Statistiques fondées sur des données administratives : Esquisse d'un cadre général

Document de travail

N° M2025-03 – Juin 2025

Heidi KOUMARIANOS
Pascal RIVIÈRE

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES
Série des documents de travail « Méthodologie Statistique »
de la Direction de la Méthodologie et de la Coordination Statistique et Internationale

M 2025/03

Statistiques fondées sur des données administratives : Esquisse d'un cadre général

Heidi KOUMARIANOS
Pascal RIVIÈRE

Insee

JUIN 2025

Remerciements :
Les auteurs souhaitent remercier Hélène Chaput, Olivier Haag, Pierre Lamarche, Joël Bizingre, Mylène Chaleix, Élise Coudin, Pascale Breuil, Éric Lesage et Corinne Prost pour leur relecture commentée et active du rapport.

Direction de la méthodologie et de la coordination statistique et internationale
Département des Méthodes Statistiques -Timbre L001 -
88 Avenue Verdier - CS 70058 - 92541 Montrouge Cedex - France -
Tél. : 33 (1) 87 69 55 00 - E-mail : DG75-L001@insee.fr - Site Web Insee : http://www.insee.fr

Résumé

La mobilisation de données administratives pour la production de statistiques et les questions qui l'accompagnent ne sont pas nouvelles dans les statistiques officielles, mais elles ne donnent pas lieu à un cadre méthodologique équivalent à ce qui existe pour les enquêtes. Cet usage peut parfois être perçu à tort comme immédiat, puisque les données « existent » déjà, et qu'elles ne nécessitent pas la mise en œuvre par le statisticien d'un processus de collecte pour les obtenir.
Cependant, cette absence de collecte n'est qu'une illusion de simplicité : elle est aussi le signe d'une perte de maîtrise sur les modalités d'obtention des données, leur temporalité, la définition des variables, des nomenclatures, les possibilités de vérification … Au total, pour ces nouvelles sources de données, c'est tout un pan du processus de production statistique qu'il faut repenser, pour tenir compte des nouvelles questions qu'elles soulèvent.

Les données administratives sont liées, par leur nature même, à un univers particulier, avec ses objectifs, son langage, ses catégories, ses dynamiques. Sous-produits de l'activité de l'administration, elles ne sont en aucun cas « données », et peuvent se révéler très éloignées de l'univers de l'utilisateur. Leur utilisation à des fins statistiques soulève des problématiques de qualité particulières, le concept de qualité étant subordonné à l'usage. Elle requiert une transition du monde administratif au monde statistique, un détachement de l'un pour se réattacher à l'autre. Cela ne peut se faire sans difficultés, sans frottements : c'est la notion de *data friction*.

Pour effectuer cette transition rigoureusement et dans de bonnes conditions, une grille d'analyse est nécessaire. Elle se présente sous la forme de 5 axes : objet (ou unité statistique), population et champ, variable, domaine (ou catégorisation), temporalités.

On propose ici une démarche fondée sur cette grille, et qui se décompose en 3 phases : acquisition, transformation, traitement statistique.

- La phase d'acquisition part des données du SI administratif, non conçu pour les statistiques, car hétérogène, épars, mouvant, lié à un usage métier. Elle vise à rassembler, documenter, filtrer mais aussi figer, pour arriver à une « source administrative » utilisable.
- La phase de transformation vise à passer de l'univers administratif à l'univers statistique.
- La phase de traitement statistique correspond à une étape classique, que l'on retrouve aussi dans les enquêtes.

Pour chacune des phases, on met en évidence l'importance de « boucles de rétroaction », dans l'esprit du *data tracking* : pour assurer la qualité des données, donc des statistiques produites, il faut effectuer des retours arrière. C'est d'autant plus difficile que, contrairement aux enquêtes, on n'a pas la pleine maîtrise des concepts. Il en découle de nombreuses vérifications, à plusieurs niveaux : en cas d'anomalie, cela conduit à remonter en amont dans le processus, y compris jusqu'à la source. Ces boucles sont de natures très différentes selon les phases. Au total, en raison du besoin de qualité et d'explicabilité des résultats, l'enchaînement des phases n'a rien de linéaire.

Si le document propose un cadre général, sa mise en pratique peut nécessiter certains ajustements en lien avec des contraintes pratiques (de volume, ou de temps, par exemple). Adapté à une situation de mono-source administrative, il peut s'étendre : les principes proposés s'appliquent aussi en bonne partie aux données privées, avec des difficultés supplémentaires (champ, confidentialité, conventions, coût, ...). Ils peuvent aussi être replacés dans un contexte multi-sources, en associant données administratives et enquêtes, ce qui pose là aussi de nouvelles questions.

Mots clés : données administratives, qualification, métadonnées, processus, transformation, boucle de rétroaction, GSBPM

Classification JEL : C80, C40

Table des matières

Introduction

La formalisation de l'échantillonnage probabiliste, dans les années 1930 (Neyman 1934) a rendu possible la réalisation d'enquêtes à coûts raisonnables, car appliquées à de petites populations (Ardilly 2006). Cette démarche s'est largement répandue, devenant la norme pour les statisticiens publics, en parallèle à la pratique des recensements (Desrosières 1993). Dans le monde entier, un vaste savoir-faire s'est peu à peu échafaudé, enrichi, consolidé, autour d'un corpus de méthodes (d'échantillonnage, de conception de questionnaire, de traitement de valeurs manquantes, …), de pratiques, reconnu par la communauté académique, officialisé dans les instances internationales de statisticiens. Un « métier de statisticien » a ainsi progressivement émergé (Volle 1980).

Parallèlement, chaque pays cherchait à exploiter ses propres données administratives à des fins de statistique publique, en complément des enquêtes. Certains, notamment les pays nordiques, les ont intégrées, historiquement, de façon massive (Wallgren, Wallgren 2007). En France, l'Insee mobilise des données administratives depuis sa création. Les déclarations de données sociales, notamment, constituent une source majeure largement et régulièrement utilisée depuis les années 50. De manière générale, dans un contexte où les enquêtes sont de plus en plus difficiles à réaliser (Meyer, Mok, Sullivan 2015), les données administratives jouent un rôle essentiel en statistique publique. Elles permettent : de construire des répertoires de référence (Wallgren, Wallgren 2016, Lefebvre 2024), d'apparier avec des données d'enquêtes et ainsi d'enrichir la palette des données (Koumarianos, Lefebvre, Malherbe 2024, Brion, 2011), de s'inscrire dans un système général d'acquisition de données pour un usage statistique (Bakker, Van Rooijen, Van Toor 2014), de pallier des non-réponses (Bycroft, Mateson-Dunning 2020), d'évaluer des sources de biais (D'Aurizio, Papadia 2019), de fournir des totaux pour réaliser un calage sur marges (Deville, Särndal 1992), d'améliorer les mesures de population à un niveau géographique fin (Zhang 2021), entre autres.

Un ensemble de données administratives peut aussi être utilisé seul pour l'élaboration de statistiques. C'est sur ce sujet précis que va se concentrer ce document, afin de bien caractériser les composantes de ce processus, en tenant compte des propriétés particulières des données administratives. On exclut donc ici toute réflexion sur les appariements, sur les combinaisons de sources, sur les contrôles, le calcul de précision… De même, on laissera de côté toute une série de thèmes régulièrement attachés aux données administratives dans la littérature académique : confidentialité, acceptabilité, éthique, cadre légal, ...

Lorsqu'on élabore des statistiques à partir de données administratives, on se trouve dans une situation étrange où « les données existent déjà », ce qui semble, en première approche, faciliter considérablement le travail en comparaison à une enquête : il n'y a ni questionnaire à construire, ni échantillon à sélectionner, ni processus de collecte à dérouler. En réalité, on se retrouve dans un contexte fondamentalement distinct, où les exigences sont d'une autre nature et où les processus génériques ne sont pas extrêmement cadrés, contrairement à ce qui se passe pour les enquêtes. Pour celles-ci, le General Statistical Business Process Model (GSBPM) propose désormais un modèle reconnu et partagé par les instituts nationaux de statistique (Unece 2019), dont seule une partie est pertinente pour le cas des sources administratives.

On constate ainsi que s'il existe de nombreux papiers parlant de statistiques produites à partir de données administratives sur des jeux de données précis[1], si l'on rencontre quelques papiers très généraux (Hand 2018, Zhang 2012, Rouppert 2005), on ne trouve pas, ou peu, de véritable cadre conceptuel pour les statistiques fondées sur des sources administratives en général. Or l'utilisation de nouvelles sources en lieu et place de données d'enquête constitue un changement beaucoup plus

1 Sujet souvent abordé aux Journées de méthodologie statistique de l'Insee, par exemple.

profond qu'il n'y paraît (Elbaum 2018), qui est susceptible d'offrir de réels avantages, mais qui présente des limites, et qui nécessite de remettre en cause des idées préconçues et une partie des méthodes de travail. On trouve ainsi chez Salgado et Oancea (2020) une comparaison poussée des différents types de sources (données d'enquête, administratives, numériques) sur le plan méthodologique, mais aussi en termes de qualité, de technologie et d'accessibilité.

Le présent document de travail propose une ébauche de réflexion en ce sens, avec d'abord trois parties posant les bases du sujet, en caractérisant données d'enquête et données administratives, puis trois autres portant sur la démarche proposée, et enfin une synthèse.

Le chapitre 1 traitera des statistiques « classiques », issues d'enquêtes, en cherchant à mettre en évidence les propriétés qu'elles embarquent, explicites ou non.

Avoir ainsi un cadre d'analyse permettra, dans un 2e chapitre, d'aborder les « données administratives» et la base de leurs principales propriétés : on pourra ainsi les mettre en regard des données d'enquête, pour mieux analyser leurs avantages et leurs limites d'utilisation. Mais ces données ne sont jamais isolées : elles font partie de ce qu'on nomme « source administrative », sans bien savoir de quoi il s'agit.

Le chapitre 3 montrera qu'une telle source découle en réalité d'un « système d'information administratif », incorporant toute une vision métier, objet méconnu, insaisissable, et pourtant décisif. Or passer de l'univers administratif à un autre univers (statistique, ou recherche, par exemple), ne se fait pas sans frictions et sans transformations, pouvant même remettre en question la pertinence de l'adjectif « brut ». On déduira de tout cela une représentation du processus de production statistique issu de données administratives en trois grandes phases : acquisition, transformation, traitement statistique.

Le 4e chapitre sera consacré à la phase d'acquisition, c'est-à-dire de récupération de sources de données administratives conformes, bien structurées, documentées. On insistera particulièrement sur la notion de qualité, la démarche de qualification pour mieux l'appréhender et la nécessaire boucle de rétroaction pour l'améliorer. L'acquisition est probablement la phase la plus méconnue et la moins formalisée dans la littérature.

Le chapitre 5 portera sur la phase de transformation, qui permet de passer de l'univers administratif (avec son langage, ses objectifs opérationnels, sa sémantique) à l'univers statistique. Il s'agit ici d'une succession d'étapes que l'on va chercher à automatiser, ce qui n'enlève pas l'intérêt de vérifications, d'une autre nature. A l'issue de cette phase, on aura donc un jeu de données qui entre parfaitement dans la logique d'un usage statistique.

La phase de traitement statistique sera décrite de manière beaucoup plus succincte dans un 6e chapitre : en effet, ce sont les traitements statistiques habituels, comme pour une enquête, et l'on devra simplement déterminer où le caractère administratif des données peut avoir quelques impacts.
Enfin, dans une dernière partie, après avoir synthétisé la démarche, on s'interrogera sur ce qu'on peut tirer de tout cela sur un plan pratique, tout en ouvrant quelques pistes prospectives.

A. Caractérisation des données d’enquête et des données administratives

1. Production de statistiques à partir d'enquêtes : propriétés implicites et explicites

Pour Lessler et Kalsbeek (1992), une enquête est une « étude scientifique sur une population existante d'unités représentées par des personnes, institutions ou objets physiques ».

