Un projet de recherche collaboratif entre

Historiens, économistes, démographes

Archivistes

Experts en intelligence artificielle

ined
INSTITUT NATIONAL D'ÉTUDES DÉMOGRAPHIQUES

PARIS SCHOOL OF ECONOMICS
ECOLE D'ECONOMIE DE PARIS

FranceArchives
PORTAIL NATIONAL DES ARCHIVES

TEKLIA

Et les services d'archives départementales et municipales

Lionel Kesztenbaum

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

SOCFACE

Un projet de recherche collaboratif entre

Historiens, économistes, démographes

Archivistes

Experts en intelligence artificielle

ined
INSTITUT NATIONAL D'ÉTUDES DÉMOGRAPHIQUES

PARIS SCHOOL OF ECONOMICS
ECOLE D'ÉCONOMIE DE PARIS

FranceArchives
PORTAIL NATIONAL DES ARCHIVES

TEKLIA

Et les services d'archives départementales et municipales

Financement sur **3 ans ½**

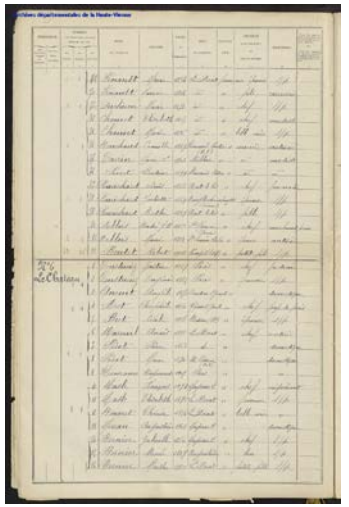
ANR AGENCE NATIONALE DE LA RECHERCHE

SocFace:

The local face of social change: one century of French social structure seen from the ground, 1836–1936

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

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.



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.

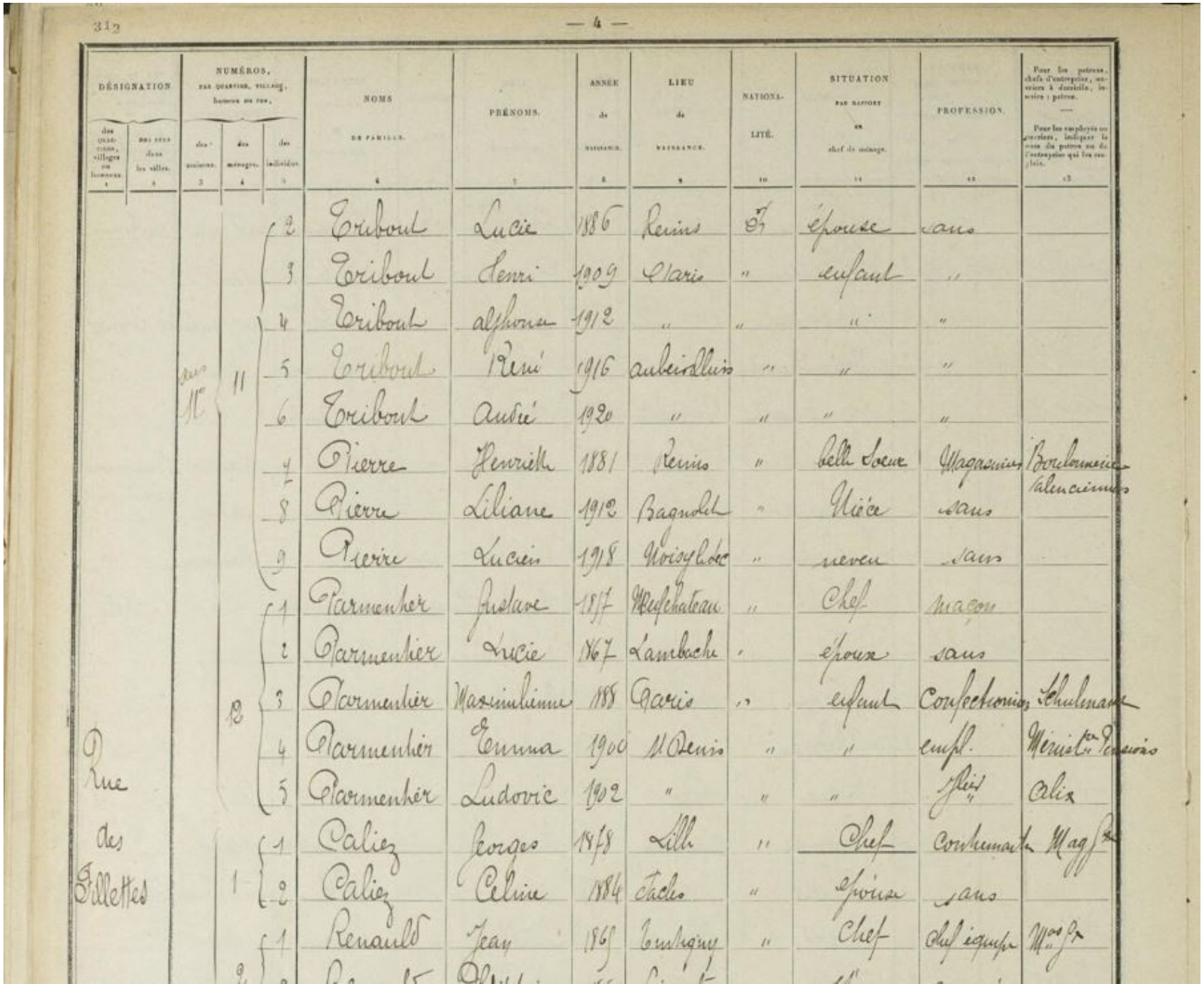
Produire des données individuelles en masse à l'échelle nationale

Développer les technologies de traitement de vaste ensemble de sources historiques

Diffuser les informations obtenues au grand public comme aux chercheurs

Une liste nominative:

