

The logo for SBC FACE, with 'SBC' in a black cursive font and 'FACE' in a bold, orange, sans-serif font.

SBC FACE



Un projet de recherche collaboratif entre

Historiens, économistes, démographes

Archivistes

Experts en intelligence artificielle



FranceArchives
PORTAIL NATIONAL DES ARCHIVES

*Archives départementales
et municipales*

TEKLIA

Christopher Kermorvant
Lionel Kesztenbaum

anr[©]

6^e session du CHEC
Musée de l'histoire de l'immigration
10/10/2024



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**

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 monographies plus locales.
 - ❖ Une période décisive dans l'histoire de la France: industrialisation, urbanisation, transformations sociales profondes.
- Différentes pistes poursuivies
 - ❖ 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.
 - ❖ Liens avec la période contemporaine.

LES LISTES NOMINATIVES

Défis et objectifs



A handwritten nominative list on a grid. The text is written in a cursive script. The list is organized into columns, with names and other identifying information. The handwriting is dense and fills most of the grid cells.



A handwritten nominative list on a grid. The text is written in a cursive script. The list is organized into columns, with names and other identifying information. The handwriting is dense and fills most of the grid cells.



A handwritten nominative list on a grid. The text is written in a cursive script. The list is organized into columns, with names and other identifying information. The handwriting is dense and fills most of the grid cells.

DESIGNATION		NUMÉROS, PAR QUARTIER, VILLAGES, HAMEAUX ou VILLES			NOMS	PRÉNOMS	ANNÉE	LIEU	NATIONALITÉ	SITUATION PAR RAPPORT	PROFESSION	Notes
des quartiers, villages ou hameaux	des rues ou des villes	des numéros	des ménages	des habités	DE FAMILLE		de naissance	de	LITTÉ.	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	Alphons	1912	"	"	"	"	
				5	Eribout	Reni	1916	Aubervilliers	"	"	"	
				6	Eribout	Audé	1920	"	"	"	"	
				7	Pierre	Henriette	1881	Reims	"	belle sœur	Magasinier	Boulonnais Valenciennes
				8	Pierre	Liliane	1912	Bagnollet	"	filie	sans	
				9	Pierre	Lucien	1918	Woisylot	"	neveu	sans	
				10	Parmentier	Julave	1877	Mesfiteau	"	chef	maçon	
				11	Parmentier	Lucie	1867	Lambach	"	épouse	sans	
				12	Parmentier	Maximilien	1888	Paris	"	enfant	Confectionnier	Chulman
				13	Parmentier	Emma	1900	W. Denis	"	"	empl.	Minist. Pensions
				14	Parmentier	Ludovic	1902	"	"	"	époux	Calix
				15	Caliez	Georges	1878	Lille	"	chef	Commerçant	Magist
				16	Caliez	Céline	1884	Taches	"	épouse	sans	
				17	Renauld	Jean	1869	Combrigny	"	chef	chef équipe	Magist
				18	Renauld	Olivia	1875	Combrigny	"	"	"	