L'Insee, et plus généralement le Service statistique public (SSP), a une longue expérience des enquêtes, et de la fabrication de statistiques fondées sur des données d'enquêtes. Au fil des années, des méthodes ont été conçues, documentées, enrichies, des outils ont été développés sur cette base, des organisations ont été mises en œuvre, notamment pour la réalisation de la collecte, qu'il s'agisse d'enquêtes auprès des ménages ou des entreprises. Des connaissances pratiques, des habitudes de travail, des tours de main se sont construits.

Pour les comprendre, les mettre en évidence simplement, on va ici volontairement s'appuyer sur une représentation très simplifiée, facilitant l'explication et la comparaison entre statistiques d'enquête et statistiques issues de données administratives.

1.a. Représentation mathématique classique… et grille d'analyse

Le point de départ est simple : on est confronté à une situation dans laquelle on veut estimer le total d'une variable sur une certaine population. Par exemple : le chiffre d'affaires total de chaque secteur d'activité, le nombre de chômeurs par région. On a souvent à estimer une moyenne, ce qui est la même chose en divisant par le nombre d'individus (par exemple : âge moyen dans une commune), ce qui revient à un ratio d'agrégats.

Ramenons-nous pour simplifier à un total. On veut donc estimer le total d'une variable X sur une population U pour une période / date de référence t :

$$X^t = \sum_{i \in U} X_i^t \qquad (1)$$

Lorsqu'on réalise une enquête, on connaît souvent la population U, pour laquelle on dispose d'une base de sondage. Dans cette base, on tire un échantillon S aléatoirement, avec des probabilités d'inclusion connues, la théorie des sondages nous permet de construire un estimateur permettant d'approcher ce total :

$$\hat{X}^t = \sum_{i \in S \subset U} w_i X_i^t \qquad (2)$$

où, dans l'estimateur d'Horvitz-Thompson (1952), les pondérations w_i sont les inverses des probabilités d'inclusion.

Dans les faits l'expression (1) est trop générale. En premier lieu, les statistiques calculées le sont pour une certaine « case » de tableau, par exemple toutes les personnes ayant entre 25 et 34 ans, ou bien toutes les entreprises relevant du commerce de détail : on s'intéresse donc aux entités relevant d'une certaine sous-population (ou catégorie) C_k. Par ailleurs, on cherche le total de X à une certaine date, par exemple le chiffre d'affaires total dans le commerce de détail en 2023, X^{2023}.

Le total qu'on veut approcher s'exprime donc plutôt sous la forme :

$$X^t_{C_k} = \sum_{\substack{i \in U \\ i \in C_k}} X^t_i \quad (3)$$

où t est la période, par exemple une année.

Pour l'estimer, on va tenir compte de l'échantillonnage (et introduire, donc, les poids de sondage), mais aussi du fait que certaines valeurs X^t_i sont manquantes ou aberrantes, ce qui conduit à les imputer[2], sur la base d'un modèle d'imputation. On note $\widetilde{X}^t_i$ la valeur imputée.

L'estimateur s'écrira :

$$\hat{X}^t_{C_k} = \sum_{\substack{i \in S \subset U \\ i \in C_k}} w_i \widetilde{X}^t_i \qquad (4)$$

Dans cette formule très simple, $\widetilde{X}^t_i = X^t_i$ si la valeur est conservée[3], et s'appuie sinon sur un modèle. On pourrait évidemment complexifier, en ajoutant le calage sur marges, par exemple.

La représentation probabiliste associée aux enquêtes rend techniquement possible un calcul de variance de l'estimateur, et plus généralement un calcul de *total survey error*[4] (Groves, Lyberg 2010 ; Lyberg, Weisberg 2016 ; Biemer et al. 2017) : ainsi, dans la formule précédente, l'échantillon *S* et la valeur imputée $\widetilde{X}^t_i$ sont des variables aléatoires[5]. On peut également chercher à estimer une évolution, et là aussi on peut calculer une précision, qui fait intervenir les probabilités jointes d'inclusion : en d'autres termes, d'une manière ou d'une autre, on sait peu ou prou quantifier l'incertitude, la marge d'erreur[6].

Pour matérialiser tout cela, et obtenir en pratique la population, les observations, l'appartenance aux catégories, … on va mettre en œuvre un processus d'observation, pour approcher au mieux les différents aspects évoqués qui figurent dans la formule initiale de l'agrégat à estimer. La formule (4) montre qu'on a besoin de cinq composantes (Figure 1):

1. L'unité (i)
2. La population, le champ (U)
3. La variable (X)
4. La catégorisation (C)
5. La temporalité (t)

***Figure 1* – Grille d'analyse des données d'enquête**

2 Bien entendu on pourrait non pas imputer, mais repondérer.

3 Elle n'est pas conservée lorsqu'on effectue une imputation, par exemple.

4 L'idée de la *total survey error* est de construire une mesure d'erreur qui aille au-delà du calcul de la variance d'échantillonnage, mais tienne compte d'autres composantes de l'erreur (ex. erreur d'imputation). Cela mériterait un développement en soi, qui sort du cadre du présent document. On se borne ici à citer des références.

5 On présente ici une formule simplifiée, avec des composantes majeures de l'erreur, en premier lieu l'erreur d'échantillonnage, mais la littérature académique sur le sujet en distingue bien d'autres.

6 Hand (2018) : « statistics is the technology of extracting meaning from data and of handling uncertainty. »

1.b. Appareil d'observation en deux temps

Le cadre méthodologique de la statistique d'enquête ne se limite pas à un attirail de formules mathématiques. Il s'agit, plus généralement, de réaliser un *appareil d'observation*, comme il en existe dans d'autres domaines : en archéologie, en biologie, en astronomie … (Borgman 2015). Chacun de ses dispositifs présente ses propres spécificités, mais aussi ses propres fragilités, comme le soulignent (Borgman et *al.* 2016) pour l'astronomie, par exemple.

Qu'en est-il de l'observation de la réalité socio-économique à partir d'une démarche d'enquête ? La particularité de l'appareil d'observation statistique, c'est qu'il vise à observer des phénomènes « macro », sur toute une population (emploi, pauvreté, délinquance, santé, …). Pour cette raison, il fonctionne en deux temps : d'abord le recueil et la mise au point d'un « grand » nombre d'observations individuelles, pour en déduire dans un deuxième temps des usages statistiques, avec en particulier des agrégats, qui sont en quelque sorte des observations macroscopiques.

Pour obtenir une observation individuelle, on utilise un *instrument* d'observation qui s'appelle le *questionnaire* (en biologie, ce serait le microscope, par exemple). Et ce qu'on va appeler ici le *système* d'observation, c'est ce qui produit l'ensemble de ces observations individuelles, pour les unités de *l'échantillon*.

Ce système nécessite, pour chaque enquête, un long travail préalable de conception, puis de construction (comme si on devait reconstruire un télescope à chaque fois), et enfin de mise en œuvre opérationnelle, que l'on appelle la *collecte*, cette mise en œuvre s'appliquant à la fois à la population[7] et aux variables à observer.

On a ainsi :

- La conception : questionnaire (on élabore un moyen de trouver les X_i *via* un questionnement), élaboration de la base de sondage, échantillonnage (les unités i), choix du mode de collecte …
- La construction : réalisation du (des) support(s) de collecte, mise au point du protocole[8]
- La mise en œuvre : la collecte elle-même.

Le 1er niveau d'observation, c'est l'ensemble des données individuelles obtenues : c'est le résultat de la collecte, qui se matérialise par un ou plusieurs fichiers. C'est en quelque sorte un produit intermédiaire, à certains égards, caractérisé par une représentation tabulaire : un « rectangle » de *p* variables pour *n* individus. Pour aboutir à cela, toute une série d'itérations sont nécessaires. Bien sûr, c'est un peu plus compliqué que cela en pratique, mais on peut se limiter à cette vision simplifiée pour notre propos sans perdre en généralité.

Le 2e niveau d'observation a comme point de départ ce fichier, ce rectangle, et a comme résultat un ensemble de statistiques[9] (agrégats, moyennes, ratios, quantiles, caractéristiques de distribution …). Il s'agit en premier lieu de le compléter, de le finaliser (ou presque) : l'imputation des données

7 En toute rigueur, il faut distinguer l'unité de collecte (auprès de laquelle on collecte) et l'unité statistique (au sujet de laquelle on collecte) : pour les enquêtes auprès des entreprises, l'unité de collecte peut être un cabinet comptable.

8 Y compris des protocoles complexes, comme c'est le cas avec le multimode (Beck et al 2022).

9 Bien entendu il y a d'autres usages, notamment économétriques, mais dans une démarche simplificatrice, on se centre ici sur les statistiques descriptives.

manquantes ou aberrantes, mais aussi la catégorisation (C_k)[10], dont on a vu l'importance dans la partie précédente ; par exemple, le codage d'une cause de décès (Coudin, Robert 2024), à partir de plusieurs libellés et d'autres informations, tout cela exigeant un important travail d'interprétation. Sur cette base-là on construit ensuite les statistiques elles-mêmes, avec des méthodes statistiques d'estimation, de traitement de la non-réponse totale, de calage sur marges,…

Le fait que le processus de traitement repose sur des modèles statistiques permet, au moins en principe, d'estimer une variance, avec ses différentes composantes. En d'autres termes, ayant tous les leviers du processus d'élaboration des statistiques, ayant un cadre mathématique, on est capable de mesurer l'erreur, dans ses différentes dimensions : la statistique d'enquête permet, et c'est là sa grande force, la maîtrise de l'erreur (Deville 1997). À noter que dans le cas des statistiques en évolution, toute cette mécanique nécessite une certaine constance dans le temps de la grille d'analyse : les variables X, les unités i, les catégories C_k, la population U, etc.

1.c. Besoin de vérifications à différents niveaux

Un des aspects centraux du processus d'enquête, c'est le vaste travail de vérification, automatique et manuel, qu'il requiert, et ce à différentes étapes. Pourquoi ? Les statistiques publiques peuvent avoir de nombreux utilisateurs, qui très naturellement vont être tentés de les « mettre en regard » avec d'autres statistiques (en comparant avec des périodes antérieures, ou bien avec d'autres statistiques équivalentes). Assurer cette comparabilité dans l'espace et le temps[11], et donc donner des gages de cohérence, fait partie des exigences de la statistique publique, de même que la capacité à expliquer et à justifier les statistiques, dans un souci de transparence.

Il existe donc naturellement de nombreuses vérifications, de natures et surtout de temporalités diverses, associées ou non à des imputations automatiques : c'est ce qu'on regroupe sous le terme de *data editing*[12] (De Waal, Pannekoek, Scholtus 2011). On pense d'abord à la vérification à la source, qui s'effectue lors de la collecte par enquêteur : la médiation par un enquêteur permet par exemple de vérifier la bonne compréhension de la question, de voir si la personne enquêtée est bien sûre de sa réponse, ou d'effectuer la bonne interprétation quand on catégorise. Le fait de collecter par voie électronique permet aussi d'effectuer des contrôles élémentaires automatiquement (contrôles de type, de forme, voire de cohérence inter-variables), intégrés au support de collecte, et qui permettent d'éviter une partie des erreurs[13].

Dans les enquêtes auprès des entreprises, la vérification par interaction avec l'enquêté peut s'effectuer ultérieurement, mais toujours dans une temporalité courte, en rappelant l'entreprise enquêtée pour vérifier des réponses qui ont pu paraître anormales. Ce travail des gestionnaires d'enquête permet d'effectuer cette mise au point, qui est essentielle.

10 Qui peut être présente dans la base de sondage (par exemple l'APE pour les enquêtes entreprises), ou obtenue via le 1er niveau d'observation.

