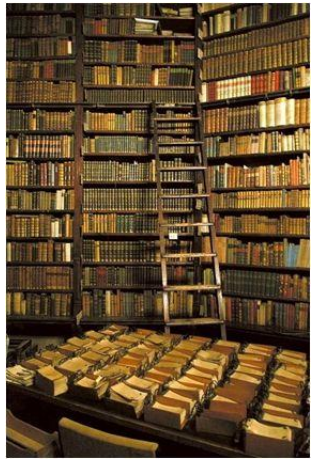
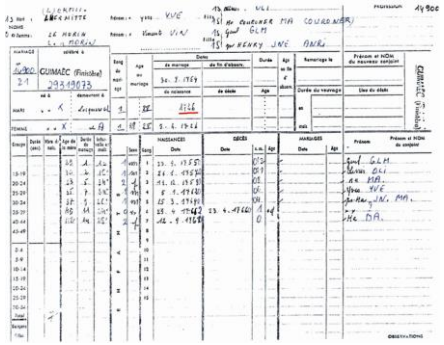
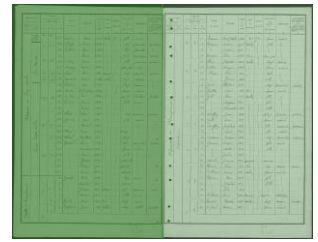
Construction de bases de données historiques:

Retours d’expérience







Lionel Kesztenbaum (INED et PSE)

Gilles Postel-Vinay (PSE)

Ateliers SoDUCo-BNF 4e séance

06/11/2023

Introduction: l'enquête historique, d'hier à aujourd'hui

- De la source (ou des sources) à la (ou aux) base(s) de données
  - Enjeux et défis archivistiques: interprétations et limites des sources, risques téléologiques, etc.
  - Enjeux et défis techniques: difficultés de la collecte, de la constitution et du maintien des bases de données, etc.
  - Enjeux et défis historiographiques: agrégation et normalisation, maintien et perte d'information etc.
- Pérennité et cumulativité
  - Un quatrième enjeu qui a pris dans les dernières décennies une importance croissante c'est la documentation des données et des choix opérés, explicités ou non. Sachant que ces choix portent en eux
    - le risque d'homogénéisation incontrôlée et de perte du lien à la source.
    - Mais qu'ils peuvent aussi permettre, sous certaines conditions, la mise en compatibilité entre les bases, leur transférabilité et partant la cumulativité des recherches.
- Retour d'expérience à de trois exemples de travaux sur l'histoire (quantitative) de la France
  - Numériser et diffuser les enquêtes industrielles du XIXe siècle: enquêtes industrielles.
  - Constituer un échantillon représentatif de la population française sur 150 ans: enquête TRA.
  - Numériser un siècle de recensement français: projet Socface.

Données historiques numériques - évaluations critiques et perspectives d'utilisations

- Nous sommes invités à nous interroger
  - sur le mode production de données historiques numériques à un moment où ce type de données connait des changements très rapides. Alors que depuis les années 70, elle reposait surtout sur une saisie classique d'informations extraites de divers corpus, elle prend désormais une autre ampleur en recourant à des processus automatisés susceptibles d'extraction voire d'enrichissement des sources initiales.
  - Et sur les conditions d'appropriation et de réutilisation de ces nouvelles « sources numériques ». Comment et à quelles conditions peut-on s'en emparer efficacement ? Comment les comprendre, évaluer leur « qualité » et leur rapport avec les sources initiales ? Comment comparer ou rendre cumulable les travaux très divers qui peuvent être issus de l'exploitation de ces jeux de données ?

Le cas d'une enquête des années 1980-90…

- Je vais admettre que les questions que posaient hier la constitution et la circulation de bases de données historiques ne sont pas très différentes de celles qui se posent aujourd'hui ; mieux, que se retourner, fût-ce brièvement, vers elles offre un miroir grossissant aux questions d'aujourd'hui
- Je ne vais pas remonter jusqu'aux époques héroïques des années 1960-70, au temps des bordereaux de dépouillement, des cartes perforées et des grands ordinateurs
- mais pour amorcer la discussion, je vais commencer par évoquer la constitution d'une de ces «sources numériques » réalisée dans les années 1980 sur les enquêtes « industrielles » de la France au XIXe siècle.

Pourquoi les enquêtes industrielles de la France du XIXe

- J'étudiais à l'époque le fonctionnement des marchés du travail dans la France d'alors, c'est-à-dire dans un monde où il était encore courant de basculer d'un emploi agricole à un emploi industriel au fil des saisons et pour cela j'avais cherché avec plusieurs collègues à construire et à analyser des corpus de données numériques élaborés à partir des enquêtes réalisées par la SGF sur l'agriculture et sur l'industrie.
- S'agissant de l'industrie, en particulier, la tâche paraissait simple puisque, dans les années 1840 puis dans les années 60, la SGF avait réalisé 2 enquêtes qui toutes deux répondaient à un questionnaire très proche et qui l'une et l'autre étaient publiées sous formes de tableaux.
  - Elles visaient un recensement exhaustif du secteur industriel.
  - Elles étaient organisées à l'échelle de l'établissement.
  - et chaque établissements était défini par un ensemble de variables quasi identiques dont une part sur l'emploi et sa saisonnalité

Constituer à partir de la source initiale imprimée la nouvelle «source numérique» paraissait particulièrement simple : transformer les 2 enquêtes en des tableaux croisés portant sur une dizaine de milliers d'établissements définis par une vingtaine de variables.