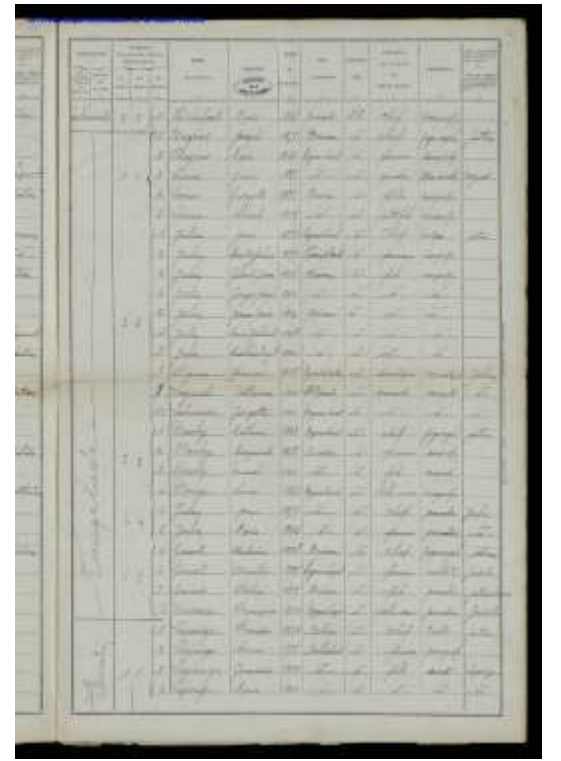
Rue
des
Ellelles

Une liste nominative:
Aubervilliers, 1921

REGISTRATION		NUMEROUS, the number, being shown in it.			NAME OR SURNAME	YEAR	SEX	NATIONALITY	SITUATION OR OCCUPATION	PROFESSION	Date of arrival in the country, or date of departure from the country, or date of return from the country (if applicable)
1	2	3	4	5							
		II	2	Eribout	Lucie	1886	Reine	F	épouse	ans	
			3	Eribout	Samy	1909	Clara	"	enfant	"	
			4	Eribout	Alphonse	1912		"	"	"	
			5	Eribout	Reni	1916	Antoinette	"	"	"	
			6	Eribout	Audré	1920	"	"	"	"	
			7	Pierre	Jennette	1881	Renée	"	bellevue	Magasin	Bellevue Mauricie
		8	Pierre	Liliane	1912	Barthé	"	Uice	ans		
		9	Pierre	Lucien	1918	Wojtyla	"	veuve	ans		

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.



Défis et obstacles

➤ Collecte des données

- ❖ Les sources et les images sont conservées dans les archives départementales et municipales (100 dépôts d'archives).
- ❖ Très forte hétérogénéité dans la conservation des sources.

➤ Transcription du texte

- ❖ Très grande quantité de scripteurs différents, avec des pratiques spécifiques.
- ❖ Très grande diversité des entrées, en particulier pour les abréviations ('idem').

➤ Appariement des individus à travers le temps et l'espace

- ❖ Manques d'information: il y a des archives détruites, les ménages collectifs manquent, etc.
- ❖ Faiblesse de l'information individuelle disponible (ainsi souvent il n'y a qu'un seul prénom).

➤ Et en sciences sociales?

- ❖ Information sur la catégorie sociale limitée: uniquement la profession.
- ❖ Importants vides: Paris est absente (jusqu'en 1926), de larges zones sont manquantes, etc.

→ Un défi commun: **la taille**, des millions d'images, des centaines de millions d'observations...

Les défis comme objectifs

➤ Collecte des données

- ❖ Un des objectifs du projet est de quantifier précisément l'état des collections en France.
- ❖ Le projet est incitatif pour que les Archives développent et améliorent leurs collections.

➤ Transcription du texte

- ❖ La diversité des cas n'est pas simplement une question d'écriture, mais aussi de pratiques, d'habitudes, etc.
- ❖ C'est un argument fort pour justifier la collaboration entre démographes, historiens et experts en informatique.

➤ Appariement des individus à travers le temps et l'espace

- ❖ Certaines particularités du recensement français peuvent aider à améliorer les appariements (nom de jeune fille).
- ❖ Permet d'évaluer comment la reconnaissance automatique influence l'appariement.

➤ Et en sciences sociales?

- ❖ Permet de déplacer la focale au-delà de zone très étudiées (hors de Paris, banlieue, etc.).
- ❖ Une base de données qui servira de base à d'autres études, autour d'autres sources.