11 La comparabilité est l'une des dimensions du cadre qualité de la statistique européenne.

12 Le *data editing* porte avant tout sur le travail de vérification de données, automatique ou manuel, et ce qui découle de cette vérification, à savoir la validation des données ou au contraire leur invalidation. Cette dernière situation peut conduire à des modifications manuelles, par des gestionnaires, ou à des traitements automatiques (imputation fondée sur des modèles). Ainsi, le *data editing* (qu'on peut traduire en français par contrôle redressement, peu ou prou) n'inclut les imputations que dans la mesure où elles dérivent d'une vérification infructueuse.

13 Ces contrôles de cohérence sont rarement bloquants, mais l'outil de collecte électronique contraint le format de réponse.

Les contrôles peuvent aussi avoir lieu bien plus tard, en phase de traitement statistique, en analysant de façon sélective les unités pour lesquelles l'impact sur les statistiques est fort, ce qui conduit à de nouvelles vérifications individuelles (Lawrence, McKenzie 2000).

Dans ces processus de *data editing*, les vérifications commencent en général par des contrôles automatiques, qui permettent de trouver des anomalies (une réponse en euros au lieu de k€ par exemple). Elles peuvent se poursuivre par des vérifications manuelles lorsqu'il y a anomalie dont l'ampleur est jugée statistiquement significative et pour laquelle une correction automatique serait une source d'erreur potentielle trop importante, et conduisent à rétroagir sur la valeur … ou pas (car ce qui apparaît comme une anomalie n'est pas nécessairement une erreur). La démarche de vérification induit ainsi naturellement une « boucle de rétroaction » pouvant conduire à modifier des données. Ce principe de rétroaction, de *feedback loop*, présente un caractère très général et vaut pour le contrôle de tout processus (Wiener 1948). Cette idée centrale de la cybernétique, science des systèmes commandés, s'applique naturellement au processus de production statistique.

Ainsi, qu'elle ait lieu au moment de la collecte ou après[14], la rétroaction s'effectue *au niveau de l'unité statistique* : les modifications porteront *sur des micro-données*. On verra que c'est tout à fait différent dans le cas des données administratives.

En toute fin d'enquête, on procède à des contrôles qui portent cette fois sur les agrégats eux-mêmes (*output editing*) : le fait qu'il y ait une évolution forte dans le temps de telle ou telle statistique n'est pas nécessairement une erreur, mais nécessite que cette évolution soit détectée, bien comprise et documentée, voire subdivisée en composantes (évolution des unités pérennes, évolution liée à la démographie des unités du champ, etc.). L'output editing est parfois réalisé informellement par les chargés d'études eux-mêmes lorsqu'ils sont les premiers utilisateurs de résultats et détectent des incohérences. Même dans ce cas-là, la boucle de rétroaction induira, s'il y a lieu, une modification de micro-données[15].

Ces vérifications multiples, qu'il faut savoir doser (Granquist 1997), jouent un rôle indispensable dans le processus d'enquête, avec pour objectif d'améliorer la qualité des données individuelles et, partant, des agrégats. Pour les corrections automatiques (imputation notamment) il est également possible de mesurer leur impact sur la précision des agrégats fournis.

1.d. Caractérisation des statistiques issues d'enquête

La représentation mathématique de la partie 1a a mis en évidence différentes dimensions (unités statistiques, variables, champ, catégories, temporalités) que l'on peut reprendre ici. Au fond, ce qui caractérise le processus de production statistique d'enquête, c'est l'existence d'un certain degré de maîtrise sur les 5 éléments (Nordbotten 2010)[16] de notre grille d'analyse :

- les unités statistiques à observer, définies très tôt, et qui sont souvent (mais non exclusivement) des individus, des ménages, des entreprises, ou des établissements[17],

14 Ce qui peut être utilement modélisé dans le GSBPM (cf. § 1.e).

15 Il ne faut pas imaginer pour autant que le cadre des enquêtes soit idyllique, il existe aussi des situations dans lesquelles la modification de micro-données n'est plus possible : c'est le cas par exemple lorsqu'on mobilise des données d'enquête dans des séries longues. Certains travaux (rétropolation de changement de concepts, de nomenclatures) s'effectuent de façon agrégée.

16 En tant que piliers de base, trois de ses composantes à savoir unité, population et temps, sont un triptyque classique depuis longtemps : Nordbotten en parle dans son article de 2010, qui renvoie à un papier de sa part des années 60.

- le champ d'observation : défini de telle sorte qu'il ait du sens, et, de préférence, afin qu'il soit accessible avant la collecte

  Par exemple, pour une enquête auprès des entreprises, on peut définir le champ en fonction du secteur d'activité, de la tranche d'effectif, mais pas en fonction du montant d'investissements réalisé, ceci en fonction des informations disponibles dans la base de sondage[18].

- les variables à observer : définies après concertation avec les différentes parties prenantes (utilisateurs mais aussi fournisseurs des données[19]), dans le cadre d'instances dédiées à cela (Anxionnaz, Maurel 2020),

- les principes de catégorisation : des nomenclatures partagées (Guibert, Laganier, Volle 1971),

  Exemple : la nomenclature des professions (Amossé 2020), ou la nomenclature des infractions (Camus 2022),

- enfin, une double temporalité fixée et qui s'applique uniformément à toutes les unités : temporalité d'acquisition (période durant laquelle on effectue la collecte), temporalité de référence des données collectées (date ou période considérée[20]). Cette dernière caractéristique est tout à fait essentielle et représente une grande différence avec la situation des données administratives.

En termes de processus, le cadre classique de l'enquête permet :

- de maîtriser le protocole d'observation : il peut être complexe pour certains types d'enquête, par exemple les enquêtes multi-mode ou les enquêtes à carnet, ou encore l'enquête sans domicile pour laquelle on ne dispose pas de base de sondage,

- de vérifier le cas échéant les données à la source (interaction enquêteur – enquêté, gestionnaire[21] – enquêté),

- d'estimer les erreurs grâce à la connaissance de la base de sondage, et du processus aléatoire d'échantillonnage, mais aussi en prenant en compte les modèles d'imputation (ou de repondération) pour les non-réponses partielles ou totales.

17 Dans les dispositifs comme les enquêtes usagers de la Drees, on s'intéresse à des séjours hospitaliers avec une triple approche (établissement, médecin, patient)

18 La base de sondage est une contrainte qui peut éventuellement limiter le champ d'observation.

19 On pense ici par exemple aux entreprises, représentées par des fédérations, ou organisations patronales.

20 Par exemple, les enquêtes structurelles auprès des entreprises collectées l'année N portent souvent sur les valeurs de l'année N-1, et en pratique sur le dernier exercice comptable. Le chiffre d'affaires annuel n'est connu qu'en fin d'année et ne peut donc être collecté que l'année suivante.

21 L'activité de gestionnaire d'enquête *(survey clerk* en anglais) comporte de nombreux aspects, comme la vérification manuelle des données (micro ou macro), le suivi, la relance des non-répondants, … Les gestionnaires d'enquête jouent un rôle essentiel pour assurer la qualité des statistiques publiques.

1.e. Représentation du processus : le GSBPM

La statistique officielle fait l'objet d'une représentation normalisée, offrant un cadre partagé au niveau international pour le processus de production des statistiques officielles : le Generic Statistical Business Process Model, ou GSBPM (Unece 2019).

Le GSBPM explicite les étapes du processus, avec un découpage en 8 phases, qui constitue un modèle de référence. Il est conçu pour que chaque institut statistique l'applique à sa manière, en fonction de son organisation et de ses contraintes, selon ce qui leur semble approprié (Erikson 2020).

Les 8 étapes sont les suivantes[22] :

1. **Définition des besoins**
2. **Conception**
   dont conception du descriptif de variables, de l'échantillon, de la collecte, des produits, du système de production
3. **Élaboration**
   dont élaboration de l'instrument de collecte, des composantes du processus, mise à l'essai du processus
4. **Collecte**
   création d'un cadre et sélection de l'échantillon, organisation, réalisation et finalisation de la collecte
5. **Traitement**
   dont intégration, codage, data editing, imputation, calcul des variables dérivées, finalisation du fichier de données
6. Analyse
7. Diffusion
8. Évaluation

Si l'on se limite aux 4 premières étapes, antérieures au traitement, la fonction du processus de production statistique consiste, dans le cas des enquêtes, à traiter des difficultés de multiples natures, telles que : converger sur un besoin (et donc des variables et des classifications), prendre connaissance des conventions ou des modèles théoriques sous-jacents dans le domaine concerné, exprimer des questions cohérentes avec les buts recherchés, choisir la base de sondage la plus adaptée, sélectionner un échantillon apportant une précision suffisante pour les buts recherchés, bâtir un support de collecte efficace, mettre au point un protocole de collecte adapté à différentes sous-populations et configurations, traiter les non-réponses (partielles ou totales), affecter de façon pertinente un code de la nomenclature sur la base d'un libellé,… La maîtrise du processus permet de construire toutes les métadonnées (Sundgren 1993, Bonnans 2019) nécessaires au fur et à mesure des étapes.

Pour les données administratives, les problématiques et les difficultés rencontrées seront de toute autre nature et remettront en question ce découpage, tout du moins dans les premières phases, jusqu'au traitement, qui est la 5e étape du GSBPM.

22 En gras, celles qui concernent ce document de travail.

2. Produire des statistiques à partir de données administratives

L'utilisation de vastes sources de données (y compris privées, d'ailleurs) peut apparaître à première vue comme une sorte d'eldorado pour la statistique publique : une forme d'exhaustivité, permettant d'accéder à des sous-populations plus fines, une variance réduite à zéro, un gain économique très significatif, etc. (Hand 2018) montre que cela ne va pas de soi, et qu'**il serait trop rapide de considérer que l'absence de coûts de collecte justifie à lui seul l'usage de sources administratives,** alors qu'il n'existe même pas de cadre théorique consensuel.

Avant toute chose, il est indispensable de mieux comprendre de quoi on parle lorsqu'on évoque les données administratives, et en premier lieu de déterminer les principales caractéristiques de ces données, ce qui va conduire à s'interroger sur leurs usages. Mais ce ne sera pas suffisant : pour appréhender la question de façon rigoureuse, il nous faudra commencer par spécifier la notion de donnée et notamment ses différentes dimensions. Celles-ci vont nous apporter un cadre conceptuel permettant d'interroger les propriétés de données administratives *utilisées à des fins statistiques*, cadre qui se révélera également utile dans les chapitres suivants.

2.a. Quelques caractéristiques des données administratives

La première dimension qui vient à l'esprit est **institutionnelle** : lorsqu'on évoque ces données, on fait allusion explicitement à leur origine et à la nature de l'unité productrice, le plus souvent une organisation gouvernementale[23]. L'OCDE les définit comme *l'ensemble des unités et données dérivant d'une entité administrative*, cette dernière étant définie comme *une unité organisationnelle responsable de l'implémentation d'une régulation administrative (ou groupe de régulations), pour laquelle l'ensemble des unités et transactions sont vues comme une source de données statistiques* (OECD 2008). Elles présentent ainsi un caractère officiel, s'appuyant sur une assise juridique, et l'existence même de ces données peut donner des droits ou des devoirs, et avoir ainsi des conséquences très concrètes sur la vie quotidienne : paiement d'un impôt, d'une amende, attribution d'une retraite, droit à bénéficier d'une subvention, … Contrairement aux données privées[24], les données administratives sont souvent associées à des obligations (payer ses cotisations sociales par exemple) et l'administration n'a pas d'intérêt commercial ou stratégique associé aux données.

Dans la littérature académique, les auteurs insistent plus particulièrement sur **l'usage** qui en est fait. Hand (2018) les définit simplement comme les « *données générées au cours d'une opération et conservées dans une base de données* ». Une définition un peu plus précise est proposée par Alain Desrosières : « *Une source administrative est issue d'une institution dont la finalité n'est pas de produire une telle information, mais dont les activités de gestion impliquent la tenue, selon des règles générales, de fichiers ou de registres individuels, dont l'agrégation n'est qu'un sous-produit, alors que les informations individuelles en sont l'élément important, notamment pour les individus ou les entreprises concernés.* » (Desrosières 2004).

