeux



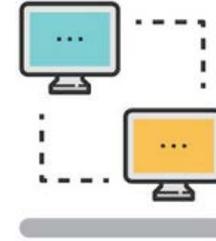
## Science

**Etudes en économie, en histoire et en démographie :** évolution du marché du travail, des inégalités, de la structure sociale, des mobilités.



## Technologie

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



## Valorisation

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

# De multiples objectifs et enjeux



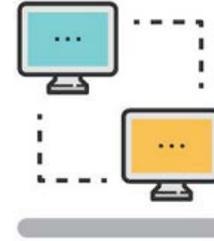
## Science

**Etudes en économie, en histoire et en démographie :** évolution du marché du travail, des inégalités, de la structure sociale, des mobilités.



## Technologie

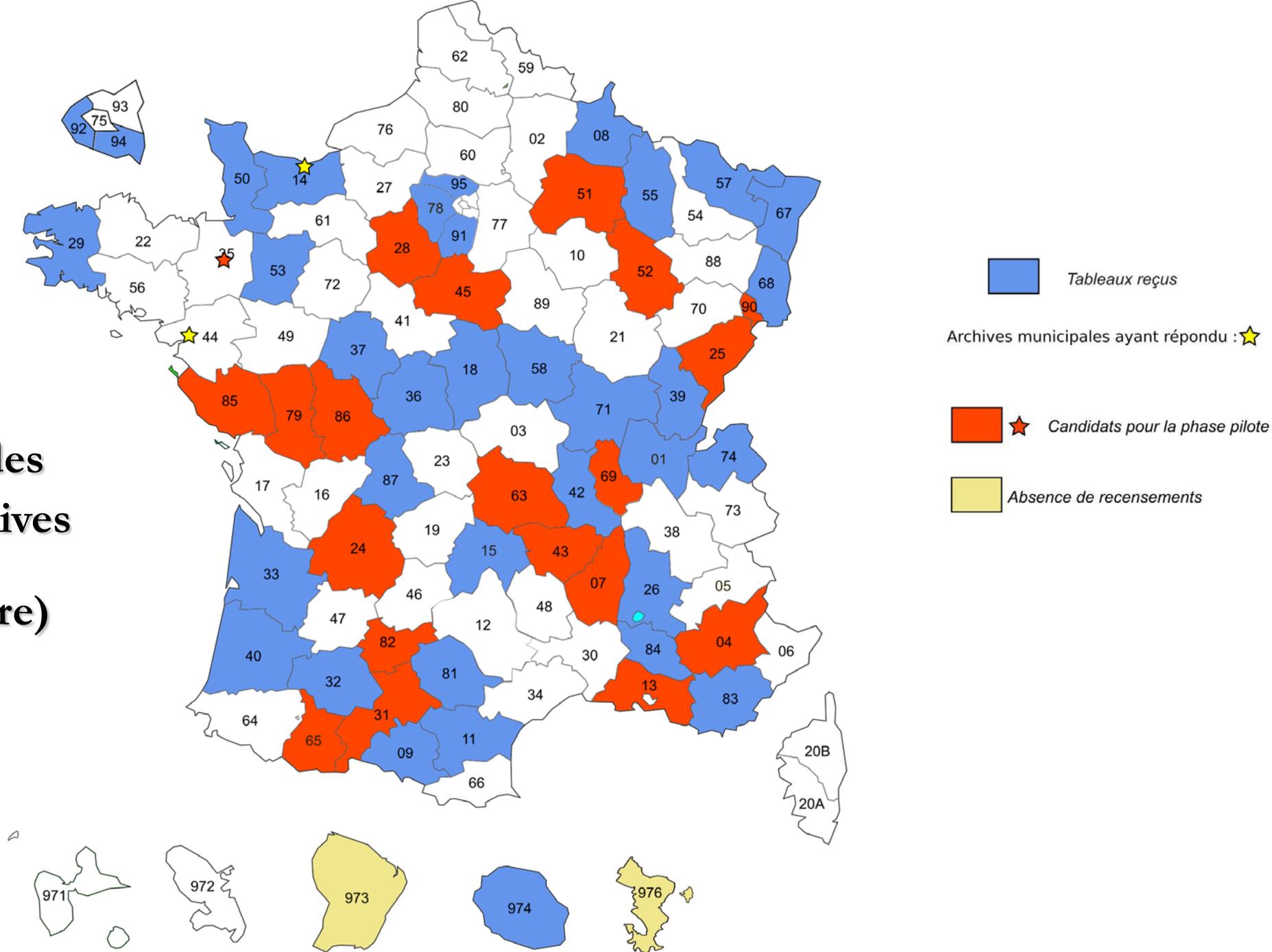
**Reconnaissance automatique d'écriture** manuscrite, analyse de tableau, traitement de plusieurs million d'images, accès à de multiples sources.