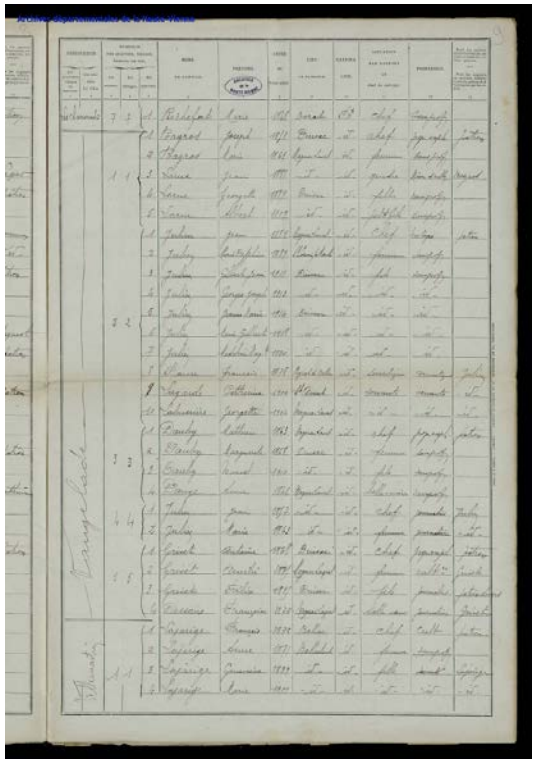
Aubervilliers, 1921



312 — 4 —

DÉSIGNATION des quartiers, villages ou hameaux.	DÉSIGNATION des rues dans les villes.	NUMÉROS par quartier, village, hameau ou rue, des maisons.	des ménages.	des individus.	NOMS DE FAMILLE.	PRÉNOMS.	ANNÉE de naissance.	LIEU de naissance.	NATIONALITÉ.	SITUATION par rapport au chef de ménage.	PROFESSION.	Pour les patrons, chefs d'entreprise, ouvriers à domicile, ouvriers : patrons. — Pour les employés ou ouvriers, indiquer le nom du patron ou de l'entreprise qui les emploie.
				2	Tribout	Lucie	1886	Reims	Fr	épouse	sans	
				3	Tribout	Henri	1909	Paris	"	enfant	"	
				4	Tribout	alphonse	1912	"	"	"	"	
		sans n°	11	5	Tribout	Rémi	1916	aubervilliers	"	"	"	
				6	Tribout	André	1920	"	"	"	"	
				7	Pierre	Henriette	1881	Reims	"	belle Sœur	Magasinier	Boulonnerie valenciennes
				8	Pierre	Liliane	1912	Bagnolet	"	Nièce	sans	
				9	Pierre	Lucien	1918	Montgeron	"	neveu	sans	
				1	Parmentier	Gustave	1877	Meusnes	"	Chef	maçon	
				2	Parmentier	Lucie	1867	Lambach	"	épouse	sans	
			12	3	Parmentier	Maximilienne	1888	Paris	"	enfant	Confectionneuse	Schulmann
				4	Parmentier	Emma	1900	St Denis	"	"	empl.	Ministère Travaux
Rue				5	Parmentier	Ludovic	1902	"	"	"	Elève	Calix
des				1	Caliez	Georges	1878	Lille	"	Chef	Contremaître	Magasin
Fillettes			1	2	Caliez	Céline	1884	Tacles	"	épouse	sans	
				1	Renauld	Jean	1865	Courbigny	"	Chef	Chef équipe	Mines
			2									

Pourquoi les listes nominatives?

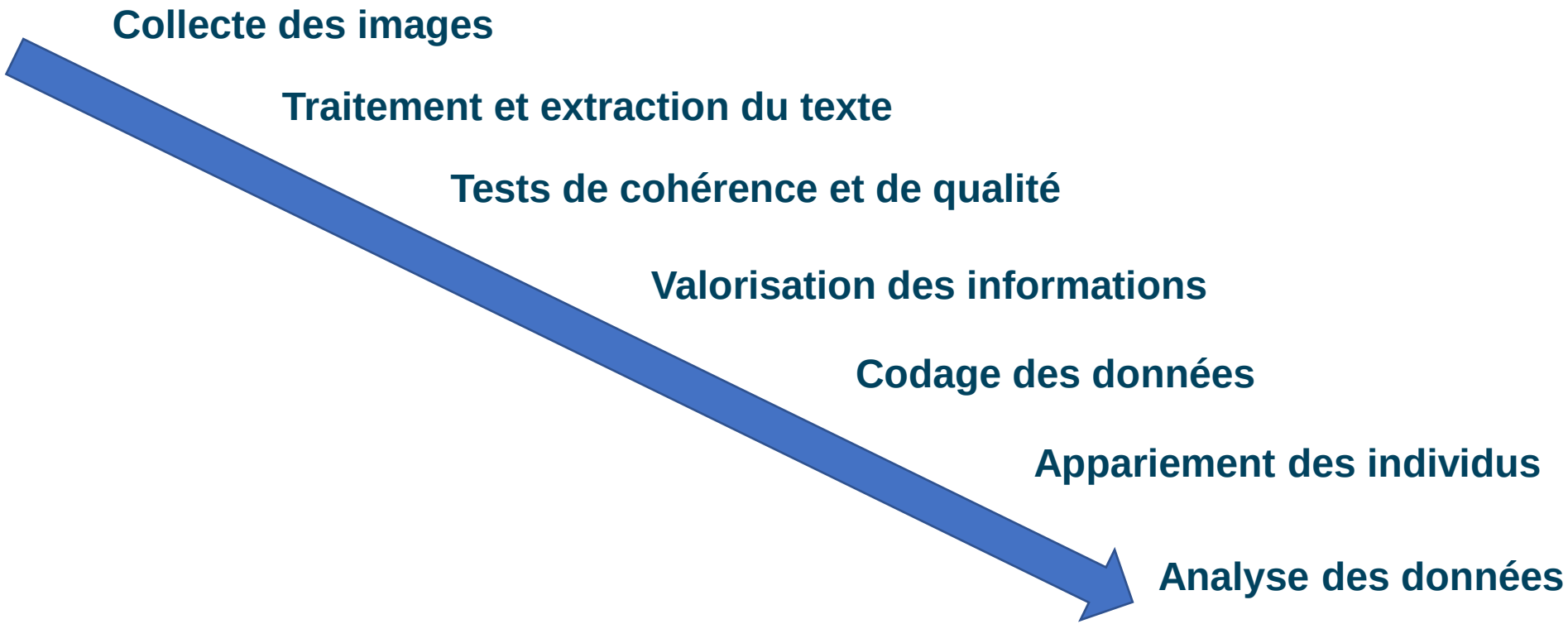


- Une source abondante, simple, stable et standardisée.
- Une source régulière dans le temps.
- Déjà photographiée par beaucoup de services d'archives.
- Source nationale, à l'identique.
- Permet de construire un échantillon de toute la France (ou presque…).