Dans leur essence, les données administratives se distinguent donc des données d'enquête avant tout par leur **finalité** : elles « naissent » en lien avec une intention de gestion, de mise en œuvre d'un processus administratif, alors que dans une enquête l'intention est d'observer. Pour l'administration concernée, la constitution de bases de données n'est pas un objectif en soi, elle supporte, permet de

23 Pour désigner l'institution en question, on utilisera indifféremment, tout au long du document, les termes « administration », « entité administrative », « entité publique », ou entité chargée d'une mission de service public..

24 On pense ici aux données provenant d'entreprises privées, comme les données de caisse (Leclair 2019).

piloter un processus opérationnel. Les données embarquent alors leur propre univers, leurs propres concepts, liés au processus supporté et à ses exigences : dans l'exemple de données hospitalières, les identifiants de personnels médicaux et soignants, les parcours patient, les nomenclatures de groupes homogènes de malades en vue de remboursement par l'assurance maladie, ...

Il en découle que la donnée administrative n'est pas nécessairement une observation : elle peut par exemple refléter une décision de gestion, comme l'attribution d'une prestation sociale.

> Exemple : pour l'assurance maladie (AM), une simple consultation pour vérifier la vue chez un ophtalmologue peut se traduire par deux résultats différents dans le système d'information : une consultation en ophtalmologie, ou bien plusieurs actes médicaux (ceux que le médecin a fait pendant la consultation). Les montants associés à ces deux codifications étant différents, chaque médecin peut être amené à optimiser son codage pour avoir la meilleure rémunération. Ainsi, le médecin en secteur 2 pourra avoir tendance à enregistrer une consultation ; celui en secteur 1 les différents actes.

Ces logiques d'optimisation peuvent de plus varier dans le temps, et au sein de l'organisation. Cela s'avère d'autant plus vrai lorsque la gestion du processus administratif est déconcentrée.

Plus généralement les données administratives peuvent être vues comme des « déclencheurs d'action » et non comme des informations passives. Elles sont construites pour un usage précis (par exemple : rémunération d'un médecin, attribution d'une prestation sociale, prélèvement d'impôt à bon droit...), mais pas en première instance pour l'observation : pour Hand, elles sont finalement des **co-produits de processus opérationnels**, voire des « résidus » de ceux-ci. Il parle ainsi de « *data exhaust* [25]» ou « *ce qui reste après que la machine administrative ait utilisé les données pour son activité propre* ».

L'usage statistique de ces données ne constitue pas en général une finalité de ces données[26], contrairement aux données d'enquête qui sont justement conçues à cette fin, mais une utilisation secondaire, une réutilisation. Il n'en demeure pas moins que certaines d'entre elles (les déclarations annuelles de données sociales ou les données fiscales notamment) ont été utilisées pour produire des statistiques durant de nombreuses années (les données fiscales sont ainsi utilisées par l'Insee depuis 1950), cet usage étant facilité, en France, par un accès garanti par la loi, pour le service statistique public[27].

Les données administratives présentent également un gros défaut : **l'absence de garantie de leur pérennité**. Dépendantes des processus qu'elles soutiennent, les données administratives peuvent évoluer fortement en matière de champ d'intérêt, de concepts, et peuvent même tout simplement ne plus exister. Les modifications récentes sur la fiscalité (suppression de la taxe d'habitation – TH[28] - sur les résidences principales) ont entraîné la disparition des données liées à ce processus, qui jouaient un rôle majeur pour la statistique publique : l'Insee utilisait ces informations pour définir les ménages dans la base de sondage des enquêtes. La suppression de la TH (Bach et al 2023) a donc été une grosse perte qui a, de fait, été imposée aux statisticiens publics sans qu'ils aient la possibilité d'agir, en perdant du jour au lendemain le lien entre les individus et leur logement.

25 Ce terme fait référence à l'épuisement d'une ressource, ou également aux gaz d'échappement : ce qui reste de la matière première « donnée » une fois que l'on a tiré tout ce qui était possible pour son usage premier.

26 Il en résulte que toutes les données ne sont pas utiles à la statistique, très loin de là, et qu'il existe donc de nombreuses données dites « secondaires ».

27 Article 7 bis, Loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques.

28 Les différentes sources évoquées dans ce document sont décrites succinctement dans l'annexe 2.

Sans suppression totale d'une source, il peut arriver qu'une variable soit supprimée ou modifiée alors qu'elle est centrale dans un processus.

> Exemple : l'exonération fiscale des heures supplémentaires peut impliquer qu'elles ne sont plus déclarées dans le salaire net fiscal[29]… or c'est une variable de base pour produire les évolutions de salaire.

De façon générale, le fait que ces données ne soient évidemment pas sous la responsabilité du système statistique conduit à un **déficit de maîtrise** : on ne peut les placer sur le même plan que les données d'enquête. Dans ce qui suit, on va préciser en quoi, en commençant par formaliser les différents aspects de la notion de donnée, dans l'optique d'un usage statistique.

2.b. Notion de donnée … et nouvelle grille d'analyse

Il existe de nombreuses définitions de la notion de donnée dans la littérature, par exemple (Kitchin 2014), (Borgman 2015), ou (Caron et al 2020). Mais on va plutôt utiliser ici la définition, plus opérationnelle, de Thomas Redman (1997), qui se base lui-même sur une définition traditionnelle en informatique.

On peut définir une donnée comme étant un triplet (concept, domaine, valeur)[30].

Le **concept** représente la signification de la donnée, que l'on va décrire par un libellé : par exemple, « la hauteur de la tour Eiffel », « le genre d'une personne », « la date de début d'un contrat », « l'activité principale d'une entreprise ». Il est décrit comme l'attribut d'un **objet**, ou, pour reprendre un terme plus traditionnel, par une **variable** associée à un objet. Dans ces exemples, les objets sont la Tour Eiffel, une personne (une personne bien précise), un contrat (idem), une entreprise. Et les attributs, ou variables, sont ici : hauteur, genre, date de début, activité principale.

On appelle **domaine** l'ensemble de valeurs possibles associées au concept, avec explicitation du sens. Par exemple, pour la hauteur, ce sera un nombre entier positif, exprimé en mètres, mettons. Pour le genre, ce sera par exemple l'ensemble {0 , 1}, lui-même complété d'explications (par exemple 0 = homme, 1 = femme), ou {0, 1, 2} avec un genre « autre ». Pour la date, il faudra déterminer dans quel calendrier on se situe, et intégrer dans la définition du domaine les diverses contraintes (ex : le jour est inférieur ou égal à 31, …). Soulignons ici que le *format* n'est qu'une implémentation de la représentation d'un domaine de valeurs. Pour une date par exemple, il existe plusieurs formats possibles (JJMMAA, JJ/MM/AAAA, …), et le choix de format ne change rien au domaine.

Et la **valeur** … c'est la valeur, et c'est ce qu'on a tendance à appeler « donnée » habituellement.

Concept et domaine font ainsi, et c'est le point essentiel ici, partie intégrante de la notion de donnée : si on lit « 324 », par exemple, ce n'est pas une donnée, c'est juste un nombre. On peut lui associer le concept « hauteur de la tour Eiffel », « nombre d'habitants de tel village », « poids de telle moto », et dans ce cas on a affaire à des données différentes, mais on a besoin de la caractérisation du domaine, avec par exemple l'unité de mesure de la hauteur (pieds ? mètres ?).

29 En pratique, les consignes de déclaration évoluent dans le temps : ainsi les heures supplémentaires étaient incluses avant 2020, puis ne l'étaient plus entre 2021 et 2023, et le sont à nouveau en 2024.

30 Voir aussi à ce sujet (Rivière 2020).

Si l'on a en tête un usage statistique, et les caractérisations proposées dans la partie sur les données d'enquête, deux autres dimensions de la notion de donnée vont jouer un rôle important.

En premier lieu, le **champ**, c'est-à-dire une restriction de l'ensemble des possibles pour l'objet considéré : on aura, par exemple, une mesure administrative qui ne s'appliquera qu'aux entreprises de moins de 20 salariés créées après telle date, ou aux transactions de tel type effectuées telle année, ou aux retraites versées à telle période, inférieures à tel seuil, pour des personnes ayant telles caractéristiques (âge, durée de cotisation, …).

En second lieu, les **temporalités** : une donnée se caractérise souvent par une *date (ou période) de référence* (par exemple, tel arrêt maladie a eu lieu tel jour), qui est en soi une métadonnée (Gartner 2016)[31] fort utile. Mais cette date de référence est différente de la *date (ou période) d'acquisition*, i.e. la date à laquelle l'information a été obtenue : par exemple, une enquête auprès des entreprises menée en 2024 au sujet de l'exercice 2023.

Pour mieux appréhender les données administratives, on va, au total, s'appuyer sur une grille semblable à celle des données d'enquête, et comportant comme elle 5 dimensions (Froeschl et Grossmann 2000)[32] :

1. L'objet
2. Le champ
3. La variable
4. Le domaine de valeurs
5. Les temporalités

***Figure 2* - Grille d'analyse des données administratives**

Cette grille présente de légères différences celle relative aux enquêtes : on ne parle plus d'unité d'observation mais d'objet[33] ; la variable est ici un attribut de l'objet ; le domaine de valeurs ne se réfère pas toujours à une nomenclature ; la notion de champ est similaire à ce qui se passe pour une enquête, mais elle est ici spécifique aux besoins de l'administration ; enfin, les temporalités sont, dans la majorité des cas, très différentes de celles d'une enquête.

2.c. Données administratives en vue d'un usage statistique

Pour chacune des cinq dimensions proposées (objet, variable, domaine, champ, temporalité), on va indiquer les spécificités d'une donnée administrative et comparer avec les données d'enquête. On trouvera aussi dans (De Broe et *al* 2020) une comparaison de type avantages – inconvénients entre ces deux approches.

31 On peut définir une métadonnée comme une « donnée fournissant de l'information sur une donnée, ou des données ».

32 Les auteurs effectuent une distinction similaire, en étant à mi-chemin entre mondes statistique et administratif : *populations and populations units*, *variables and value sets*, *coverage*, *data production dynamics* .

33 Objet spécifique à l'administration, mais qui ne sera pas nécessairement associé à une démarche d'observation.

- ***Objet***

L'objectif d'une administration est de mener à bien différents processus opérationnels, permettant de rendre des services à des usagers, de collecter des impôts, d'octroyer des droits, de verser des prestations, ce qui conduit à enregistrer les évènements correspondants. Les objets associés, parfois simples, peuvent se révéler complexes, en raison de nombreuses liaisons entre ces objets : ainsi dans un hôpital, le parcours patient fait intervenir plusieurs « objets » en plus du patient lui-même, ayant chacun sa propre dynamique et ses propres attributs : médecin, prescription, diagnostic, établissement, ... Ce sont des objets « de gestion », différents des objets « d'étude » statistique.

Les objets doivent être créés dans le système d'information (nouveau patient, nouveau médecin, nouvel établissement). Cela peut passer par la déclaration d'un évènement simple, comme une naissance, cet événement physique conduisant à créer un nouvel enregistrement dans le répertoire des individus. Dans le cas d'une entreprise, c'est à l'inverse la création de l'objet entreprise dans le système d'information (dans le répertoire) qui va officialiser la naissance de l'entreprise.

Souvent, les données administratives vont porter sur des individus[34], ce qui peut donner une illusion de simplicité : car en tant que support d'un processus administratif, les données portent plutôt sur un **rôle** d'un individu.

> Exemples : une personne physique pourra être considérée, selon les domaines administratifs, les usages et les personnes, comme un usager, un électeur, un contribuable, un bénéficiaire de prestation, un étudiant, un professionnel de santé, un garant… avec dans chaque cas, un angle d'analyse différent et des variables différentes.

Ainsi, le lien n'est pas nécessairement bijectif entre objet administratif et unité statistique.

> Un individu, par exemple, peut correspondre à plusieurs contribuables, dans plusieurs départements, en lien avec le niveau de gestion administratif. Et inversement, un individu peut en représenter plusieurs avec les notions d'ouvrant-droit / ayant-droit par exemple.