Lecture en 2 dimensions

312 — 4 —

DÉSIGNATION		NUMÉROS, PAR QUARTIER, VILLAGES, HAMEAUX ou TERRES			NOMS		ANNEE	LIEU	NATIONAL.	SITUATION	PROFESSION	
des villages ou hameaux	des quartiers ou villages	des maisons	des maisons	des maisons	DE FAMILLE	PRÉNOMS	de NAISSANCE	de NAISSANCE	LIBRE	en chef de ménage		
1	2	3	4	5	6	7	8	9	10	11	12	13
Rue des Filleuls	115	11	2	2	Erribout	Lucie	1886	Reims	F	épouse	ans	
			3	3	Erribout	Léoni	1909	Paris	"	enfant	"	
			4	4	Erribout	Alphonse	1912	"	"	"	"	
			5	5	Erribout	Reni	1916	Aubervilliers	"	"	"	
			6	6	Erribout	Audré	1920	"	"	"	"	
		4	7	Pierre	Henriette	1881	Reims	"	beau frère	Magasin	Parabromure bromocinnol	
		8	8	Pierre	Liliane	1918	Bagnollet	"	filie	ans		
		3	9	Pierre	Lucien	1918	Wingholte	"	veuve	ans		
		1	10	Parmenier	Juliane	1877	Meschesnois	"	chef	maçon		
		2	11	Parmenier	Lucie	1867	Lambach	"	époux	ans		
		3	12	Parmenier	Maximilien	1885	Paris	"	enfant	Confectionner	Chulman	
		4	13	Parmenier	Emma	1900	St Denis	"	"	empl.	Mémoire Paris	
		5	14	Parmenier	Ludovic	1902	"	"	"	"	chef	Calix
		1	15	Caliez	Jorges	1878	Lille	"	chef	Confectionner	Mag	
		2	16	Caliez	Celine	1884	Stache	"	époux	ans		
1	17	Renault	Jean	1869	Combray	"	chef	chef équipe	Mag			

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



TRAITEMENT

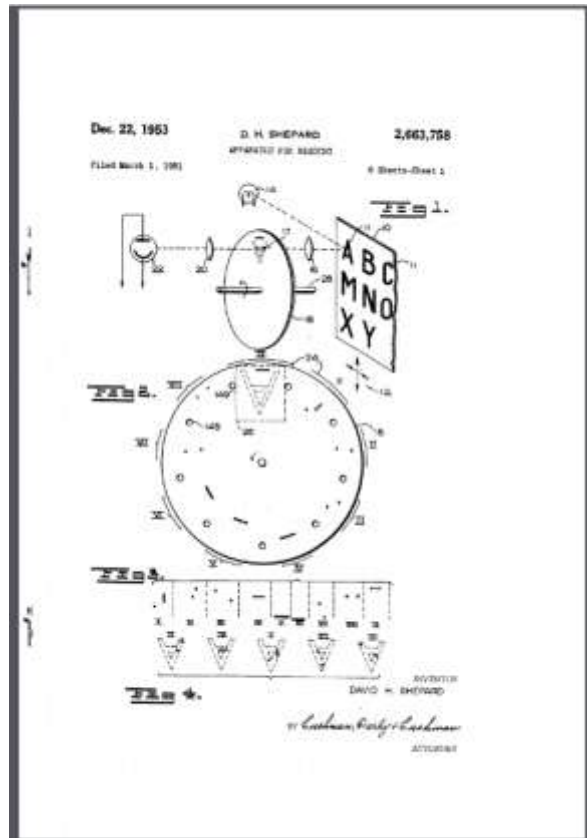
Extraire l'information des listes

A handwritten ledger page with multiple columns and rows. The text is written in cursive and appears to be a list of items or transactions. The columns are not clearly defined but seem to contain names, dates, and numerical values.

A handwritten ledger page with multiple columns and rows. The text is written in cursive and appears to be a list of items or transactions. The columns are not clearly defined but seem to contain names, dates, and numerical values.

A handwritten ledger page with multiple columns and rows. The text is written in cursive and appears to be a list of items or transactions. The columns are not clearly defined but seem to contain names, dates, and numerical values.