→ 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.

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 perspective 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 perspective considérables pour la recherche en sciences sociales.

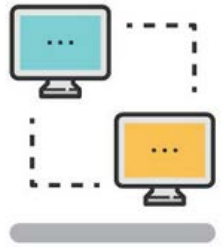
De multiples objectifs et enjeux

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.

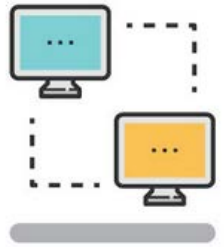
De multiples objectifs et enjeux

Science

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

Technologie

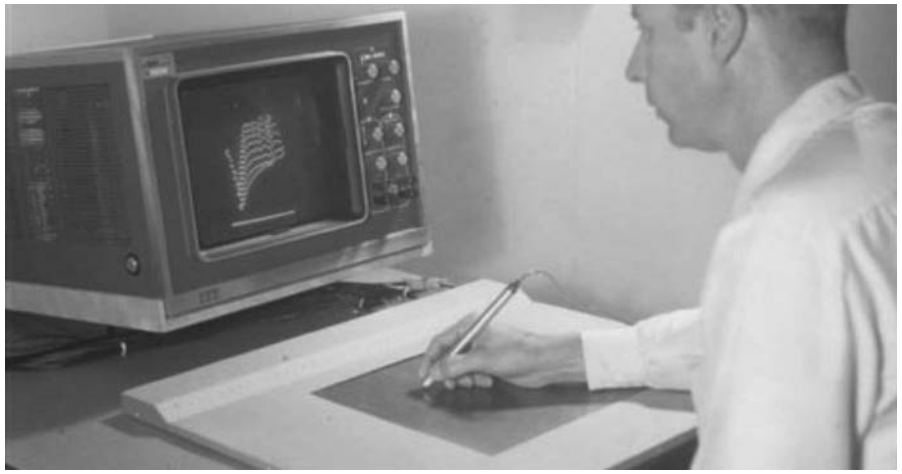
Reconnaissance automatique d'écriture manuscrite, analyse de tableau, traitement de plusieurs million d'images, accès à de multiples sources.



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



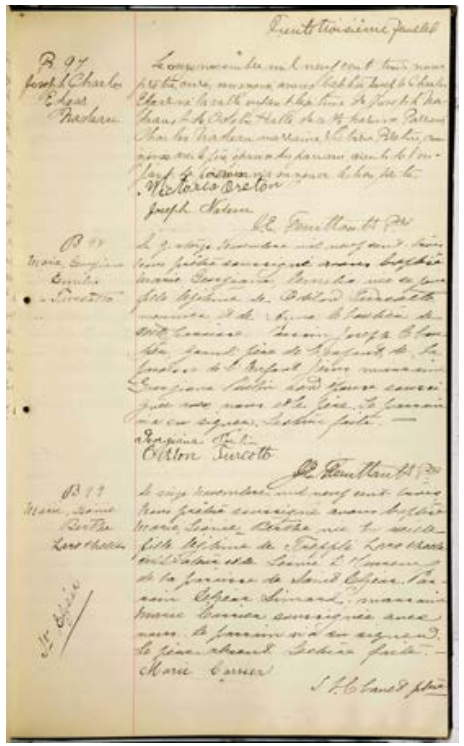
RAND corporation, 1960

The MNIST database
"The drosophila of machine learning"
Geoffrey Hinton

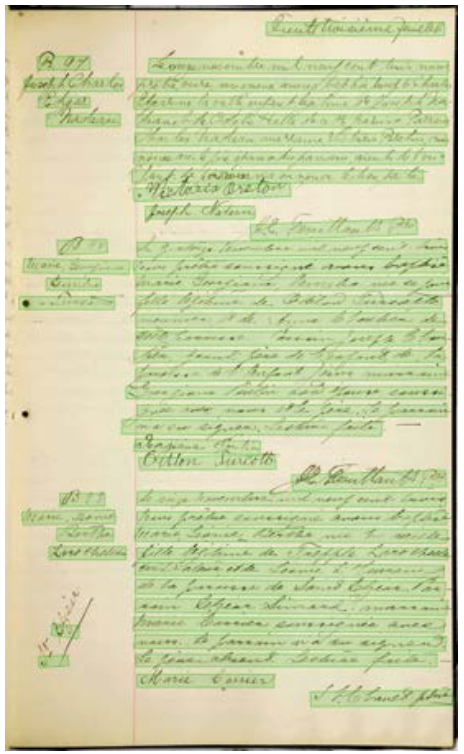
Mais les machines ne rivalisent toujours pas avec l'humain pour la lecture de documents manuscrits

Chaîne de traitement automatique

Image



Détection des lignes



Reconnaisssance

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 , Etaients 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