La multiplicité des objets considérés dans un processus administratif entraîne aussi des possibilités d'exploitation statistique variées.

> On pense aux exploitations diverses des données cadastrales, qui selon les sources peuvent être relatives à différents objets : propriétaire, logement, transaction, … (André, Meslin 2022). Mobilisées dans Résil (Lefebvre 2024) ou Fidéli (Lamarche, Lollivier 2021) à des fins de localisation des logements, elles peuvent également servir dans l'évaluation du patrimoine immobilier (André, Arnold, Meslin 2021).

À l'inverse, dans le cas des enquêtes, dans une majorité de cas le contexte est simple[35] : l'« objet » est l'individu / le ménage / l'entreprise interrogé (e), qui est « créé » dans le système d'information à travers la base de sondage et le tirage d'échantillon. La question des rôles, des liens entre objets, est moins prégnante, même si elle existe (parent – enfant, entreprise – établissement).

34 Bien entendu on a aussi des données administratives sur des entreprises, établissements, administrations, ...
35 Mais pas toujours : on peut avoir des objets plus complexes comme les entreprises profilées (Moreau 2024).

- ***Champ***

Les processus administratifs concernent souvent une partie seulement de la population : par exemple, l'ensemble des personnes *assujetties à tel impôt*, ou bien l'ensemble des entreprises *éligibles à telle subvention*.

On définit le champ « en intention », par ses caractéristiques : l'ensemble des objets vérifiant telle propriété ; par exemple, l'ensemble des établissements ayant moins de 10 salariés. Ce champ, exprimable de façon simple, permet de déduire une liste d'objets « en extension », ce qui présuppose de connaître la population totale : par exemple, le répertoire Sirene des entreprises, le répertoire national d'identification des personnes physiques (RNIPP) pour les individus (Espinasse, Roux 2022). Dans un contexte administratif, la définition du champ peut évoluer dans le temps, tout simplement parce que la législation évolue (telle mesure s'appliquait aux moins de 10 salariés, après changements elle ne s'applique plus qu'aux moins de 5).

> Exemple : la Déclaration sociale nominative (DSN) porte sur tous les employeurs, publics ou privés, qui doivent déclarer des informations de paie pour l'ensemble de leurs salariés (Humbert-Bottin 2018). Son champ a évolué ces dernières années, puisqu'il ne concernait que le secteur privé jusqu'en 2021.

Mais, on l'a vu, les processus peuvent porter sur plusieurs objets, et le champ peut aussi être défini sur ces différents objets. Lié à des objectifs de gestion, le champ du processus administratif peut donc différer du champ d'intérêt statistique, même lorsqu'il concerne des objets de même nature.

> Exemples : le champ de la source GMBI[36] (Gérer mes biens immobiliers) porte sur tous les locaux, mais ses usages statistiques envisagés ne portent que sur le champ des locaux d'habitation (logements[37]).

Last but not least, le champ supposé n'est pas nécessairement intégralement couvert par le processus. Différentes raisons peuvent conduire à des défauts de couverture : des exonérations particulières, une sous-déclaration d'individus concernés etc.

- ***Variable***

L'écart entre les variables « administratives » et « statistiques » est probablement le sujet le plus classique, le plus connu : on sait que les variables administratives n'ont aucune raison de coller aux besoins statistiques, car rien n'assure que les objectifs soient les mêmes.

> Exemple 1 : le revenu, à mettre en regard de la variable administrative « revenu imposable ». A titre d'illustration, les revenus d'un stagiaire sont non imposables, ils ne sont donc pas pris en compte dans les revenus d'un ménage pour l'administration fiscale.
>
> Exemple 2 : dans la DSN, la distinction des différentes rémunérations ne couvre pas tous les besoins statistiques. Ainsi, il n'est pas possible de distinguer précisément les primes régulières et irrégulières, ou encore les rémunérations liées aux astreintes.

Une même variable peut avoir des définitions différentes d'une source à l'autre ce qui complique leur comparaison et leur utilisation à des fins statistiques. Pour la construction d'un bateau qui sera

36 Voir (Desforges 2021), p. 1110.

37 L'usage d'un terme identique ne signifie pas que les concepts sont rigoureusement équivalents.

exporté par exemple, les avances peuvent faire partie du chiffre d'affaires à l'exportation des années qui précèdent son envoi à l'étranger pour la DGFIP alors que pour les douanes, c'est le prix total du bateau au moment où il quitte le territoire français qui sera considéré comme le CA à l'exportation.

- ***Domaine***

Le domaine est l'ensemble des valeurs admissibles de la variable, ce qui doit pouvoir être caractérisé de façon formelle, mathématisée : le *definition domain* de (Batini, Scannapieco 2016)[38]. On pense aux intervalles, pour des valeurs quantitatives, aux listes de valeurs, aux sélections par rapport à une vaste liste existante,...

Il faut préciser que c'est un ensemble *commenté*. Par exemple, il ne suffit pas de dire que pour la variable genre, la liste des valeurs possibles est {0, 1}, il faut aussi associer une signification à ces deux valeurs (0=homme et 1=femme, ou l'inverse ?). Dans le cas des données quantitatives, il faut indiquer l'unité de mesure : euros ou milliers d'euros, habitants ou milliers d'habitants, mètre ou kilomètre, ...

Comprendre le domaine, pour des données administratives, c'est aussi comprendre de quelle manière ont été notées les valeurs manquantes (-1 ? 99 ? 9999 ? NA ?), ce qui est souvent peu documenté. C'est également connaître les « valeurs refuge », i e. celles que l'on utilise quand on ne sait pas trop quoi dire (attention, ce n'est pas la même chose que les valeurs manquantes).

Dans GMBI par exemple, la variable « le local est-il meublé » est par défaut à non.

Dès lors on ne peut faire la différence entre un vrai « non » (ou une autre valeur refuge) et une non-réponse. Par ailleurs, les consignes pour ces valeurs refuge ne sont pas forcément les mêmes pour tout le monde[39].

Enfin, il arrive souvent que les « domaines » associés aux données soient des nomenclatures, celles-ci jouant un rôle essentiel dans l'activité administrative : nomenclature de catégories juridiques, d'établissements, d'activités, de grades … (Bowker, Star 2000) le soulignent ainsi : « L'attribution de catégories aux choses, aux personnes ou à leurs actions à des catégories est un élément omniprésent du travail dans l'État moderne et bureaucratique. »

La difficulté va être de bien comprendre la caractérisation du domaine (existence de documentation encore …) et de spécifier les différences avec les éventuelles nomenclatures statistiques proches. Les classifications administratives ont des finalités de gestion : par exemple, un découpage en régions reflétera simplement un découpage organisationnel. Paradoxalement, le fait que l'administration emploie une « nomenclature Insee » ne facilite pas nécessairement les choses : il peut s'agir d'une version ancienne, ou en partie agrégée, en partie transformée … il ne faut pas prendre pour argent comptant la référence annoncée à la statistique publique.

- ***Temporalités***

La double temporalité d'une enquête statistique est en général très claire : on effectue une collecte *pendant* une certaine période (*période d'acquisition*, ici *période de collecte*), connue, affichée,

38 Voir pp. 24-25.

39 Par exemple, dans le système d'information de l'assurance-maladie, pour que les patients sans médecin traitant (parce qu'ils ne parviennent pas à en trouver un du fait de la pénurie) ne soient pas pénalisés dans leur parcours de soins, les caisses primaires (CPAM) entrent un numéro de médecin traitant fictif. Les consignes ne sont pas les mêmes d'une CPAM à l'autre et ne sont pas forcément connues au niveau national.

période pendant laquelle on *prend connaissance* des informations. Elle s'effectue *relativement à* une période, ou date (*période, ou date, de référence*). Le standard de métadonnées DDI[40] (Data Documentation Initiative), cadre international largement utilisé en statistique, effectue clairement cette distinction. Pour les enquêtes, ces deux dates et/ou périodes sont, de préférence, relativement proches, et valent de façon uniforme pour toutes les unités.

Dans le cas des données administratives, la différence réside simplement dans le fait que, dans la majorité des cas, il n'y a pas de collecte (au sens d'une double temporalité uniforme). Il existe un contre-exemple à cela : les *déclarations administratives*, qui se caractérisent par une *période déclarative* et une *période de référence* sur laquelle portent les déclarations. Le processus déclaratif (déclaration d'impôts, déclaration sociale nominative) est donc similaire au processus de collecte en termes de temporalités.

Pour le reste, les données administratives sont dépendantes d'événements externes qui peuvent se dérouler à n'importe quel moment : naissance, décès, changement d'adresse, demande de départ à la retraite, demande de prestation, … Ainsi, la *date de référence* n'est absolument pas maîtrisée, et la *date d'acquisition*, ou son équivalent ici, est souvent complètement dépendante du bon vouloir de l'usager (on peut très bien transmettre sa feuille de soins très en retard …).

A noter que ces événements administratifs peuvent être des événements internes à l'administration, par exemple la décision d'attribution d'une prestation, d'une subvention, la décision de donner une amende, … Là aussi, le fonctionnement de la gestion administrative ne garantit pas toujours une bonne maîtrise des dates, ici les dates de décision (ex : décisions de justice).

Pour un usage statistique, la non-maîtrise des deux temporalités pose problème : il n'y a pas de garantie qu'on dispose de l'exhaustivité des événements, parce que ceux-ci sont connus avec un fort décalage, ou tout simplement parce que ceux-ci ne sont pas déclarés.

- ***Synthèse***

In fine, si l'on cherche à estimer le total évoqué en première partie, l'usage de données administratives nous donnera implicitement une expression très différente du cas d'une enquête, avec une formulation du type :

$$\hat{\widetilde{X}}^{t}_{\widetilde{C}_k} = \sum_{\substack{i \in \widetilde{U} \\ i \in \widetilde{C}_k}} \widetilde{X}_i^{t_i} \quad (5)$$

- où l'univers $\widetilde{U}$, la catégorisation $\widetilde{C}$, la variable $\widetilde{X}$, spécifiques à l'entité administrative, n'ont pas de raison de correspondre à U, C et X
- où, pour chaque unité *i*, les temporalités de référence t_i, supposées approcher *t*, n'ont pas de raison d'être toutes égales
- où, enfin, aucun élément de l'expression ne relève d'un modèle aléatoire prédéfini, ce qui empêche un calcul « classique » de l'erreur d'échantillonnage[41]. Il n'est pas non plus possible

40 Voir https://ddialliance.org/ddi-codebook_v2.5. Cela correspond aux balises <referencePeriod> ou <timePeriodCovered> d'une part, et <collectionDate> ou <dataCollectionPeriod> d'autre part.

41 Plus généralement, on s'intéresse à une erreur totale : dans le cadre d'une enquête, cela comprend l'erreur d'échantillonnage, mais aussi l'erreur de non-réponse (que l'on peut modéliser à partir des informations connues sur la base de sondage)... Mais tout n'est pas parfait non plus dans le contexte d'une enquête : certaines erreurs (comme l'erreur de mesure) sont difficiles à estimer.

de calculer classiquement l'erreur liée à la non réponse, en l'absence de connaissance de la base de sondage. D'autres méthodes permettent toutefois d'évaluer le défaut de couverture (méthodes de capture-recapture).
L'estimation d'une erreur totale (à l'instar de l'erreur d'enquête totale, ou *total survey error*) est complexe, mais peut être envisagée. (Reid et al 2017) proposent un cadre permettant d'identifier différents types d'erreurs. Il est toutefois difficile de modéliser celles-ci. (Berkovsky et al 2025) proposent un t*otal data quality paradigm*, prenant en compte les défis spécifiques aux données administratives, en s'autorisant à sortir du cadre probabiliste.