## Valorisation

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

# L'implication des services d'archives (au 20 septembre)



# Valorisation: retour aux archives (et au-delà?)

## ➤ Mise à disposition des données brutes aux archives

- ❖ Sur une base nationale (sur FranceArchives).
- ❖ Sur les bases des Archives Départementales.
- ❖ Idéalement avec le lien vers les photos.

## ➤ Mise à disposition des données de la recherche

- ❖ Base avec les codages des informations brutes.
- ❖ Base des appariements.

## ➤ Ouverture sur d'autres sources: un modèle pour la mise à disposition des archives nationales?

Nom

Prénom

Rechercher

Recherche avancée

Orthographe exacte



Orthographe exacte



Un  
modèle:  
le grand  
mémorial

à découvrir



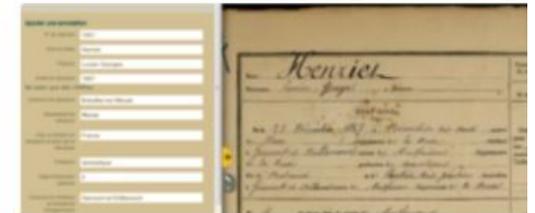
### Bases présentes dans le Grand Mémorial

Inauguré le 11 novembre 2014, le Grand Mémorial permet d'interroger les données d'indexation de...



### Qu'est-ce qu'un registre matricule militaire ?

Chaque conscrit se voyait attribuer un numéro de matricule, correspondant au numéro de la page d...



### Indexation collaborative et commémoration

Préalable indispensable à leur intégration dans le Grand Mémorial, les registres matricules doiv...



# Travail avec les services de l'Ined

Un projet avec une entreprise privée (PRCE): spécificité des accords à mettre en place, organisation des échanges, etc.

Un projet avec les archives nationales (SIAF): prise en charge des questions juridiques et de l'interaction avec les archives.

Un projet avec une masse considérable de données: besoins spécifiques?

# Travail avec les services de l'Ined (1)

## ➤ DRIP

- ❖ Soutien dans le dépôt du projet.
- ❖ Suivi de la convention ANR.
- ❖ Préparation de l'accord de consortium.

## ➤ Service des ressources humaines

- ❖ Recrutement d'un.e post-doc courant 2022 (pour l'appariement).

## ➤ SES

- ❖ Soutien à la diffusion des données à terme.
- ❖ Par exemple, Datalab?

## ➤ SMS

- ❖ Conseil et aide statistiques.

# Travail avec les services de l'Ined (2)

## ➤ Service informatique

- ❖ Equipements éventuels (post-doc/contractuels et chercheurs).
- ❖ Conseils sur le stockage des données.

## ➤ Communication

- ❖ Le logo et les visuels.
- ❖ Le site du projet: <https://socface.site.ined.fr/>.
- ❖ Diffusion des résultats.

## ➤ DPD et service juridique

- ❖ Aide à la préparation du PGD/DMP.
- ❖ Conseils sur la question de la déclaration des bases de données.

**Merci beaucoup!!!**

**<https://socface.site.ined.fr/>**

**[lionel.kesztenbaum@ined.fr](mailto:lionel.kesztenbaum@ined.fr)**