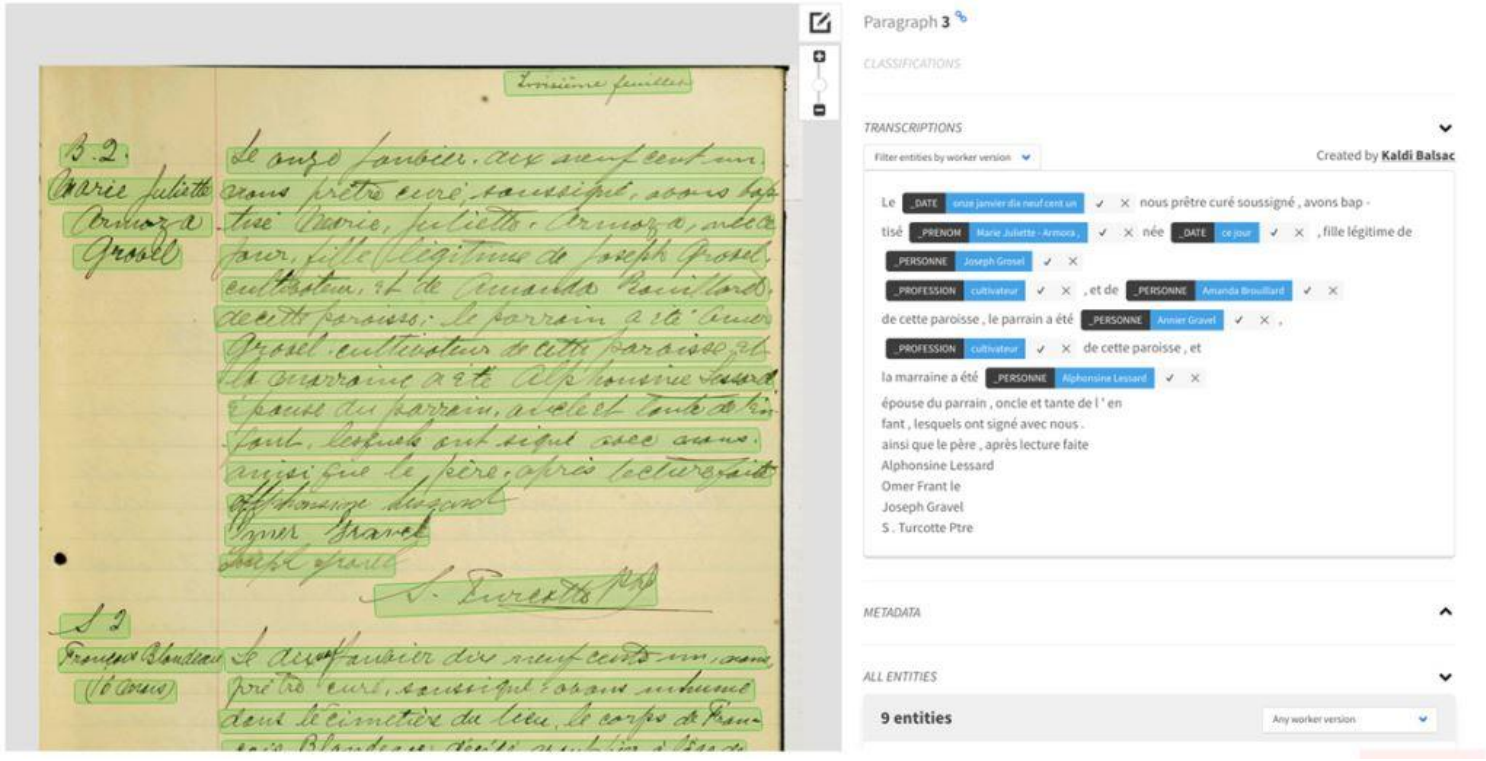
Extraction d'entités

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 , Etaients 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 entrainés à partir d'exemples

Est-ce que ca fonctionne ?

UQAC
Université du Québec
à Chicoutimi



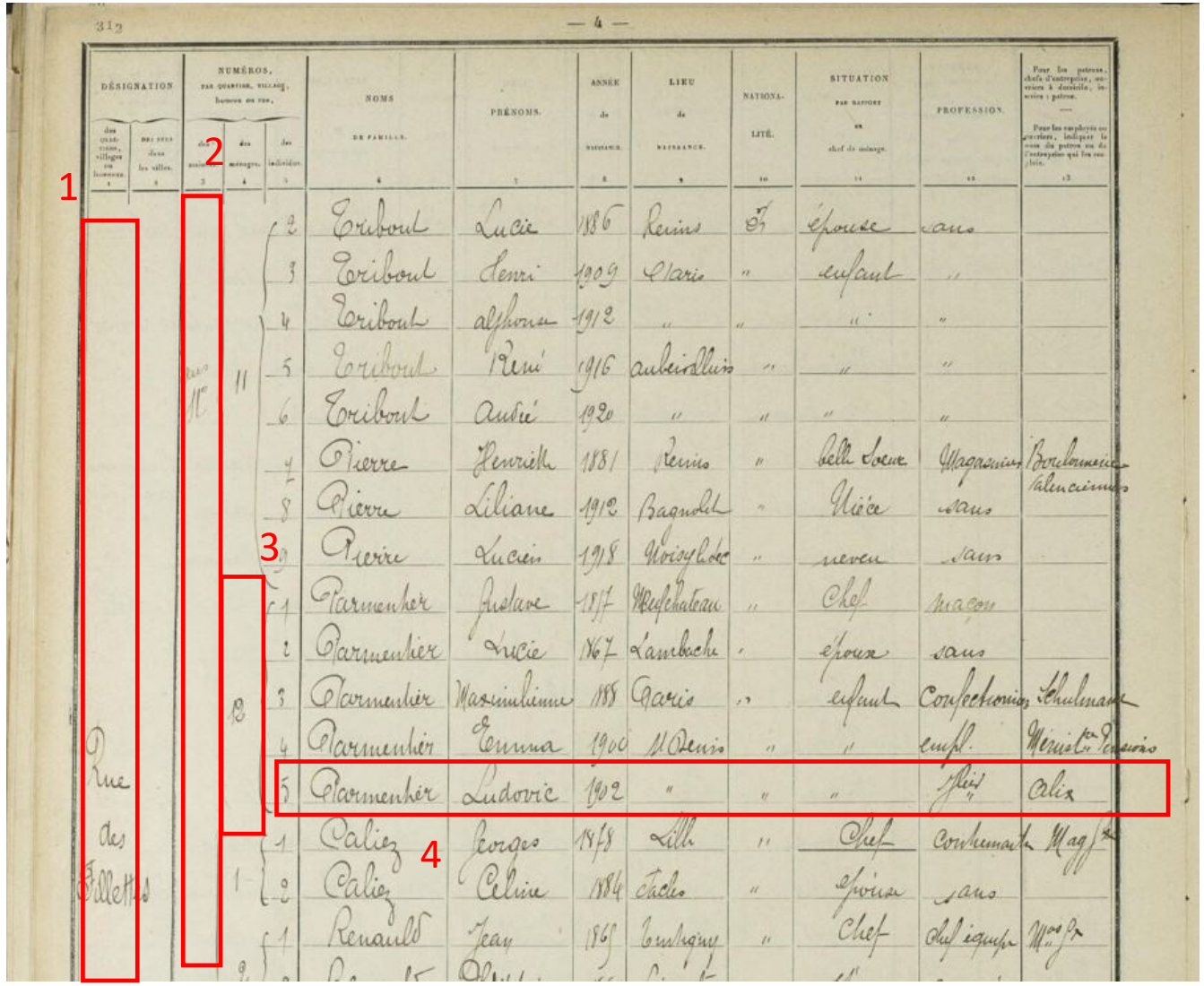
- 2.7M d'images de registres paroissiaux de 1850 à 1920
- HTR, segmentation en acte, detection et typage des entités
- Taux d'erreur caractère moyen : 6.4%