2.d. Statistiques issues de données administratives *vs* statistiques issues d'enquêtes

Vis-à-vis des données d'enquête, les données administratives offrent plusieurs avantages évidents pour le statisticien (De Broe et al 2020) :

- disparition du coût de collecte … tout du moins pour la statistique publique
- disparition de la charge d'enquête[42]
- possibilité d'analyses à granularité très fine (au niveau géographique, sectoriel, tranches d'âge)
- de meilleures possibilités d'appariement : si un identifiant commun (le NIR ou le Siren par exemple) n'est pas présent, des éléments d'identification (nom/ prénom / date de naissance, ou adresse / raison sociale) le seront souvent et permettront un rapprochement des fichiers.

Mais comme on vient de le voir, la donnée administrative a un statut, une finalité et un cycle de vie (liés à son usage) très différents d'une donnée issue d'une enquête, ce qui a des conséquences sur les propriétés des statistiques qui en découlent. Si l'on reprend les 5 dimensions de notre grille d'analyse, il apparaît surtout que la statistique publique en **perd la maîtrise**, en bonne partie.

- Le champ, les variables, les domaines de valeurs sont décidés par et pour la gestion administrative, la plupart du temps sans prise en compte des besoins d'information statistique. Toute la difficulté sera donc de comprendre comment ces éléments sont caractérisés pour se donner la possibilité ultérieure de transformer les données en vue d'un usage statistique. La qualité de ce travail est aussi fortement liée aux métadonnées des sources administratives qui ne sont pas toujours bien documentées ou difficilement accessibles[43].

- Les données ne portent pas toujours sur les objets de connaissance habituels du statisticien (individu, ménage, entreprise, établissement), mais sur des objets de gestion (transaction, facture, local, compteur électrique, ...).

- Il arrive souvent qu'on perde la maîtrise et l'homogénéité de la double temporalité (temporalité de référence / d'acquisition), à l'exception notable des données administratives issues de déclarations à période déclarative fixée.

Plus globalement c'est le protocole d'obtention des données qui échappe aux statisticiens, y compris la capacité à vérifier les données auprès de leur émetteur, puisqu'on perd le lien avec l'enquête[44]. N'ayant pas défini *a priori* de cadre probabiliste (plan de sondage, modèles d'imputation, …), on

42 … mais la charge administrative (remplissage de formulaire) peut s'y substituer.
43 Par exemple, le secret fiscal empêche parfois d'avoir une vision précise des traitements mis en œuvre pour calculer certaines caractéristiques.

perd aussi la possibilité de mesurer des erreurs à partir de tels cadres … ce qui n'empêche pas d'en définir d'autres, mais cette fois *a posteriori*. C'est là un point-clé.

2.e. Conséquence sur le processus, comparé aux enquêtes

Préalable : le but de cette partie est de mettre en parallèle simplement le processus d'enquête et le processus de traitement des données administratives pour un usage statistique, en s'appuyant sur le GSBPM. Mais il faut souligner que plusieurs travaux ont eu lieu au niveau européen sur l'adaptation du GSBPM aux données administratives, au moins sur certains aspects. Des projets d'Eurostat ont ainsi cherché à mettre en regard les étapes du GSBPM et le traitement de données administratives (Eurostat 2017c[45]).

Les premières étapes du GSBPM étaient : définition des besoins, conception, élaboration, collecte.

Dans le cas des données administratives, la collecte au sens habituel disparaît, ce qui fait disparaître également la conception de l'échantillon et celle du protocole de collecte, l'élaboration du support de collecte, et la collecte elle-même. Cela n'empêche pas qu'il y ait un travail de définition des besoins (même si celui-ci est contraint par le contenu des données administratives), la conception de quelque chose, et l'élaboration de ce quelque chose.

Si l'on présente les deux possibilités, enquête ou acquisition de sources administratives, on obtient, pour le processus de production statistique, une forme « **en Y** », où :

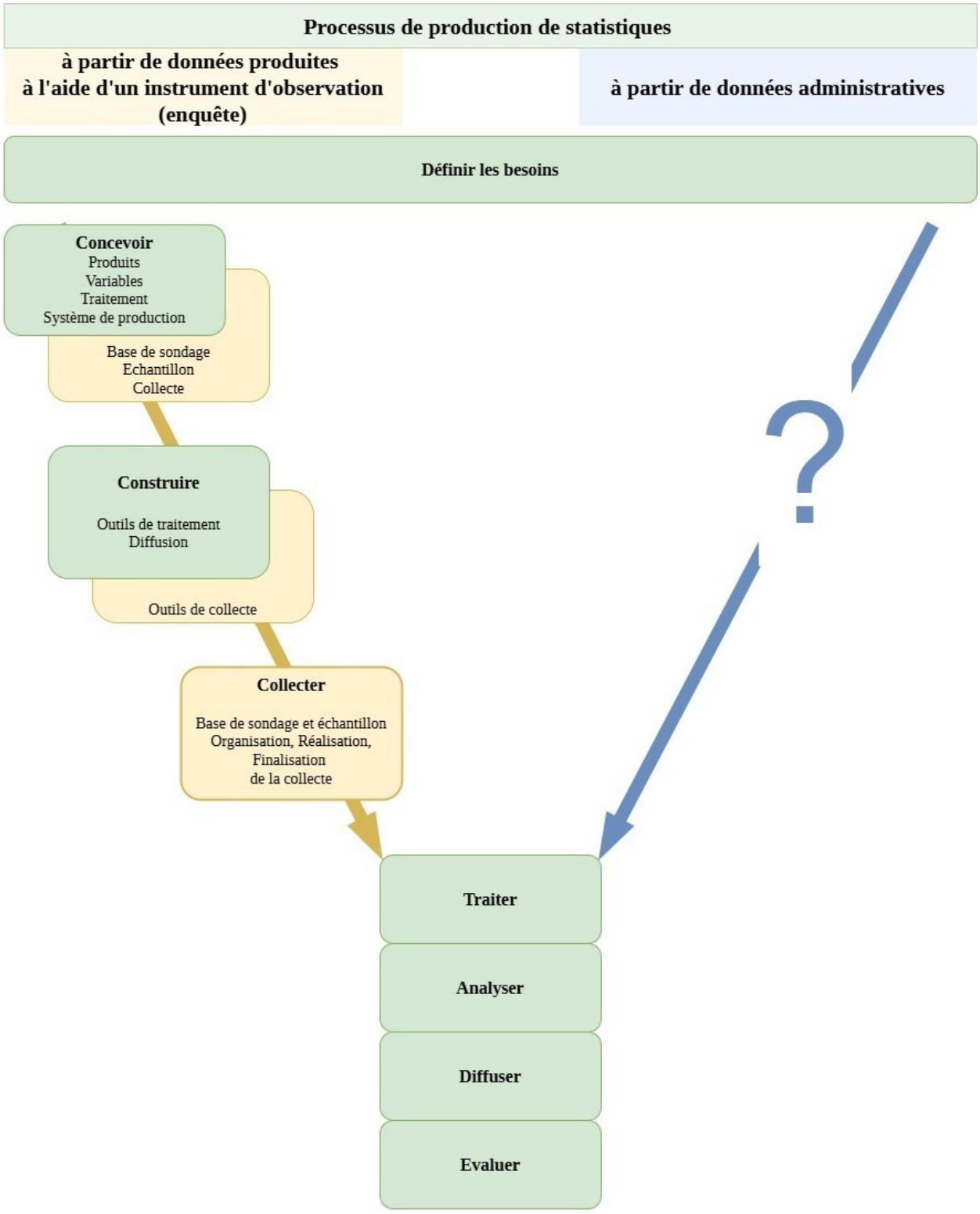
- la branche du Y en haut à gauche correspond aux 4 premières étapes du GSBPM,
- la branche commune du Y en bas représente les étapes suivantes du processus, en commençant par « traitement »,
- la branche du Y en haut à droite correspond, pour les sources administratives, au processus préalable au traitement statistique, qu'il s'agit d'expliciter dans le présent document.

Le processus à concevoir et élaborer consistera, pour le statisticien public, à se ramener à « ses » objets, champs, variables, domaines, temporalités. Il s'agira donc de construire un processus de *conditionnement* des données à partir de ce que fournit l'administration, à savoir des fichiers sur lesquels la statistique n'a que très peu (ou pas du tout) de maîtrise, et qui sont … tels qu'ils sont : les « *it is what it is* » data sets (Lothian, Holmberg, Seyb 2019).

Ce nouveau processus s'applique non pas à une donnée isolée, mais à un ensemble de données plus vaste : une « **source administrative** ». Mais existe-t-il une vision partagée de ce qu'on entend par là, une définition reconnue du concept ? Le chapitre suivant va nous montrer que cela ne va nullement de soi.

44 Y compris l'existence même des données, car rien ne garantit que la source administrative va rester disponible *ad vitam aeternam*. Ce qui a des conséquences sur la pérennité / comparabilité des statistiques.

45 La partie 2 s'intitule justement « *Applying the GSBPM to the usages characterised by integrated administrative data* ». La partie 3 détaille l'étape de traitement.



Ce schéma représente les phases du GSBPM, en mettant en exergue quelques sous-phases. On distingue les sous-phases liées aux statistiques envisagées (la finalité) des sous-phases plus concentrées sur les moyens d'y parvenir. Et notamment, pour ces dernières, celles liées à la conception, la construction et la mise en œuvre de l'instrument d'observation dans un contexte d'enquête (en jaune sur le schéma). Ce sont ces sous-phases pour lesquelles on va chercher à décrire ce qui se passe dans le contexte d'utilisation de données administratives à des fins statistiques.

3. Système d'information administratif et sources administratives

Dans une administration, les données ne découlent pas d'un processus de collecte à l'issue duquel elles seraient conservées quelque part. Non : elles sont en réalité dans un *système d'information*, qui est en quelque sorte leur habitat naturel. C'est un environnement complexe dans lequel les données ne cessent d'évoluer, en fonction des activités de gestion. La notion de source est alors plutôt une construction, une invention du statisticien pour ses propres besoins.

Dans ce qui suit, on va décrire brièvement en quoi consiste en système d'information (SI) administratif, puis on essaiera de comprendre comment on peut « figer » les données, avant de tenter une caractérisation de la source administrative comme découlant du SI. On verra ensuite que le fait de changer d'univers, i.e. de passer d'un univers métier (ici administratif) à un autre (ici, statistique) est une opération délicate. Ceci nous conduira à une démarche en plusieurs temps.

3.a. La notion de système d'information administratif

Un système d'information (SI) est un ensemble organisé de ressources (matérielles, logicielles, humaines, procédurales) qui permet de collecter, classifier, traiter, stocker, diffuser et rendre accessible l'information au sein d'une organisation. Il a pour vocation de soutenir et améliorer le fonctionnement, la prise de décision, la coordination et le contrôle des activités, afin d'atteindre les objectifs stratégiques et opérationnels de l'organisation (Kalika et *al* 2023).

Dans une administration, dans une entreprise, un système d'information se décompose en de multiples sous-systèmes d'information (ex : SIRH, SI des achats, SI de la relation client, ...) ayant chacun des finalités, et au sein desquels s'opèrent des processus opérationnels (par exemple, dans un SIRH, le processus de recrutement), souvent interconnectés.

C'est donc un monde en soi, complexe, enchevêtré, évolutif, et hétéroclite du fait de sa construction au fil du temps. Il embarque avec lui une certaine technicité, des notions métier, un jargon, une culture, des objectifs ... et va naturellement générer et mobiliser de nombreuses données, pour ses propres besoins. Ces données, imbriquées dans le SI, sont pour l'essentiel inutilisables telles quelles pour un usage de recherche, d'analyse, de statistique, et notamment parce que ces données évoluent en permanence, en lien avec la gestion de tel ou tel processus administratif. Elles sont là en tant que support d'un processus métier.