Dans un 2e temps, assortir ce premier travail d'un descriptif assez précis pour permettre dans de bonnes conditions l'appropriation et la réutilisation de ces nouvelles « sources numériques » paraissait aussi assez simple.

- De là un livre de présentation publié en 2000 (Chanut et al. Paris, Ed de l'EHESS, 2000) complété par un dvd puisque c'était à l'époque le seul moyen pour rendre accessible le travail à des tiers…

Mais, à l'expérience, constituer la nouvelle « source numérique » et la rendre réappropriable s'est révélé être une opération très loin d'aller de soi (1)

Les problèmes rencontrés ont été nombreux :

- Ces enquêtes ne sont pas exhaustives (cf. Paris en 1840)(cf. les mines…)
- Elles peinent à définir le champ qu'elles doivent couvrir. En particulier:
  - elles peinent à tracer la frontière entre industrie et artisanat. Cf. C. Lemercier et l'enquête de la Chambre de Commerce.
  - elles ont eu du mal à décider si elles devaient recenser tous les établissements ou fixer un seuil au-delà d'une certaine taille et, si oui, quel seuil (>1 ouvrier, ou à 10 ? Ou à 20 ?).
- Elles ont hésité sur le nombre de questions pertinentes.
- Enfin –et c'est peut-être plus surprenant – elles ont hésité sur la frontière entre « industrie » et « agriculture », une frontière très poreuse quand il s'agissait de l'emploi car, à l'époque, , comme je l'ai dit, une partie souvent importante de la main d'œuvre bascule d'un secteur à l'autre au rythme des saisons (au moment des récoltes).

Mais, à l'expérience constituer la nouvelle « source numérique » et la rendre réappropriable s'est révélé être une opération très loin d'aller de soi (2)

- Parmi les problèmes rencontrés certains ont été assez simples à documenter.
  - Les seuils.
  - Les frontières internes du secteur de l'industrie (définition des branches) etc…
- D'autres ont soulevé plus de nombreux problèmes. Pour les comprendre il faut se souvenir de ce que Jean-Claude Perrot appelait de ses vœux : une histoire matérielle de l'abstraction.

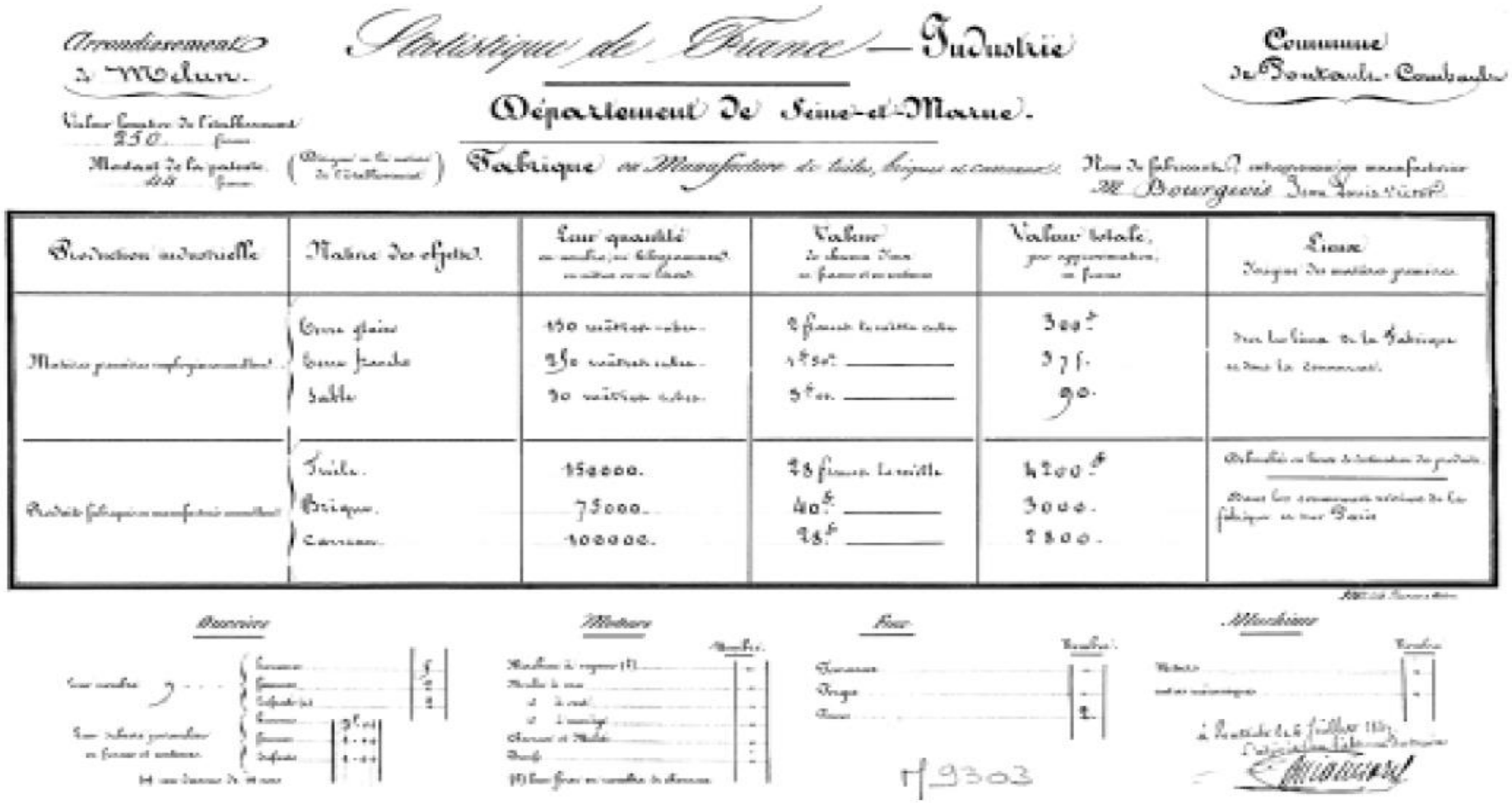
Une histoire matérielle de l'abstraction (1)