Listes nominatives de recensement

Défis du projet:

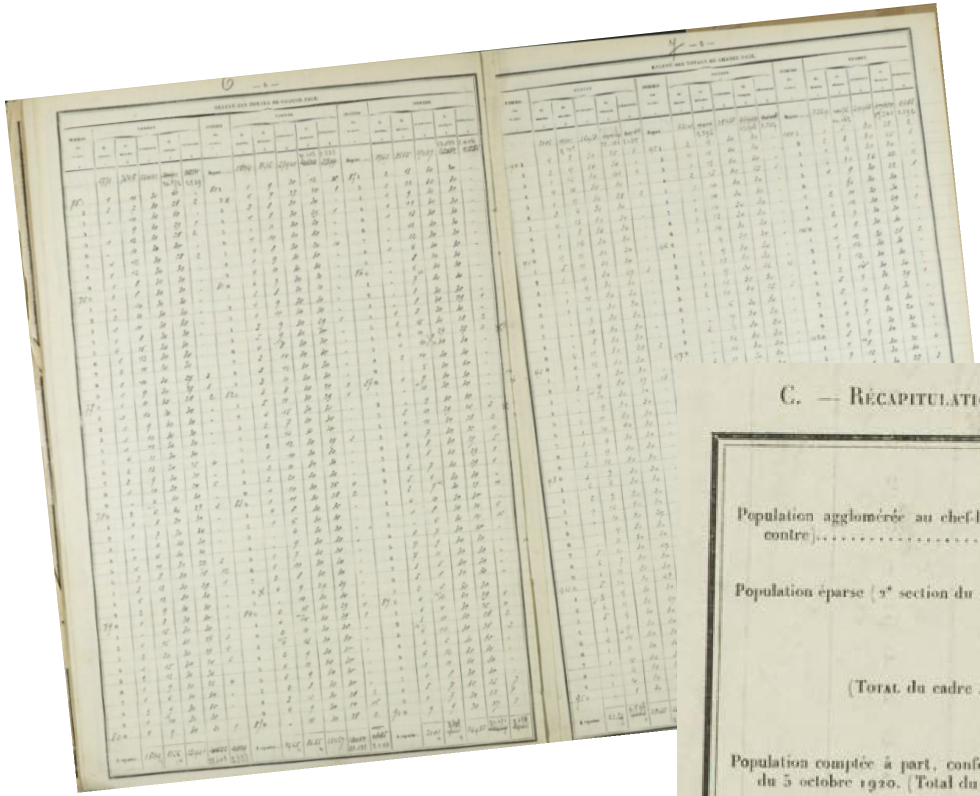
- Documents décentralisés
- Volume inconnu mais important (entre 5 et 15 M d’images ?)
- Conditions de conservation/numérisation variables
- Lecture en 2 dimensions
- Evaluer l’erreur
- Structurer, consolider, croiser l’information