Reconnaissance d'écriture : un défi pour l'IA



Mais les performances restent loin derrière les performances humaines

<https://patents.google.com/patent/US2663758A/en>

Derrière les sigles :

OCR

Optical Character Recognition

Écriture imprimée
dactylographiée

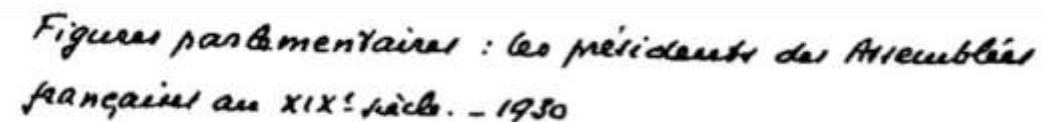


DÉBATS LÉGISLATIFS & POLITIQUES

HTR

Handwritten Text Recognition

Écriture manuscrite
cursive



*Figures parlementaires : les présidents des Assemblées
françaises au XIX^e siècle. - 1930*

ATR : Automatic Text Recognition

Dépasser la reconnaissance de caractères

OCR

ÉTAT NOMINATIF

Segmentation



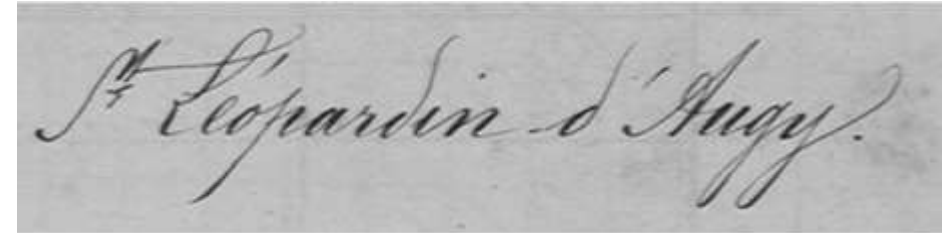
É T A N O M I N F
T A I
T

Reconnaissance



ETAT NOMINATIF

HTR

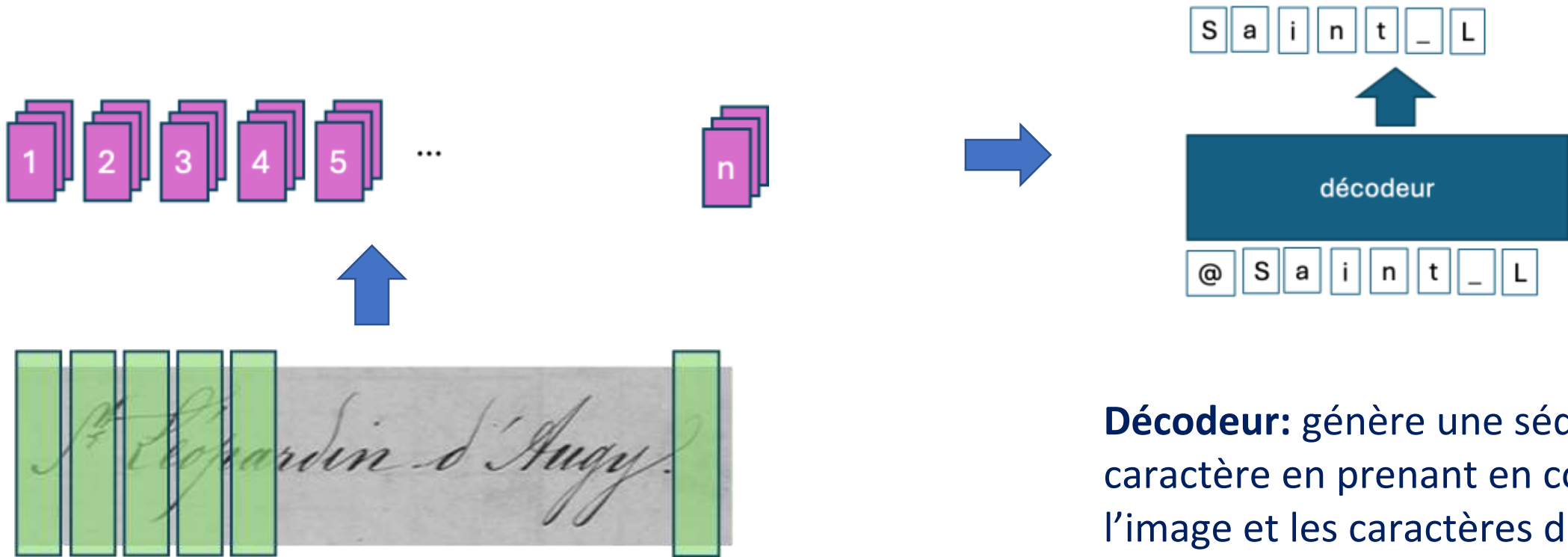


Pour reconnaître les lettres, il faut les segmenter mais pour les segmenter, il faut les reconnaître