- Parenthèse : Lucien Febvre 1932 « **Les textes … mais tous les textes** … »
- « Les textes » : jusqu'ici on en a donné une définition restrictive : on est parti des volumes imprimés.
- « Mais tous les textes »… Et, pour cela, remonter de l'imprimé à la source primaire et d'abord aux questionnaires (quitte à aller aux AD… où ils ne sont pas tjs conservés). Pb :
- Certaines questions posées par les questionnaires pouvaient être immédiatement transcrites dans les colonnes des tableaux publiés. Dans ce cas tout était simple.
- Mais d'autres réponses ont dû être ré-élaborées en bureau par les responsables des enquêtes (ex: les branches)… selon des modalités qui n'allaient pas de soi…
- Enfin, les questionnaires comportaient plus de questions que celles reprises dans les publications (par exemple les questions sur l'origine des matières premières et la destination des produits. Le passage à l'imprimé a donc imposé des choix à la fois intellectuels et matériels : quelles variables retenir dans l'imprimé compte tenu de l'espace disponible (une page du volume comporte un nombre fini de colonnes).

Une histoire matérielle de l'abstraction (2) Des questions initiales à la source publiée

					Renseignement présent au questionnaire		Renseignement obtenu par agrégation en bureau	
					1840	1860	1840	1860
Département, Arrondissement					OUI	NON	non sauf exceptions	OUI
Commune					OUI	OUI	non sauf exceptions	NON
Nom du fabricant					OUI	NON	non sauf exceptions	NON
branche d'activité					NON	NON	OUI	OUI
Valeur locative et montant de la patente					OUI	OUI	non sauf exceptions	OUI
Matières premières		nature/quantités/valeur			valeur seulement	OUI	non sauf exceptions	OUI
Produit		nature/quantités/valeur			valeur seulement	OUI	non sauf exceptions	OUI
Débouché ou lieu de destination des produits					OUI	NON	NON	NON
Mois chômés					NON	OUI	NON	OUI
Ouvriers		H/F/E nb et salaire/jour			OUI	OUI	non sauf exceptions	OUI
Moteurs		eau, vent, animaux, vapeur			OUI	OUI	non sauf exceptions	OUI
Feux					OUI	OUI	non sauf exceptions	OUI
Machines					OUI	OUI	non sauf exceptions	OUI

Bulletin individuel de la 1e enquête



Arrondissement de Melun.

Statistique de France — Industrie

Département de Seine-et-Marne.

Commune de Fontaine-Couronne

Valeur locative de l'établissement 250 francs

Montant de la patente 28 francs (Rayer la mention inutile de l'établissement)

Fabrique ou Manufacture de tuile, brique et carreau

Nom du fabricant, entrepreneur ou manufacturier M. Bourgeois Jean Louis Victor

Production industrielle	Nature des objets	Leur quantité en nombre, par chargement ou mètre cube ou linéaire	Valeur de chacun d'eux en francs et centimes	Valeur totale, par approximation, en francs	Lieux Origine des matières premières
Matières premières employées	Terre glaise Terre franche Sable	150 mètres cubes 250 mètres cubes 30 mètres cubes	2 francs le mètre cube 1f50c 5f en	300f 375f 90	Sur les lieux de la fabrique et dans la commune
Produits fabriqués ou manufacturés vendus	Tuile Brique Carreau	150000 75000 100000	28 francs le mille 40f 28f	4200f 3000 2800	Recueillis au lieu de fabrication des produits Dans les environnements voisins de la fabrique et sur Paris

Ouvriers

Moteurs

Feu

Machines

Nº 9303

Bulletin individuel de la 2e enquête

Département de Seine et Marne

Arrondissement de Fontainebleau

Canton de Chateaulandon

Commune de Beaumont

MINISTÈRE DE L'AGRICULTURE, DU COMMERCE ET DES TRAVAUX PUBLICS.

DÉNOMBREMENT DÉCENNAL DE L'INDUSTRIE MANUFACTURIÈRE.

ANNÉE 1860.

BULLETIN INDIVIDUEL.