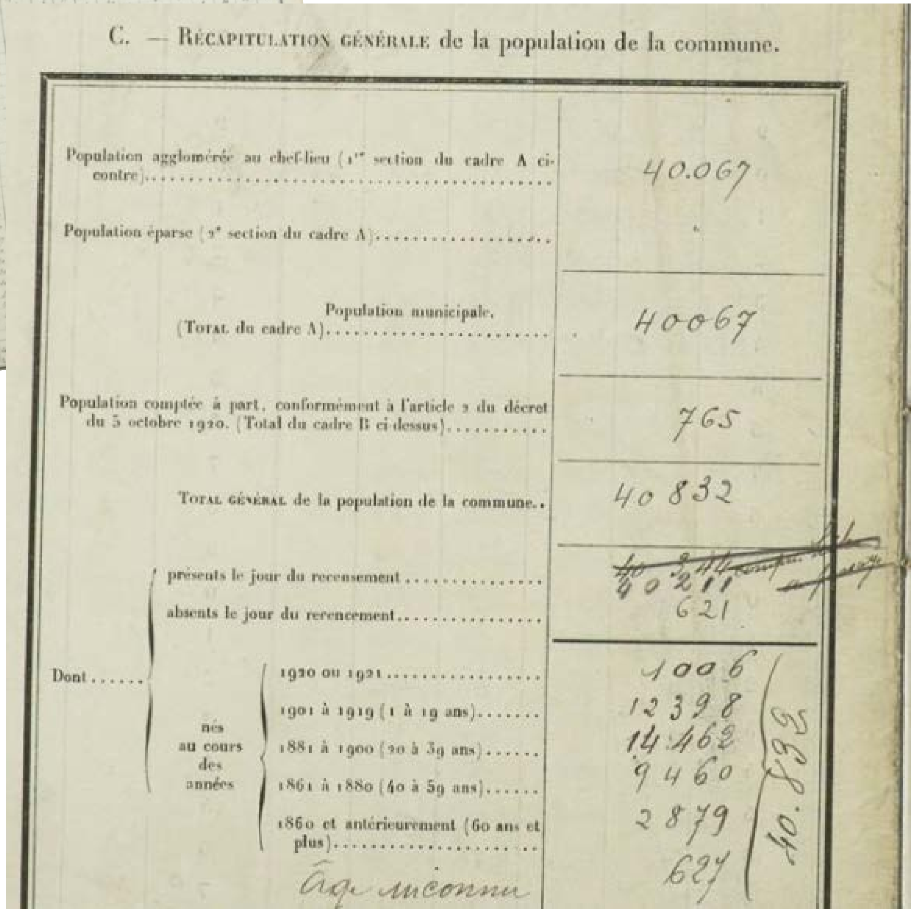
Lecture en 2 dimensions



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

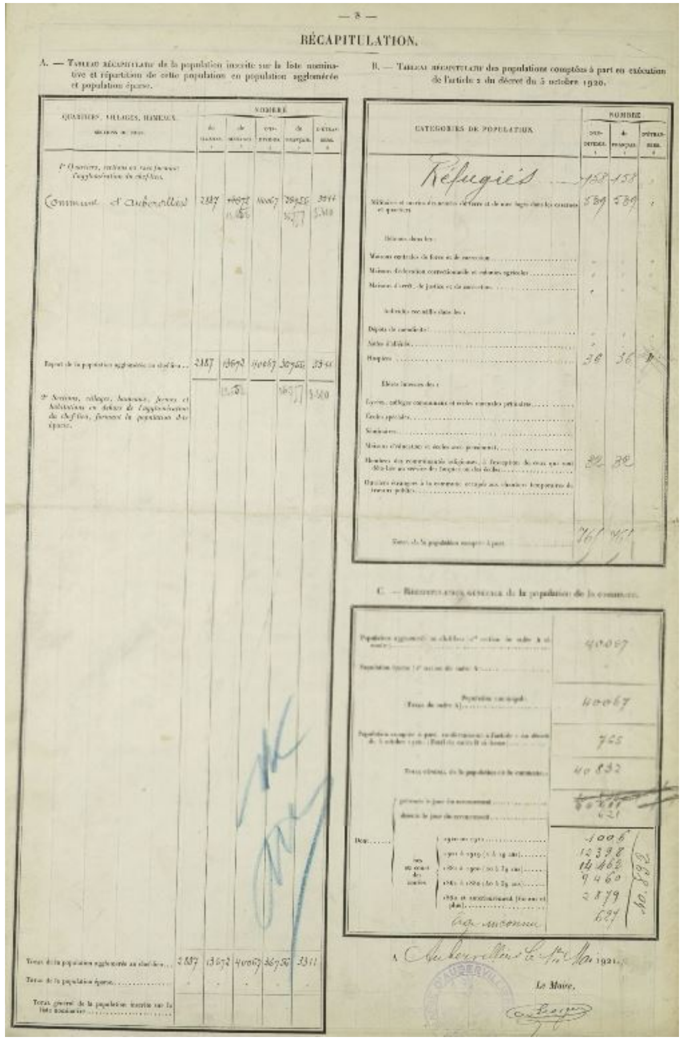
Tests de cohérence et de cohésion





C. — RÉCAPITULATION GÉNÉRALE de la population de la commune.

Population agglomérée au chef-lieu (1re section du cadre A ci-contre)	40.067
Population éparse (2e section du cadre A)	»
Population municipale. (TOTAL du cadre A)	40067
Population comptée à part, conformément à l'article 2 du décret du 5 octobre 1920. (Total du cadre B ci-dessus)	765
TOTAL GÉNÉRAL de la population de la commune.	40 832
Dont présents le jour du recensement	~~40 344~~ 40 211
Dont absents le jour du recensement	621
Dont nés au cours des années 1920 ou 1921	1006
Dont nés au cours des années 1901 à 1919 (1 à 19 ans)	12398
Dont nés au cours des années 1881 à 1900 (20 à 39 ans)	14462
Dont nés au cours des années 1861 à 1880 (40 à 59 ans)	9460
Dont nés au cours des années 1860 et antérieurement (60 ans et plus)	2879
âge inconnu	627
Total	40.832



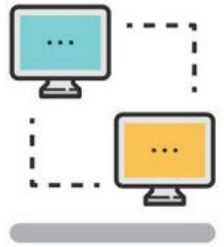
De multiples objectifs et enjeux

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.