Ainsi, il faut garder à l'esprit que les données administratives ne sont pas, du moins pour l'essentiel, des données immuables, fixes, mais au contraire des données vivantes, subissant des modifications de diverses natures, de façon continue (ajout d'un objet, suppression d'un objet, modification d'un attribut d'un objet, évènement concernant un ou plusieurs objets…)[46]. Elles se trouvent à l'intérieur de *bases de données*, qui sont justement des structures permettant de supporter des évolutions permanentes sans compromettre l'intégrité de l'ensemble (Elmasri, Navathe 2016). Dès lors, sans transformation préalable, on ne peut s'en servir à des fins statistiques : en raison de leur caractère évolutif, n'importe quelle agrégation produirait un résultat différent, pertinent uniquement pour le moment où il a été produit[47].

46 Il faut toujours distinguer les évolutions au niveau « instance » (ajout, suppression de tel individu) et évolutions au niveau du modèle de données, donc au niveau « définition » (ajout de la notion d'ayant-droit, par exemple).

47 On peut simplement souligner qu'il y a toujours peu ou prou un sous-système d'information dédié au pilotage, qui contient des données figées et rejoint pour partie des préoccupations statistiques.

On parle ici de SI *d'une entité publique*. C'est cette entité qui, *via* des unités opérationnelles qui sont chargées de cela, va transmettre ce que l'on nomme « sources » administratives. Il ne s'agit en général pas d'une collecte, et les données ne sont pas nécessairement des observations : les statisticiens publics peuvent avoir tendance ici à projeter leurs propres représentations, à tort.

Le SI administratif, vaste, protéiforme, souvent évolutif, ne peut à aucun moment être une telle source, et il est pratiquement toujours inconnu des statisticiens : pour que ceux-ci puissent travailler, il faut un fichier, comportant des données figées. Afin de produire des statistiques, et une fois assimilé les aspects importants du SI, il faudra qu'aient été définies des modalités pour **figer** les divers ensembles de données pertinents du SI.

A l'inverse, dans un processus d'enquête, on organise un processus d'observation ponctuel, produisant par construction même un jeu de données figé, comportant le plus souvent une valeur unique et datée pour chaque variable d'intérêt.

Mais alors, comment figer ce qui peut bouger en permanence ? Et en partant de quoi ?

3.b. Figer les données : une construction nécessaire, des choix décisifs

Pour produire des statistiques, on ne peut pas partir de données dont les valeurs sont susceptibles de changer tout le temps. D'une manière ou d'une autre il faut disposer d'un fichier, dans lequel les données sont donc fixes, et il est ainsi impératif de savoir « prendre une photo » de ces données mouvantes, c'est-à-dire de **figer** les données. Si cette photo n'a pas vocation à être prise par les statisticiens publics, il leur faut en revanche comprendre *comment* elle a été prise, car elle implique un certain nombre de choix[48].

On propose de distinguer 4 cas de figure pour ce figeage[49], dépendant de l'organisation du SI, en allant du plus simple au plus complexe :

- ***Lorsque la source administrative repose sur des déclarations administratives à temporalité maîtrisée***

Dans un processus déclaratif (une déclaration sociale, par exemple), le figeage des données est naturel, au sens où ce processus fournit par construction, indépendamment des besoins des statisticiens, des données qui ne bougent pas : ce qui est déclaré à une date t[50].

Un sous-produit naturel de ce processus est le fichier contenant l'ensemble des déclarations et des déclarants pour la période de collecte, telle qu'elle est définie par le processus de collecte. Il s'agit en quelque sorte d'un processus de **figeage « canonique »**, qui a lieu pour chaque période déclarative, en lien avec l'organisation de la gestion, sans aucune autre considération.

48 … comme une véritable photo, d'ailleurs : sujet, date, lieu, éclairage, orientation, précision, ...

49 Le *figeage* est donc l'action de figer. C'est un terme qu'on trouve dans d'autres domaines, en chimie par exemple, et qu'on va utiliser dans le reste du document. On pourrait aussi parler de « gel » des données. Le mot « extraction » pourrait aussi être employé mais il ne met pas assez en évidence l'action de fixer / figer la temporalité et tout ce que ceci implique.

50 Même si on peut imaginer d'avoir des déclarations rectificatives ou des décalages de paie (par exemple des primes versées en décalé que l'on pourrait avoir besoin de rapprocher de la bonne période d'activité).

Dans la pratique, la photo n'est pas non plus parfaite ; d'une période de collecte à l'autre, il peut y avoir des modifications rétrospectives (c'est le cas pour la DSN) ; mais cela reste marginal, et pour l'essentiel la valeur relative à une période est unique. Les données déclarées sont par essence figées, conditionnellement au processus déclaratif.

> Exemple : dans la source mensuelle Pasrau (Prélèvement A la Source - Revenus Autres), qui correspond aux déclarations de revenus « de remplacement » de type allocation chômage, retraites etc. (Berthelot 2020), on constate que pour obtenir 100 % des versements relatifs à un mois donné, il faut prendre en compte les 5 livraisons mensuelles consécutives.
>
> Autre exemple : les données de l'impôt sur le revenu proviennent d'une opération suivant un calendrier précis (la période de déclaration 2024 commence le 11 avril et finit le 6 juin 2024), une population bien définie[51], une période de référence claire (année 2023)… Un ensemble de données figé issu de cette source peut être obtenu « simplement » par agrégation des déclarations parvenues à l'administration pendant la période de collecte[52].

Le processus de déclaration, au sein duquel différents éléments sont définis *ex ante* et de manière rigoureuse (période de référence, population d'intérêt, les concepts…) est donc un bon cas de figure (Rivière 2018), dans lequel **la définition du figeage ne souffre (presque) pas d'ambiguïtés**. On parle bien ici de déclarations *à l'initiative de l'administration* (déclarations fiscales par exemple), avec un processus de collecte organisé.

C'est l'entité administrative qui envoie le « fichier », concaténant les déclarations, à la statistique publique, et qui en porte la responsabilité. Mais même dans ce cas idéal, tout n'est pas aussi simple en pratique : on peut avoir plusieurs livraisons, chacune caractérisée par des temporalités, un périmètre, et même parfois plusieurs fournisseurs, ce qui complique encore. C'est donc tout un **processus de livraison** qui est à mettre en place, avec tout ce que cela suppose de conventions.

- ***Lorsque l'entité concernée produit déjà régulièrement, pour ses propres besoins, des jeux de données figées***

Une administration, ou une entreprise, a très souvent besoin d'avoir une vision claire de son propre fonctionnement : effectifs, achats, budget, production réalisée… Les bases de données de gestion, vivantes, ne permettent pas telles quelles de disposer cette vue d'ensemble. L'entité est donc fréquemment amenée à produire de manière récurrente des « états », des photographies d'une situation, à partir de ces bases vivantes.

Il existe donc d'une certaine manière des « produits sur étagère», des *data products* (Meierhofer, Stadelmann, Cielibak 2019), dans lesquels on peut piocher s'il y a accord entre les deux parties. Ils peuvent parfois faciliter le travail du statisticien, à plusieurs conditions : en connaître l'existence, avoir un accord d'utilisation (convention), disposer d'un minimum de documentation, et bien sûr

51 On peut en trouver une définition ici : https://www.economie.gouv.fr/particuliers/declaration-revenus-reponses-questions#

52 On note que si le figeage ici semble simple (ce qu'il n'est pas, d'ailleurs), il ne garantit pas que les données soient adéquates à un usage statistique.

que leur contenu soit utile. La gamme de ces *data products*[53] est très variée, mais on peut citer en particulier deux types très classiques.

Il peut s'agir tout d'abord d'informations de « stock », **d'états de lieux** calculés à partir de bases de données de gestion, en fixant la date ou la période : par exemple la liste des agents au 31/12, ou bien, pour une bibliothèque, la liste des prêts effectués au cours d'une année donnée. Dans ces exemples, le processus de figeage peut paraître relativement simple.

Mais ce processus peut prendre une forme bien plus élaborée, résultant non pas d'une « photo » mais de calculs complexes à partir d'une grande variété de données : on pense naturellement aux **états comptables** (compte de résultat, bilan), ou à tous les types de **tableaux de bord** d'activité, incluant des indicateurs de production, RH, financiers, ... Dans ce cas, l'organisme produit en quelque sorte des statistiques pour ses propres besoins, ici des besoins de reporting et de pilotage. Cela va donc bien au-delà du strict figeage de données vivantes : on a affaire à des données résultant de transformations, d'agrégations, de calculs élaborés.

Ces jeux de données « sur étagère », construits, sont définis par l'entité détentrice des bases de données. Ceci permet au statisticien de récupérer un ensemble de données qui n'est pas mouvant … mais en même temps, celui-ci se voit imposer la manière de figer *et de transformer* les données, qui parfois n'est pas nécessairement bien documentée.

> On va prendre ici un exemple en dehors du domaine des données administratives, avec des données privées, mais c'est illustratif : lors de la crise sanitaire de 2020, Orange a décidé de partager gratuitement l'exploitation de ses propres données. L'Insee a utilisé ces données déjà travaillées par l'opérateur (il ne s'agissait pas des micro-données) pour effectuer des estimations de population présente sur le territoire (Tavernier 2020).

- ***Lorsque la source administrative repose sur des déclarations administratives événementielles***

L'exemple paradigmatique de la déclaration administrative est celle qui permet de calculer l'impôt, dans laquelle la période déclarative est fixée, décidée par l'administration, de même que la période de référence. Mais beaucoup d'autres déclarations échappent à ce cadre, car liées à un événement.

> On pense naturellement aux événements d'état-civil, tels que naissance (Brumberg, Dozor, Golombek 2012 ; Espinasse, Roux 2022), mariage, décès (Coudin, Robert 2024), mais aussi à ceux relatifs aux entreprises, comme un changement d'adresse ou de raison sociale.

Dans ces déclarations, aucune des deux temporalités n'est maîtrisée. En effet, la temporalité de référence correspond à la date à laquelle l'événement se produit, ce qu'on ne peut décider dès lors qu'il s'agit d'un événement « du monde réel ». Quant à la temporalité de connaissance de l'événement, elle est souvent liée à la volonté de déclarer… or on peut tout à fait décider de le faire avec un grand retard, par exemple pour transmettre à l'assurance maladie une feuille de soins.

53 Il faut souligner ici que le principe des *data products* va en réalité bien plus loin, au-delà des données elles-mêmes : "*A data product is defined as the application of a unique blend of skills from analytics, engineering & communication aiming at generating value from the data itself to provide benefit to another entity*." Le concept de *data product* est aussi associé au concept de *data contract* qui reprend et développe l'idée de contrat d'interface historiquement utilisée dans les systèmes d'échange.

Ainsi, avec les déclarations événementielles, si l'on obtient bien des données figées (l'événement que l'on décrit ne va pas changer, il est fixé une fois pour toutes), la non-maîtrise des temporalités a de fortes conséquences sur la maîtrise du champ : si la personne oublie de déclarer, on ne le sait pas (à l'inverse, dans un processus organisé et décidé par l'entité administrative, on pouvait repérer l'absence d'une déclaration « attendue »).

> Exemple : la cessation d'activité d'une entreprise. Un tel événement peut être connu du répertoire Sirène plusieurs mois / années après la réelle fin d'activité de la société. En effet les procédures légales de liquidation d'une société peuvent être extrêmement longues.
>
> Exemple similaire sur les décès : le RNIPP contient des individus de plus de 120 ans. En effet si la personne décède à l'étranger et que l'acte de décès n'est jamais envoyé en France, cette personne n'aura jamais le statut de décédé dans le RNIPP.

On observe ainsi une fréquente dissymétrie, dans les répertoires, entre naissances et cessations / décès. Pour pallier ces problèmes de couverture, le statisticien va devoir créer des concepts de « cessations statistiques » qui vont permettre de ne pas prendre en compte dans le champ des unités statistiques encore administrativement « vivantes » mais que l'on considère comme décédées ou sans aucune activité économique pour le cas des entreprises.

Au-delà même de ces questions de retard de déclaration, il faut toujours avoir en tête que les finalités de la gestion administrative sont différentes de celles de la statistique, et en particulier visent rarement une cohérence à l'instant t (voire même ne visent pas de cohérence du tout)[54].