Nom du propriétaire de l'établissement ou de la société qui l'exploite

Rousselot Armand

Nature de l'industrie :

Fabrique de pâte de guimauve

Observation importante. Le présent recensement industriel est fait en exécution des décisions du Gouvernement, qui ont prescrit que cette vaste opération aurait lieu tous les dix ans. Le dernier remontant à 1850, le moment est venu de procéder à celui de 1860. Le dénombrement de cette année offrira un intérêt tout particulier que n'avaient pas les précédents, en permettant de constater, en 1870, les effets qu'aura pu produire sur l'industrie manufacturière en France, l'exécution du traité de commerce avec l'Angleterre. A ce titre surtout, il appelle l'attention de nos fabricants qui s'empresseront, nous n'en doutons pas, de fournir, avec toute l'exactitude possible, les renseignements demandés par le présent bulletin, renseignements destinés à être contrôlés par les chambres de commerce et les chambres consultatives des arts et manufactures.

1. Par valeur vénale de l'établissement, il faut entendre le prix qu'il atteindrait probablement, s'il était mis en vente.
2. Il s'agit ici des enfants de moins de quinze ans, employés dans les manufactures à un titre quelconque et recevant un salaire quelque faible qu'il soit.
3. Indiquer ces quantités en mesures métriques, et, autant que possible, au poids.
4. S'il s'agit, par exemple, d'une fabrique de drap, indiquer séparément dans la colonne d'observations, les quantités de laines françaises et étrangères employées annuellement, à cet effet afin de faire concorder ces diverses quantités avec le total de la colonne précédente.
5. Indiquer spécialement dans cette colonne, le nombre des ouvriers qui mettent en mouvement des moteurs quelconques et ne prennent part que sous cette forme à la fabrication.
6. Il s'agit ici des animaux (chevaux, bœufs, etc.) qui mettent en mouvement un moteur quelconque (manège, etc.).
7. On devra considérer comme machine, tout appareil mécanique qui fabrique directement ou contribue à fabriquer tout ou partie d'un produit et reçoit l'impulsion d'un moteur à eau, à air, à feu, à bras, etc.
8. Les énumérer dans la colonne d'observations et, au besoin, au verso du présent bulletin.

Strasbourg, imprimerie de Ve Berger-Levrault. — 2000.

VALEUR vénale de l'établissement. (1)	NOMBRE MOYEN ORDINAIRE DES OUVRIERS ADULTES. Hommes travaillant à la journée	Hommes travaillant à la tâche (exclusivement pour la maison)	Hommes travaillant à la tâche (pas exclusivement pour la maison et pour d'autres patrons)	Femmes travaillant à la journée	Femmes travaillant à la tâche (exclusivement pour la maison)	Femmes travaillant à la tâche (pas exclusivement pour la maison et pour d'autres patrons)	SALAIRE DES OUVRIERS ADULTES NON NOURRIS travaillant à la journée. Maximum. Hommes	Maximum. Femmes	Minimum. Hommes	Minimum. Femmes	Moyen ou ordinaire. Hommes	Moyen ou ordinaire. Femmes	NOMBRE des enfants employés dans l'établissement et nourris ordinairement. (2) Garçons	Filles	SALAIRE HABITUEL de ces enfants. Garçons	Filles	QUANTITÉS moyennes des matières premières employées annuellement. (3)	QUANTITÉS d'après les PROVENANCES. France	Étranger. (4)	VALEUR à l'unité de ces matières.	CHARBONS. Quantités moyennes des consommations annuellement faites. (5)	VALEUR réelle à l'usine de ces produits (déduction faite de l'emploi).	NATURE et poids moyen annuel du combustible employé.	DÉPENSE moyenne annuelle en combustible.	MOTEURS. Moulins à eau. Nombre	Force réelle en chevaux-vapeur	Moulins à vent. Nombre	Force réelle en chevaux-vapeur	Manèges. Nombre	Force réelle en chevaux-vapeur	Machines à vapeur. Nombre	Force en chevaux-vapeur. (6)	Forces animées. Nombre des Chevaux. (7)	Animaux. (8)	MACHINES. (9) Leur nombre	Leur valeur. (10)	Nombre des broches des machines à filer.	NOMBRE DES HAUTS FOURNEAUX.	CHOMAGES. Circonstances ayant amené des chômages réguliers	Époque habituelle de ces chômages.	Leur durée moyenne.	Nombre total des ouvriers pendant leur durée.	OBSERVATIONS.
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
10000	8	"	"	1	6	"	4,50	4,50	2	1,50	3,50	3,50	2	"	1,50	"	17,000	"	"	"	3,000 Charbon	2400,00	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	"	Les de Chômage

17,000 Charbon

Une histoire matérielle de l'abstraction (3)

- Exemple 1/ Quand ont été publiés les résultats de la première enquête, une page imprimée ne pouvait contenir toutes les variables et il a fallu choisir entre publier le nom des propriétaires et leur adresse (deux colonnes) **ou** publier les données sur la localisation des consommations intermédiaires et des produits :
  - La décision a été prise de renoncer aux secondes pour publier l'identité et le domicile des propriétaire, puisque la logique d'ensemble reposait sur le choix de publier les information établissement par établissement.
  - Conséquences :
    - Certaines positives : on observe ainsi non seulement les établissements mais aussi les entreprises (puisque l'enquête identifie son propriétaire) et son statut juridique.
    - D'autres négatives : ont ainsi surgi de nombreux conflits avec les enquêtés (très hostiles à ce que les renseignements recueillis soient rendus publics).
    - Si nombreux que c'est la décision inverse qui est prise en 1860 quitte à changer la présentation des résultats et ne les publier qu'agrégés pour les établissements d'une même branche dans un même arrondissement. Ce faisant, on perd notamment la distinction entre établissement et entreprise.