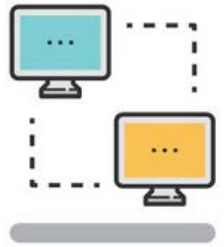
De multiples objectifs et enjeux

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

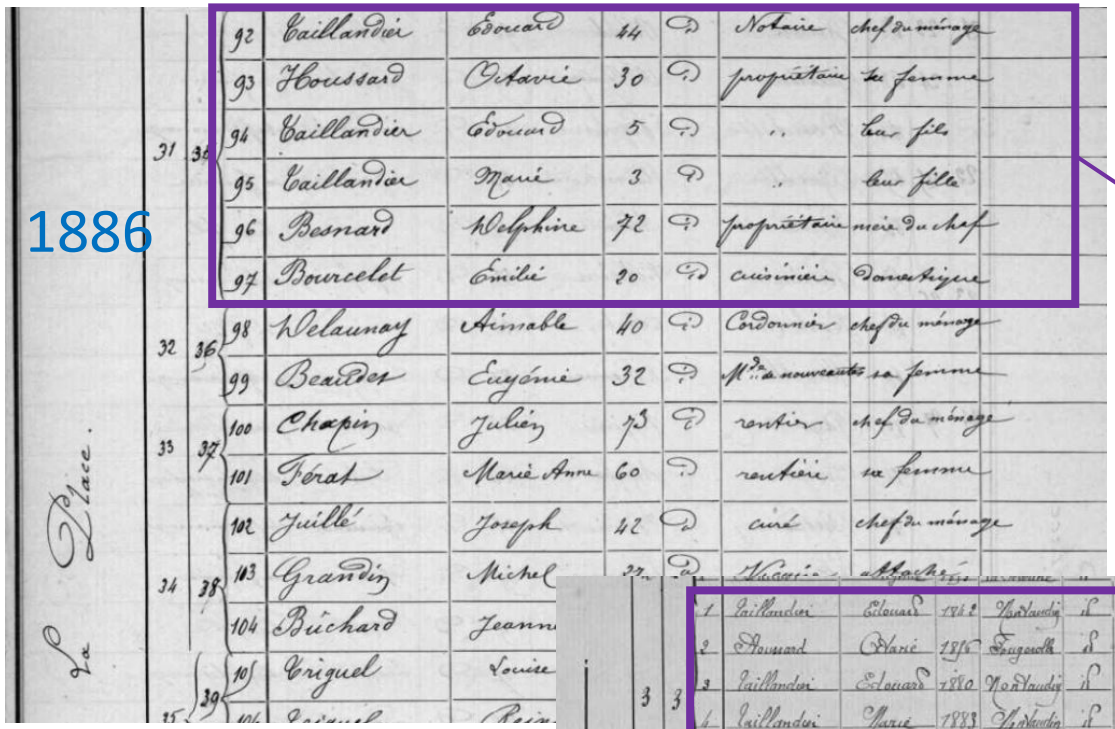
- Changement de structure dans la durée
  - Evolution du marché du travail: situation selon l'état marital, le genre, etc.
  - Hétérogénéité spatiale des phénomènes.
- Chocs et modifications ponctuelles
  - Conséquences à court, moyen et long terme.
  - Par exemple: crise du phylloxera; Première guerre 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.

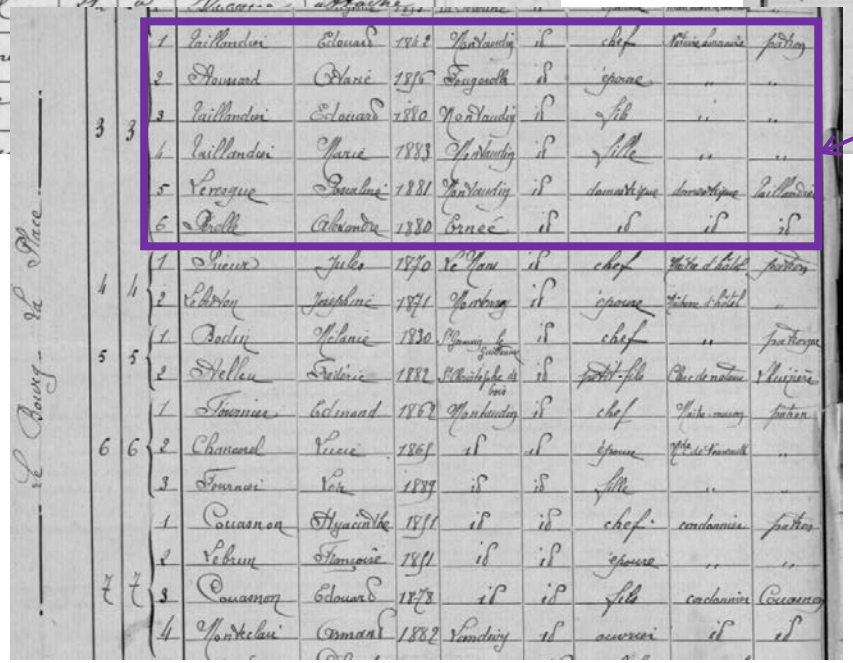
L'apport des appariements: étudier la mobilité

1886



1896

1906



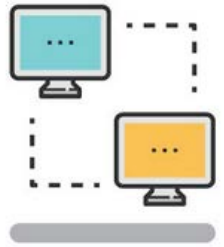
De multiples objectifs et enjeux

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.

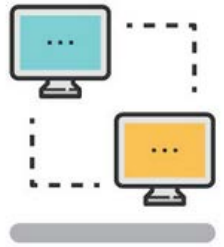
De multiples objectifs et enjeux

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.