Paradoxe de Sayre

Reconnaissance d'écriture manuscrite

IA, Deep Learning, Réseaux de neurones, Transformers

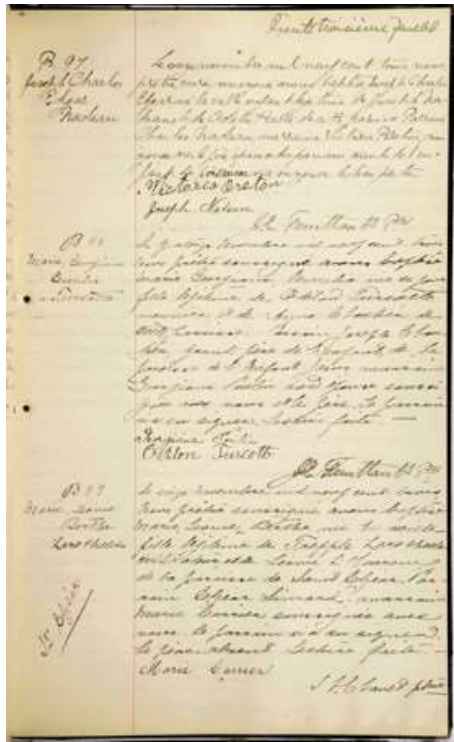


Encodeur: crée une représentation en haute dimension de l'image

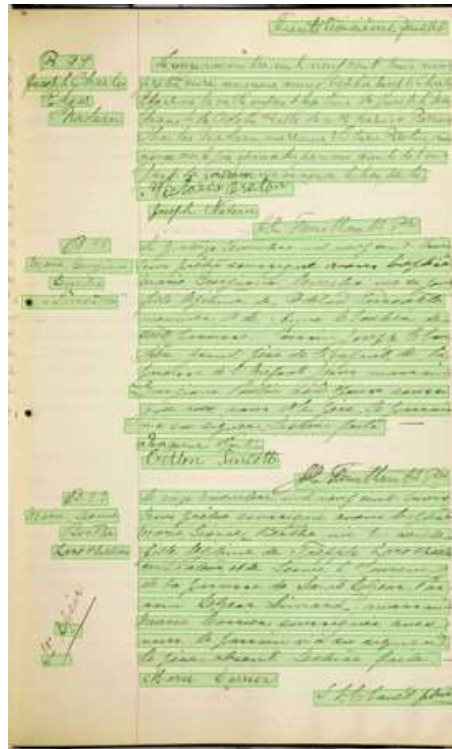
Décodeur: génère une séquence de caractère en prenant en compte l'image et les caractères déjà générés

Chaîne de traitement séquentiel :

Image



Détection des lignes



Reconnaissance

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 . Lechoe 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 , Etaient de l ' en - Vant . Le parrain n ' a su signer . Lechoe 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

Chaîne de traitement intégrée :

NOM		PRÉNOM		AGE	PROFESSION	ÉTAT	REMARQUES
4	Robert	Juliette	8	néant	l. fille		
5	Robert	Albertine	3				
6	Johan	Jules	19	domestique	chauger		
7	Belouet	Jeanne	22				
8	Beignet	Charles	32	les	chef		
9	Bezault	Hermance	27		épouse		
10	Bugnet	René	6		l. fils		
11	Biignet	Juliette	3		l. fille		
1	Portin	Jean	50	leer	chef		
2	Gronni	Marie	46		épouse		
3	Fortin	Juliette	19	néant	sa fille		
4	Portin	Georgette	12				
5	Portin	Maurice	10				
6	Vappereau	Andrée	1				
7	Grangeau	Philomèn	47	meunier	chef		