Une histoire matérielle de l'abstraction (4)

- Exemple 2/ ces deux enquêtes prennent la suite d'opérations antérieures plus ou moins abouties mais avec un questionnaire voisin.
- Elles cherchaient notamment à cerner les flux saisonniers de la main d'œuvre entre industrie et l'agriculture.
  - La 1e enquête industrielle a minoré ce point et décidé de privilégier l'information sur le nom et le domicile des propriétaires choix qui, on vient de le dire, structure la publication.
  - La 2e enquête ayant renoncé à individualiser les propriétaires des établissements, l'espace laissé libre dans la page de la publication a pu être employé à publier l'information sur l'activité saisonnière et donc les mois où est employé la main d'œuvre. De là de nombreuses difficultés pour agréger à l'échelle de chaque branche les résultats de chaque établissement tels qu'ils figuraient au questionnaire.
  - Pb intéressant, mais demandant de tels développements que j'ai fini par renoncer à les expliciter dans le livre sus mentionné…

Bilan et perspectives (1)

- Le premier avantage de ces nouvelles « sources numériques » a été de rendre accessibles les info micros sur les établissements (ou au moins sur les établissements d'une même branche dans un même arrondissement). C'était un saut qualitatif par rapport aux recherches antérieures qui n'avaient retenu de ces enquêtes que leurs résultats agrégés à l'échelle du pays pour établir les comptabilités nationales rétrospectives, ce qui était l'un des grands objectifs de toute une génération à l'époque.
- L'approche s'est ensuite révélée fructueuse quand les recherches ont porté sur les établissements de branches les mieux cernées par les enquêtes (le textile cotonnier en particulier: cf. les articles de Pierre Sicsic et, récemment, ceux de Reka Juhasz (reliant pour ce type d'établissements les enquêtes qui se sont succédées de l'Empire à celle de 1840)(effet durables du blocus continental; quel rythme d'adoption de techniques nouvelles).

- Mais on peut imaginer bien d'autres objectifs. Pour cela, à nouveau revenons à Lucien Febvre. « **Les textes** » dit-il (ici les enquêtes publiées)…
- puis il précise « **Les textes …mais tous les textes** » (ici les questionnaires originaux)…
- Mais il ajoute :

Bilan et perspectives (2)

- « **Les textes, sans doute : mais tous les textes… Les textes, évidemment : mais pas rien que les textes. Les documents aussi, quelle qu'en soit la nature »** ; et Febvre de citer comme exemple les documents qu'utilise … la linguistique … ou la palynologie **« en quête de pollen millénaire ». « L'histoire fait son miel avec lui. L'histoire qui s'édifie, sans exclusion, avec tout ce que l'ingéniosité des hommes peut inventer et combiner pour suppléer au silence des textes, aux ravages de l'oubli. »**.
  - Febvre en 1932 pense à déplacer les interrogations des historiens en recourant à des croisements entre des sources classiques et des sources nouvelles : en 1967 avec ses recherches sur le climat Le Roy Ladurie (1967) se saisit de ce programme mais aujourd'hui d'autres enquêtes deviennent possibles, en particulier en tirant parti des informations non ambigües des enquêtes.
  - Cf. l'enquête de 1840 dont les informations localisées (commune par commune) peuvent être réanalysées en les croisant avec les sources qui désormais renseignent sur les conditions climatiques locales, sur la présence d'une rivière ou d'un fleuve (plus ou moins puissance) et donc l'existence de ressources hydrauliques en un temps où c'est encore l'énergie dominante…etc.

L'enquête TRA: une enquête, des données, un programme de recherche

- Une base de données historiques qui a 40 ans d'histoire
  - Le projet initial: l'enquête « 3000 familles », portée par Jacques Dupâquier au début des années 80.
  - Différentes extensions, en premier lieu autour des données fiscales (Denis Kessler).
- Un contexte de développement rapide de bases de données historiques individuelles
  - Projets initiaux de la démographie historique: Louis Henry, Princeton European Fertility Project, The Cambridge Group for the History of Population and Social Structure, etc.
  - Grandes bases de données individuelles: Balsac et PRDH, Scania database, Eurasian project, etc.
- Changement d'échelle et de finalité
  - Vers une logique cumulative : l'enquête TRA (appariement avec les données militaires, recensement, retraités de la fonction publique etc.).
  - Une base de données individuelles représentative.
  - Avec le temps, la base a changé d'objectifs : **une matrice**.