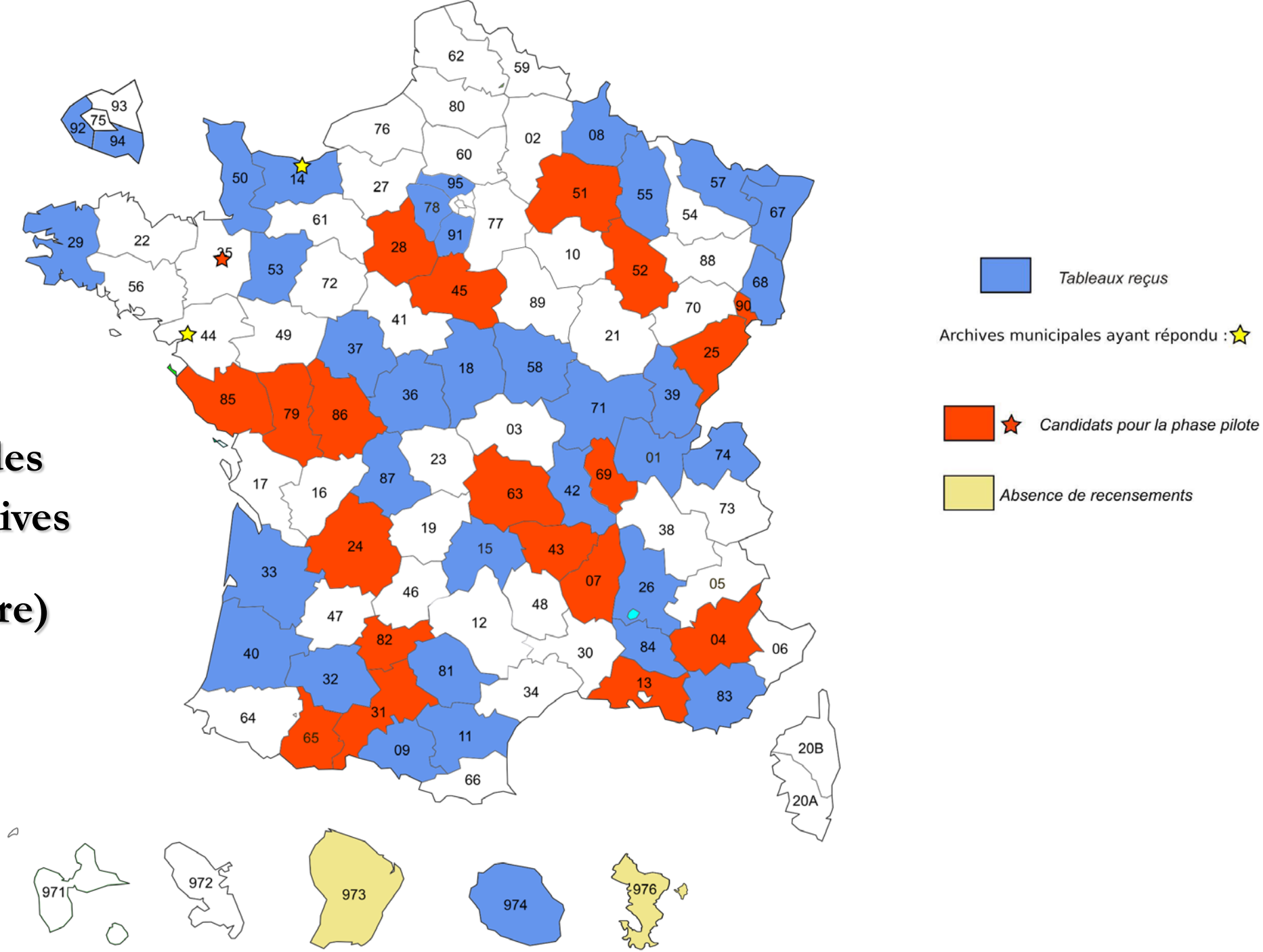


Valorisation

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)



Tableaux reçus

Archives municipales ayant répondu :

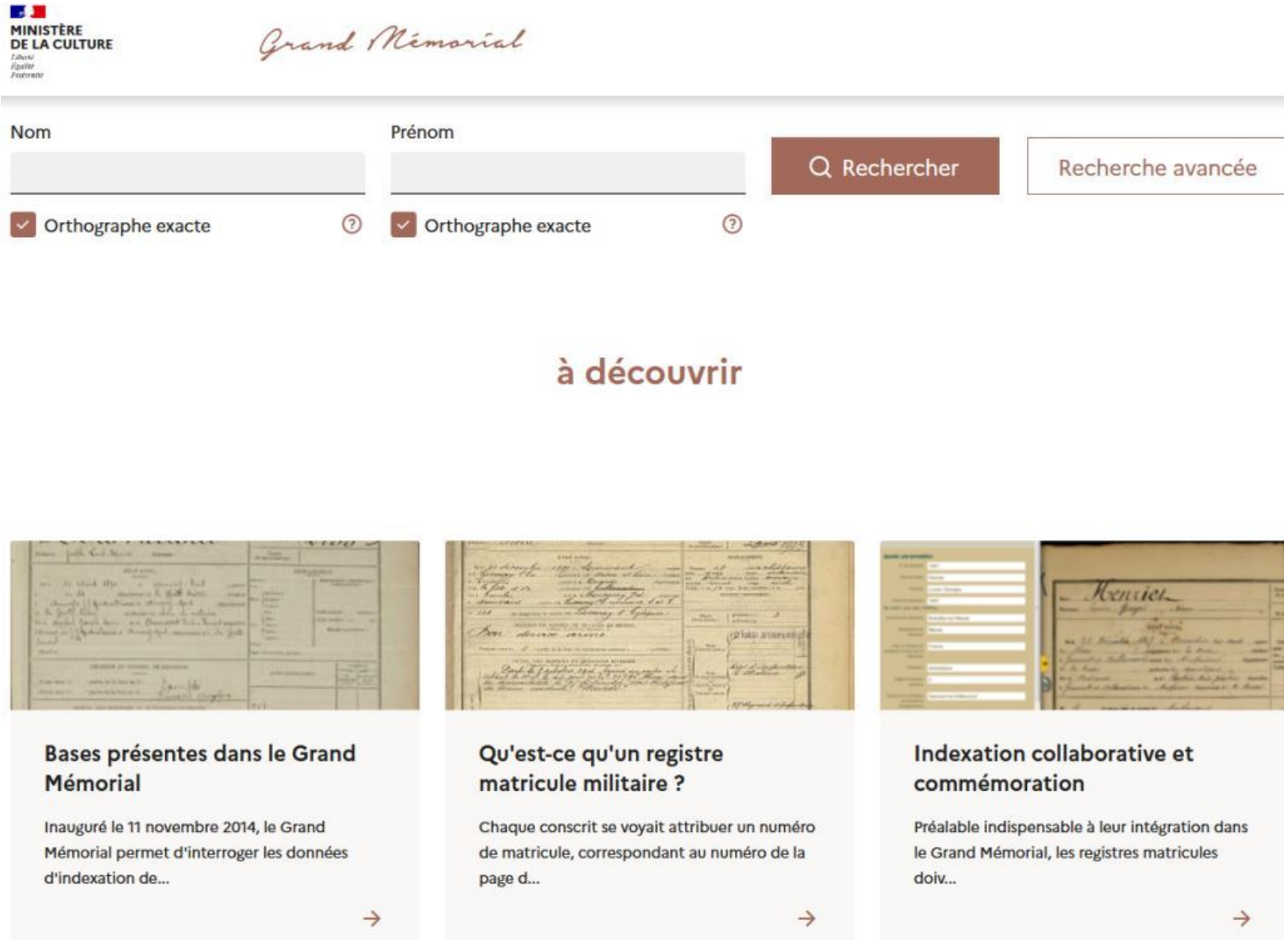
Candidats pour la phase pilote

Absence de recensements

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?

Un modèle: le grand mémorial



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