surname Robert, firstname Juliette, occupation néant, link l. fille, age 8, nationality française
 surname Robert, firstname Albertine, occupation idem, link idem, age 3, nationality idem
 surname Johan, firstname Jules, occupation domestique, link chauger, age 19, nationality idem
 surname Belouet, firstname Jeanne, occupation idem, link idem, age 22, nationality idem
 surname Beignet, firstname Charles, occupation j les, link chef, age 32, nationality idem
 surname Bezault, firstname Hermance, occupation idem, link épouse, age 27, nationality idem
 surname Bugnet, firstname René, occupation idem, link l. fils, age 6, nationality idem
 surname Biignet, firstname Juliette, occupation idem, link l. fille, age 3, nationality idem
 surname Portin, firstname Jean, occupation j leer, link chef, age 50, nationality idem
 surname Gronni, firstname Marie, occupation idem, link épouse, age 46, nationality idem
 surname Fortin, firstname Juliette, occupation néant, link sa fille, age 19, nationality idem
 surname Portin, firstname Georgette, occupation idem, link idem, age 12, nationality idem
 surname Portin, firstname Maurice, occupation idem, link idem, age 10, nationality idem
 surname Vappereau, firstname Andrée, occupation idem, link idem, age 1, nationality idem
 surname Grangeau, firstname Philomèn, occupation meunier, link chef, age 47, nationality idem

Modèle pleine page avec attention:

The image shows a page from a civil registry book with a grid of columns. The columns are labeled at the top: 'NOM ET PRÉNOM', 'NOM DE FAMILLE', 'PROFESSION', 'LIEU DE NAISSANCE', 'LIEU DE DÉCÈS', 'MARIAGE', and 'REMARQUES'. The entries are handwritten in French. A red rectangular box highlights the entry for 'Bousquet Elisabeth', which is the second row in the section starting with '176'. The entry contains the following information: '176 21 21 Bousquet Elisabeth', 'Bousquet', 'sans profession', 'Paris', 'Paris', and 'mariée'. Below the highlighted entry, there are several other entries, including '177 22 21', '178 23 21', '179 24 21', '180 25 21', '181 26 21', '182 27 21', '183 28 21', '184 29 21', '185 30 21', '186 31 21', '187 32 21', and '188 33 21'. The page is numbered '176' at the bottom left and '177' at the bottom right.

NOM ET PRÉNOM	NOM DE FAMILLE	PROFESSION	LIEU DE NAISSANCE	LIEU DE DÉCÈS	MARIAGE	REMARQUES
176 21 21	Bousquet	Elisabeth				mariée
177 22 21						
178 23 21	Couysson	Belle	Paris	Paris		
179 24 21						
180 25 21	Agard	jeune	Paris	Paris		
181 26 21						
182 27 21						
183 28 21						
184 29 21						
185 30 21						
186 31 21						
187 32 21						
188 33 21						
Année de la naissance						
189 1 21						
190 2 21						
191 3 21						
192 4 21						
193 5 21						
194 6 21						
195 7 21						
196 8 21						
197 9 21						
198 10 21						
199 11 21						
200 12 21						

○Bousquet ○Elisabeth ○femme du précédent ○Femme mariée

Identité, Altérité et IA

Les multiples biais de l'apprentissage automatique statistique influencent les représentations de l'identité et de l'altérité