L'enquête TRA : une base individuelle et nationale (1)

- Tous les individus dont le noms de famille commence par les lettres TRA.
- Deux sources principales
  - L'état civil et surtout les mariages : Jacques Dupâquier puis Jean-Pierre Pélissier et Danielle Rébaudo à l' EHESS, CNRS, INRA et à l'INED – "l'enquête des 3000 familles" 1984-1992.
  - Les sources fiscales (l'Enregistrement : taxation des successions) : Denis Kessler au CNRS et à l'EHESS, 1985-1992, puis Gilles Postel-Vinay et al. at INRA-EHESS-INED, depuis 1995.
- Un échantillon représentatif de la population française :

échelle ≈ 1 pour 1100

L'enquête TRA : une base individuelle et nationale (2)

- Pourquoi TRA?
  - Représentatif à l'échelle nationale.
  - Stable dans le temps.
  - Facile à lire.
  - Effectifs suffisants, mais pas excessifs.
- Collecte des données
  - Tous les mariages de 1802-1902 – et un sous-échantillon entre 1903-1993.
  - Sources fiscales : tous les morts entre 1798-1939 et un sous-échantillon 1940-1959.
- Reconstitutions de familles.
  - "3000 familles" : l'objectif initial à partir de 3000 couples formés entre 1800 et 1830 choisis au hasard
  - Reconstitutions en lignées masculines descendantes.

L'enquête TRA : une base individuelle et nationale (3)

- Un double objectif
  - Trajectoire de vie individuelle (naissance – mariage – décès)
  - Reconstitutions de familles sur 7 générations.
- De la démographie à l'ensemble de l'histoire économique et sociale
  - Transformations de la société française dans la longue durée.
  - Industrialisation, urbanisation, vieillissement.
  - Inégalités, mobilité sociale.
- Changement de logique : appariement avec d'autres sources nominatives, une logique cumulative.

Combiner de multiples sources

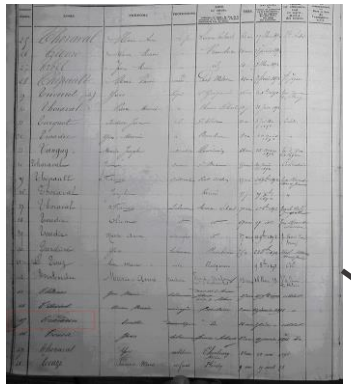
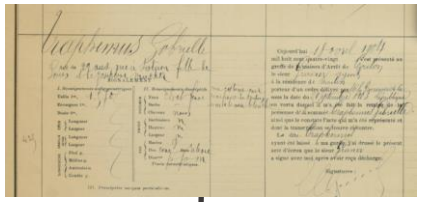
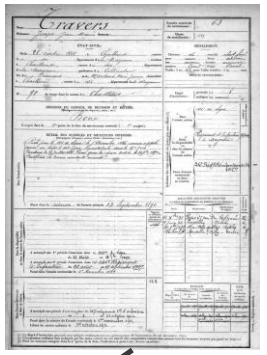


Table des successions et absences

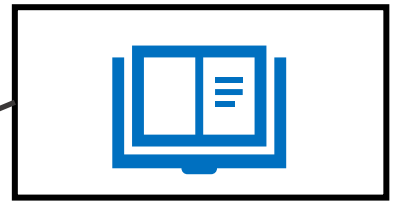
Registre de mutation par décès



Registre d'écrou



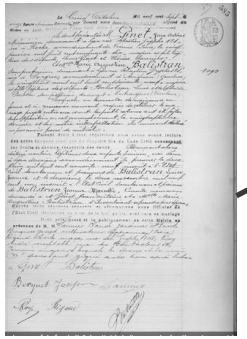
Registre matricule militaire



Autres?

Individu

Famille



Acte de mariage

Registre de pensionnés, dossier de salariés, etc.

Actes de naissance, de décès

Autres sources généalogiques

Un projet collectif

- Construction de la base de données démographiques : Jacques Dupâquier, Jean-Pierre Pélissier, Myriam Provence..... et des centaines de bénévoles (correspondants de l'enquête).
- Construction de la base de données fiscales : Denis Kessler, Eliane Monge.
- Élargir la reconstitution de la famille : Pascal Cristofoli, Clément Dherbécourt, Jean-Brieu Delbos, Viviani Huerta, Danièle Rébaudo, Eric Zolotukine.
- Élargir l'ensemble des données sur la richesse : Linda Amiri, Guillaume Bazot, Charlotte Coutant, Sarah Favre, Alena Lapatniovna.
- La reproduction familiale : Noël Bonneuil, Michel Dürr.
- Les migrations dans la durée : Didier Blanchet, Paul-André Rosental.
- La mobilité sociale à long terme : Marco Van Leeuwen, Ineke Maas.
- Micro-étude de la transmission de la richesse : Luc Arrondel, Cyril Grange.
- Evolution de la structure de la richesse et des professions : Jérôme Bourdieu, Lionel Kesztenbaum, Martha Ménendez, Gilles Postel-Vinay, Akiko Suwa-Eisenman.