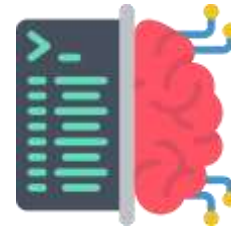
Données



Annotation



Apprentissage



Exploitation



Biais de population
Biais d'échantillon
Biais de genre
Biais de langue

Biais de
d'interprétation
Biais culturels

Biais d'algorithmes
Biais d'optimisation

Biais d'objectifs

Reconstituer la diversité des habitants de la France

Diversité originelle vs diversité artificielle



Reconstituer le monde social dans sa diversité

- Les recensements de population sont un outil idéal pour cela
 - ❖ Exhaustif: tout le monde est recensé (à quelques exceptions près + pertes d'archives).
 - ❖ Précis et riche par la masse: beaucoup d'individus observés un par un.
 - ❖ Des informations quantitatives plutôt que qualitatives: peu d'information sur beaucoup de personnes.
- Diversité des pratiques et des choix
 - ❖ La diversité des cas n'est pas simplement une question d'écriture, mais aussi d'usage, de pratiques, etc.
 - ❖ Un cultivateur pourra être « cultivateur propriétaire », « cultivateur exploitant », « agriculteur », etc.
 - ❖ Mais aussi: « cult », « cult^R », « C^{eur} », etc.

→ Diversité d'origine de la source.
- La reconnaissance automatique ajoute de la complexité
 - ❖ Ajoute de la variation à la variation: « cultivateur », « cultivrteur », « cultivathicaire », etc.

→ Diversité artificielle.

Reconstituer le monde social dans sa diversité (2) Normaliser

- L'IA produit des données brutes, pour produire de la connaissance, il faut les normaliser
 - ❖ Normaliser = lier à un référentiel.
 - ❖ Mais normaliser signifie aussi perte d'information.
- Référentiel interne : lié à la procédure de production humaine des données (historiques)
- Référentiel externe: existe ex ante et reflète des choix déjà faits
 - ❖ Géo-codage: évidence du référentiel mais complexité de la référence
 - ❖ Profession : évidence de la référence mais complexité du référentiel.

La masse... quelques exemples sur un (petit) corpus: la ville d'Orléans

- Liste nominative exhaustive pour 19 recensements de la ville d'Orléans (1836-1936, sauf 1856)

- ❖ Un peu moins d'un million de lignes (970 000).

- Diversité de l'information

❖ Nom:	180 983 dont 124 298 uniques,	top 10 = 2,64%	pour 90% = 84 223.
❖ Prénom:	55 482 dont 44 244 unique,	top 10 = 20,89%	pour 90% = 1 252.
❖ Nationalité:	628 dont 482 uniques,	top 10 = 99,75%	pour 90% = 1.
❖ Position:	9 900 dont 7 743 uniques,	top 10 = 80,50%	pour 90% = 23.
❖ Profession:	76 970 dont 64 572 uniques,	top 10 = 38,81%	pour 90% = 20 024.

Diversité d'origine vs diversité artificielle:

Les prénoms

Prénom	effectif
Françoise Julie	3
Françoïsette Vve	1
Frmnçoise	1
Frrnçoise	1
Frènçoise	1
Frènçoise Aurélie	1
Frènçoise Isabelle	1
Frénçoise	1
Félicie Françoise	2
Félicité Françoise	3
Félix Françoise	1
Hélène Françoise	4
Jauline Françoise	1
Jean Françoise	5
Jean Françoise Albertine	1
Jean Françoise Rachel	1
Jeanne Françoise	27
Vernande Françoise	1
Viançoise	1
Victoire Françoise	3
Victor Françoise Désiré	1
Virginie Françoise	1
Yvonne Françoise	1
Zean Françoise	1
Zrançoise	1

Diversité d'origine vs diversité artificielle:

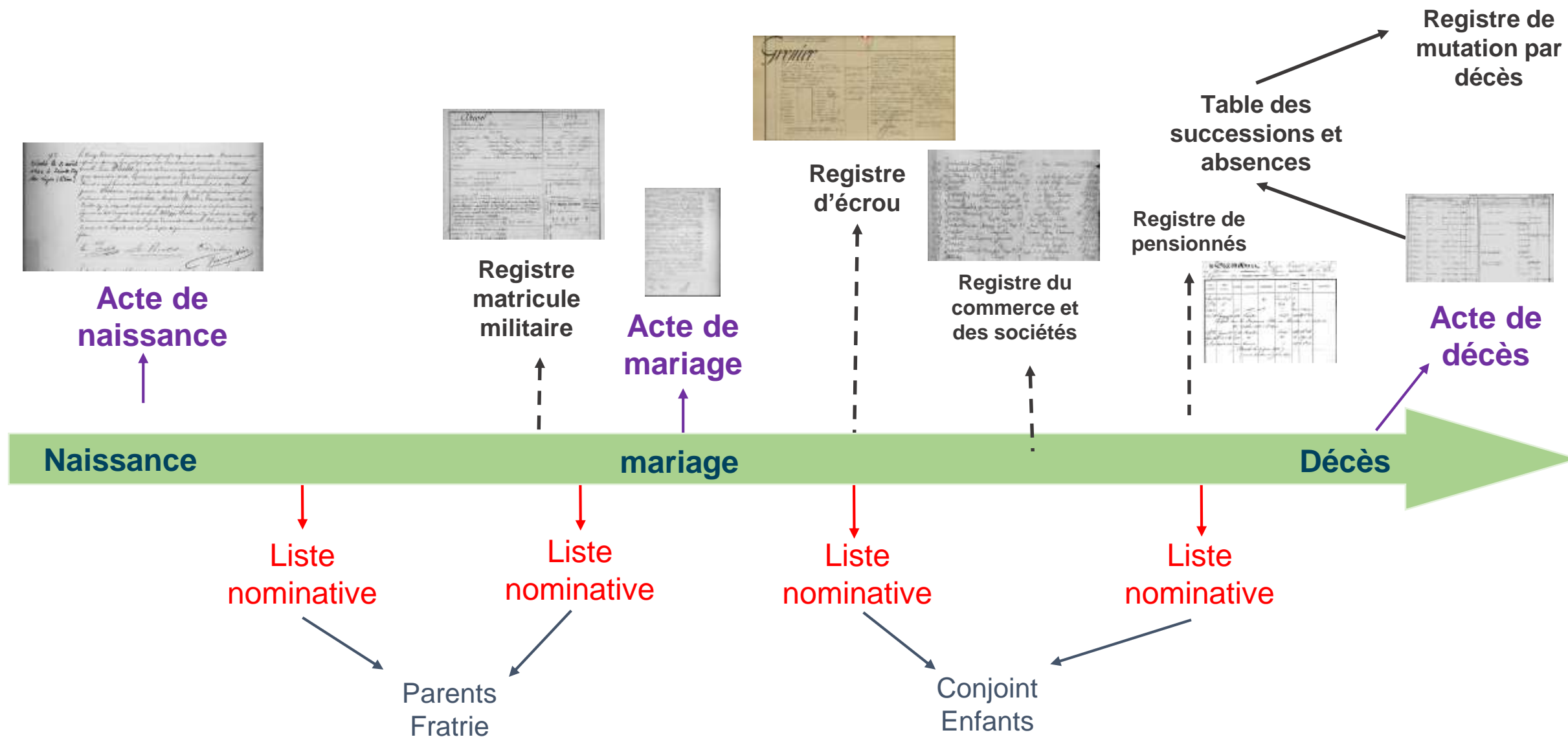
Position dans le ménage

Position	effectif
be fille	1
be nièce	1
be père	1
be soeur	1
be st Marie	1
beale fille	2
beale mère	1
beane	1
beau	1
beau chapouse	1
beau de compagnie	1
beau de muet	1
beau de pernier	1
beau de pine	1
beau des pêiné	1
beau d'Eufant	1
beau épouse	1
beau femme	1
beau fermier	1
beau fille	4
beau fils	544
beau fils ainé	1
beau fils au chef	1

Dissemination: retour aux archives, et au-delà

- Les données brutes mises à disposition par les archives
 - ❖ Sur une base nationale (*FranceArchives*), avec un moteur de recherche.
 - ❖ Sur les bases des Archives Départementales.
 - ❖ Liens directs vers les images.
- Une base de données organisées pour la recherche
 - ❖ Une base de données où les informations sont normalisées et classées (profession, lieu de naissance, etc.).
 - ❖ A terme, une base de données avec des appariements entre les individus.
- Et une ouverture vers d'autres sources: rassembler les archives individuelles autour du recensement?

Pour aller plus loin ...



SOC FACE

Merci de votre attention