Bilan et perspectives

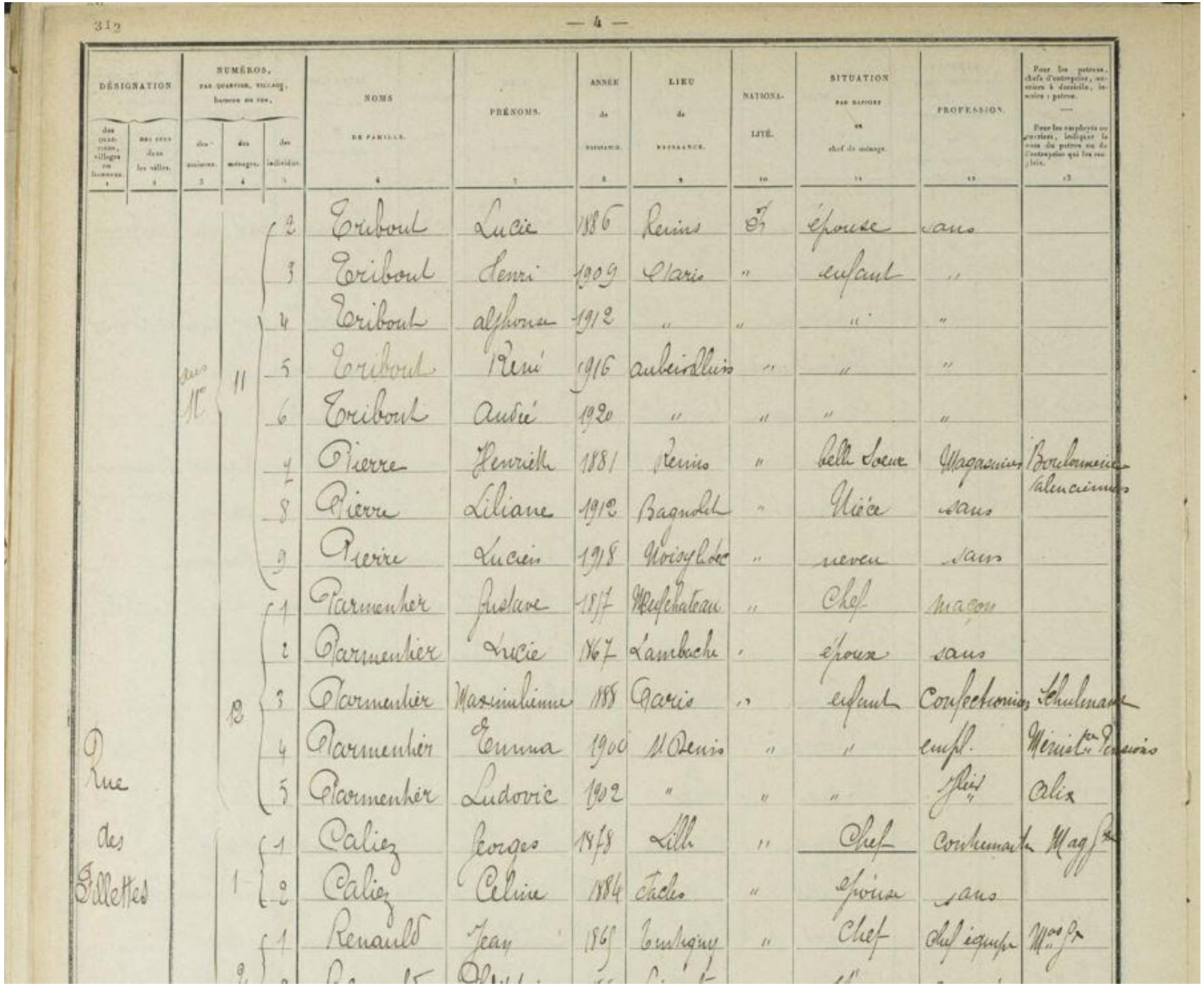
- Addition de nouvelles sources à partir de la structure nominative: une « enquête » construite pour être diffusée
  - Un socle commun qui ouvre de nombreuses possibilités d'extension (mais les limite aussi: il faut que la source soit (aisément) accessibles nominativement).
  - Et en même temps un élargissement qui implique une appropriation de l'existant, parfois complexe, souvent long.
- Suivre les individus dans le temps et l'espace
  - Extension (et généralisation) du principe initial de l'enquête.
  - Reconstitution de trajectoires individuelles et familiales.
  - Ouverture à de nombreuses thématiques.
- Mais aussi nombreux défis: instabilités des noms (et prénoms), données manquantes, stabilité des sources et du vocabulaire…

Retranscrire les listes individuelles du recensement : le projet Socface

- Le projet Socface: 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 des individus ayant vécu en France entre 1836 et 1936.
  - Construire une histoire économique et sociale à la fois individuelle et nationale.
  - Grande souplesse des approches: sur une seule année/coupe, plusieurs années combinées, avec appariement des individus entre les recensements ou non…
- Un projet de taille?
  - 20 millions d'images à traiter + décentralisation (une centaine de dépôt d'archives).
  - Un projet rendu possible par les méthodes automatisés.
  - Mais ces méthodes modifient aussi la substance du projet: façon de travailler, rendu en masse…

Une liste nominative:

Aubervilliers, 1921



312 — 4 —

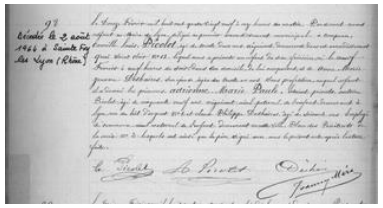
DÉSIGNATION des quartiers, villages ou hameaux.	DÉSIGNATION des rues dans les villes.	NUMÉROS par quartier, village, hameau ou rue, des maisons.	NUMÉROS des ménages.	NUMÉROS des individus.	NOMS DE FAMILLE.	PRÉNOMS.	ANNÉE de naissance.	LIEU de naissance.	NATIONALITÉ.	SITUATION par rapport au chef de ménage.	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	"	"	"	"	
			5		Tribout	Rémi	1916	Aubervilliers	"	"	"	
		sans N°	11	6	Tribout	André	1920	"	"	"	"	
				7	Pierre	Henriette	1881	Reims	"	belle Sœur	Magasinier	Boulangerie Valenciennes
				8	Pierre	Liliane	1912	Bagnolet	"	Nièce	sans	
				9	Pierre	Lucien	1918	Noisy-le-Sec	"	neveu	sans	
				1	Parmentier	Gustave	1877	Montcheteau	"	Chef	maçon	
				2	Parmentier	Lucie	1867	Lambach	"	épouse	sans	
			13	3	Parmentier	Maximilienne	1888	Paris	"	enfant	Confectionneuse	Schulmann
				4	Parmentier	Emma	1900	St Denis	"	"	empl.	Ministère Pensions
Rue				5	Parmentier	Ludovic	1902	"	"	"	élève	Alix
des				1	Caliez	Georges	1878	Lille	"	Chef	Contremaître	Magasin
Gillettes			1	2	Caliez	Céline	1884	Tarbes	"	épouse	sans	
				1	Renaud	Jean	1865	Tourcoing	"	Chef	Chef équipe	Magasin
			2									

Retranscrire les listes individuelles du recensement : le projet Socface

- Le projet Socface: 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 des individus ayant vécu en France entre 1836 et 1936.
  - Construire une histoire économique et sociale à la fois individuelle et nationale.
  - Grande souplesse des approches: sur une seule année/coupe, plusieurs années combinées, avec appariement des individus entre les recensements ou non…
- Un projet de taille?
  - 20 millions d'images à traiter + décentralisation (une centaine de dépôt d'archives).
  - Un projet rendu possible par les méthodes automatisés.
  - Mais ces méthodes modifient aussi la substance du projet: façon de travailler, rendu en masse…

Le projet Socface dans un temps long

- Le projet vise à être le point de départ des études quantitatives sur la France avant la seconde guerre mondiale.
- Peut aussi permettre de faire le lien avec les données contemporaines (seconde moitié du XXe siècle).
- Réflexion sur les conditions de réutilisation des données
  - Penser en amont la diffusion, techniquement et scientifiquement.
  - Transparence des choix qui sont fait + en faire le moins possible: tentation de « corriger les sources ».
  - Documenter au maximum les étapes de la constitution de la base.
- Développer des indicateurs, outils, marqueurs (…) pour faciliter la réutilisation
  - Indices de « qualité » aisément mobilisables par les sciences sociales.
  - Documenter l'idiosyncrasie de la source: rendre possible comparaison et réutilisation sans les préempter.



Acte de naissance

Registre matricule militaire

Acte de mariage

Registre d'écrou

Registre du commerce et des sociétés

Registre de pensionnés

Table des successions et absences

Registre de mutation par décès

Acte de décès

Naissance — mariage — Décès

Liste nominative → Parents, Fratrie

Liste nominative → Parents, Fratrie

Liste nominative → Conjoint, Enfants

Liste nominative → Conjoint, Enfants

Le projet Socface dans un temps long

- Le projet vise à être le point de départ des études quantitatives sur la France avant la seconde guerre mondiale.
- Peut aussi permettre de faire le lien avec les données contemporaines (seconde moitié du XXe siècle).
- Réflexion sur les conditions de réutilisation des données
  - Penser en amont la diffusion, techniquement et scientifiquement.
  - Transparence des choix qui sont fait + en faire le moins possible: tentation de « corriger les sources ».
  - Documenter au maximum les étapes de la constitution de la base.
- Développer des indicateurs, outils, marqueurs (…) pour faciliter la réutilisation
  - Indices de « qualité » aisément mobilisables par les sciences sociales.
  - Documenter l'idiosyncrasie de la source: rendre possible comparaison et réutilisation sans les préempter.

Conclusion (1) de la question de recherche à l'enquête, et retour

- Diversité des sources et des « enquêtes »
  - Enquête industrielle: retracer le cheminement de l'enquête.
  - Enquête TRA: comprendre les apports et limites de chaque source.
  - Projet Socface: gérer la masse et la diversité d'une source unique.
- Mais des questionnements mutuels
  - Quelle partie du monde social est dévoilée par l'enquête et comment percevoir au mieux extension et limites de l'observatoire.
  - A quelle étape faire la consolidation des informations, le codage des données, la sélection des variables, etc.
  - Linking entre les sources: objectifs et limite, sélection(s) qui en résulte.

Conclusion (2) Complémentarités et cumulativité

- Complémentarités à différentes échelles
  - Transmission et réutilisation de l'enquête.
  - Enquêtes individuelle, qualitatives et quantitatives.
  - Données contextuelles: caractéristiques des communes (population, position géographique)) et leurs évolutions (construction d'une école, d'une gare, d'une usine…).
- Effet de la reconnaissance automatique pour le développement de larges bases de données (historiques)
  - Une forme d'arbitrage entre des données riches (denses) et des données vastes.
  - Question de la représentation: qui est présent? qui est absent? pourquoi?
  - Variation dans la distance entre la source et la base de données… entre la base et les objets.
  - Déconnexion entre la production et la valorisation des données.