- ***Les autres situations dans lesquelles les modalités de figeage ne sont pas univoques***

Dans le dernier cas de figure, complémentaire des trois autres, on ne dispose pas de déclaration administrative, et il n'y a pas de produits sur étagère prêts à l'usage qui convienne : il faut donc construire le jeu de données figées utile à l'étude statistique, à partir de bases de données de gestion. C'est probablement le cas le plus fréquent dès lors qu'il n'y a pas de déclaration.

> Cela concerne les statistiques de la justice par exemple, où il faut passer « d'une donnée de gestion vivante à une donnée statistique millésimée » (Chambaz 2018), mais aussi le domaine médico-social par exemple (Berthe 2025).

Cette fois, la définition des modalités de figeage, voire des modalités de transformation, va résulter d'une **concertation** entre entité fournisseuse et entité statistique. Pour les statisticiens, il est essentiel de comprendre dans un premier temps quelles sont les données existantes dans les bases, leur signification, leur cycle de vie, leur logique d'utilisation dans les processus métier. On peut alors déterminer celles qui sont les plus pertinentes, jusqu'à aboutir à une convention entre les deux entités. Il existe ainsi en France de nombreuses conventions entre services statistiques publics (Insee ou SSM) et administrations.

La convention explicite les variables à sélectionner, le périmètre pertinent, les temporalités, et les transformations à effectuer. On arrive ainsi à des *data products* qui ne sont plus sur étagère, mais

54 Par exemple, en cas d'arrêt maladie, les heures non travaillées seront supprimées du décompte d'heures pour le mois concerné par l'absence. En revanche, si la rémunération a été versée dans son intégralité, les éventuelles corrections ne seront pas portées sur le mois concerné (rémunération réellement versée) mais plutôt sur le mois réel de prélèvement. D'où un décalage temporel possible heures travaillées vs rémunération associée, ce qui est gênant pour calculer un salaire horaire.

sur mesure, issus d'une spécification partagée entre statistique publique et fournisseur de données. On ne dépend donc plus de ce que l'administration a décidé de construire … mais le travail préalable est beaucoup plus long : la **convention**[55] qui en résulte donne un cadre contractuel aux échanges, une description formelle qui sera utile aux comparaisons dans le temps.
L'élaboration de ces conventions peut prendre beaucoup de temps, et on peut l'illustrer clairement … avec des exemples de données privées : les données de caisse, pour lesquelles 10 ans auront été nécessaires (Leclair 2019), et les données de comptes bancaires (Bonnet, Loisel 2024).

Le cas des données privées : quelques remarques

On a évoqué ici aussi bien les données publiques que privées, mais il existe cependant des différences notables, qu'on ne va pas détailler ici (Lesur 2025). L'utilisation de données privées soulève des questions de confidentialité spécifiques, peut engendrer des coûts significatifs (les entreprises voulant vendre leurs données, cf. le cas de la téléphonie mobile), et nécessite une organisation particulière (Chaleix, Mikol 2024).

Par ailleurs, le risque de disparition de la source, ou de non-comparabilité dans le temps (car les structures de données ont changé), est probablement plus important que pour les données publiques, qui, si elles sont loin d'offrir des garanties absolues, bénéficient malgré tout d'un peu plus de stabilité. Symétriquement, la notion de déclaration, avec ce qu'elle implique d'obligation formelle, est plutôt spécifique des données administratives.

3.c. La « source administrative » : un concept fragile

Le passage d'un système d'information vivant à un fichier de données structuré, associé à une période / date de référence et à une population d'intérêt déterminée, est indispensable à la production de statistiques. Mais en pratique, les systèmes d'information sont souvent complexes, enchevêtrés, mouvants, partiellement redondants, ce qui complexifie le figeage. Même les processus qui semblent les plus proches d'une enquête, comme ceux fondés sur des déclarations administratives, combinent souvent les données de déclarations avec des données plus anciennes qui décrivent la population d'intérêt.

> Exemple : le fichier d'imposition des personnes, fourni chaque année par la DGFiP. Celui-ci contient, outre les personnes ayant fait une déclaration d'impôt l'année précédente (champ d'intérêt statistique), les personnes ayant fait une déclaration lors des 3 dernières années.
>
> Autre exemple : les demandes d'aides européennes des agriculteurs. Les données de la PAC (politique agricole commune) comportent à la fois des déclarations relatives à une période donnée et des informations structurelles sur les exploitations pouvant être d'une origine plus ancienne (et pas toujours à jour).

Ainsi, le fait de figer les données fait souvent intervenir, même dans les cas les plus simples, des informations de filtrage pouvant opérer sur un champ plus large que les seuls déclarants.

55 On détaillera ce sujet au chapitre 4, dans la partie 4.b.

De manière générale, le processus de figeage est mis en œuvre pour aboutir à une « source » à partir du SI, mais il n'y a pas nécessairement unicité du processus (plusieurs démarches sont possibles), ni permanence dans le temps (la structure des données dans le SI évolue). Une conséquence importante est que **plusieurs sources peuvent découler d'un même système d'information**[56]. L'opération consistant à figer peut intervenir à différents moments, mettre en œuvre des filtres, des sélections d'information, des agrégations ou regroupements. Elle peut apparaître comme une « projection » d'un SI vivant vers une photo d'un état daté et souvent partiel de ce SI.

De multiples projections sont possibles, les sources issues d'un même système d'information pouvant différer sur un grand nombre de critères :

- champ d'intérêt (sélection des objets et des informations),
- fréquence de mise à disposition de la source,
- fraîcheur des informations,
- périodicité des informations,
- période de référence,
- granularité des informations,
- transformations opérées sur les données.

La notion de transformation[57] recouvre ici un spectre assez large, et comporte entre autres les changements d'« objet » de référence. En effet, comparés aux enquêtes portant généralement sur un petit nombre d'objets (souvent un seul objet : l'individu, l'entreprise ; ou deux objets en lien comme un ménage et les individus qui le composent), les systèmes d'information administratifs possèdent généralement plusieurs dimensions, et portent sur de nombreux objets. Par exemple, pour la DSN, les objets entreprise, salarié, contrat, versement, lieu de travail, base assujettie… et pour le cadastre, les objets propriétaire, local, transaction... La multiplicité de ces objets, et la non-unicité des liens entre eux (par exemple, un salarié peut avoir plusieurs contrats) peut ainsi rendre complexe et non univoque le processus de figeage[58].

On convient d'appeler « source de données administratives » un ensemble de données figées, caractérisé temporellement, se présentant sous forme de fichier(s) déduit(s) d'un système d'information administratif par un ensemble de règles de sélection, transformations, filtres, plus ou moins explicites et maîtrisés.

L'existence de plusieurs sources issues du même SI pose différents problèmes et rend cette notion de « source » assez fragile. Ainsi, on peut trouver de la redondance dans les informations, mais aussi des divergences : les figeages étant différents, les informations a priori communes aux deux sources peuvent également diverger. La mobilisation de plusieurs sources issues d'un même SI par le service statistique public demande donc une argumentation, et une évaluation des coûts et avantages des différentes possibilités. La « source » n'existe pas à l'état naturel[59], c'est une construction qui requiert de nombreuses opérations[60].

56 Il peut aussi arriver qu'une source découle de plusieurs SI.

57 On utilise ici le mot « transformation » pour le passage de données du SI à des données administratives figées. On le retrouvera après pour parler du passage de ces données administratives figées à des données brutes statistiques.

58 Ce qui montre au passage qu'en toute rigueur, l'existence même d'un figeage canonique dans certains cas, notamment pour les déclarations administratives, n'est possible que sous certaines conditions.

59 On peut d'ailleurs se demander si ce terme, très utilisé en pratique, est bien approprié...

60 Voir (Denis 2018), ainsi que l'interview de l'auteur à propos de son livre : https://www.nonfiction.fr/article-9517-entretien-avec-jerome-denis-a-propos-de-son-livre-le-travail-invisible.htm

Tout ceci explique la grande difficulté que pose la constitution d'un catalogue des sources administratives. Le site Datactivist, qui promeut une culture générale des données citoyenne, met en évidence ces difficultés : dans sa section « L'écosystème des sources de données publiques »[61], il pointe la nécessité de distinguer les sources, en fonction de l'entité productrice, en fonction de leur caractère primaire ou secondaire. Il stigmatise également la « jungle des sigles ».

On constate que même le nommage d'une source ne va pas de soi : la connaissance du SI administratif d'origine ou même d'un nom général n'est souvent pas suffisante pour que ce nommage soit univoque.

> Exemple : la « source POTE » produite par la Direction générale des finances publiques à partir des déclarations de l'impôt sur le revenu se décline en réalité en quatre sources utilisées par le système statistique public, certaines portant sur un échantillon seulement ou étant déjà agrégées par l'administration.

Lorsqu'on cherche à décrire et à urbaniser les processus statistiques d'un institut, il est alors nécessaire de décrire finement les sources utilisées, afin de distinguer s'il s'agit d'une unique source ou de sources différentes, et de leur associer un millésime. Et, surtout, la notion de source est étroitement liée à l'usage envisagé pour les données qu'elle contient.

> Le système statistique public utilise différentes sources de données issues du système d'information de la CNAF. Celles-ci relèvent de périodicités et dates de consolidation différentes : périodicité annuelle, consolidé 6 mois ou 2 mois après fin année n-1, sur tout le champ ; ou périodicité mensuelle, consolidé 6 mois après fin mois m-1, sur tout le champ, ou bien sur un échantillon tiré par l'Insee pour un panel de la Drees.

Ces extractions et transmissions ne sont pas sans coût pour l'administration qui les produit. Par ailleurs, elles peuvent évoluer dans le temps.

> Ainsi, la CNAF a fait évoluer sa transmission de sources, en abandonnant d'abord les livraisons consolidées à 2 mois, puis en ne conservant qu'une seule livraison de données mensuelles à l'Insee, en charge de les transmettre aux différents services du SSP.

Au total, il apparaît donc que la notion de source est délicate à caractériser, et que selon les circonstances de « naissance » de la donnée, selon les modes de figeage, il existe toute une palette de sources possibles. On trouvera en ***annexe 1*** une présentation de typologies de sources administratives existant dans la littérature et une proposition de nouvelle typologie.

3.d. La difficile transition d'un univers métier à un autre

Déplacer les données d'un environnement administratif au contexte de la statistique publique peut être perçu comme une opération purement technique de transmission de fichiers[62]. C'est pourtant bien plus que cela : il s'agit plus généralement d'un changement d'univers, d'écosystème, qui ne va nullement de soi contrairement à des idées préconçues. Considérée dans toute sa généralité, cette question a suscité depuis une quinzaine d'années une vaste littérature.

61 https://open.datactivist.coop/docs/culture-generale-donnees-section-4

62 … transmission qui ne va pas de soi non plus : dans Informatique, normes et, temps (1999), Isabelle Boydens reprend une citation d'un général américain, lors de la première guerre du Golfe en 1990, où celui-ci regrette amèrement que « la transmission des bases de données soit moins rapide que celle des objets physiques ».

Dans un récent article, (Borgman, Groth 2025) sont allés jusqu'à développer la métaphore de la « distance » entre le créateur de la donnée et l'utilisateur de la donnée, et les différentes dimensions de cette distance (méthodes, domaines, objectifs,..).

- ***Un cliché : il faut transmettre des « données brutes »***

L'idée selon laquelle certaines données (administratives, notamment) doivent être accessibles à tous, sans restrictions majeures, pour promouvoir la transparence, l'innovation et le développement socio-économique a été centrale dans le mouvement de l'Open data. Un moment clé dans l'histoire de l'Open Data fut la réunion tenue à Sebastopol, en Californie, en décembre 2007[63]. Cette conférence a rassemblé des militants, des chercheurs et des praticiens pour définir un cadre initial pour les données ouvertes, qui s'est articulé en huit principes fondamentaux : des données complètes, **primaires**, opportunes, accessibles, lisibles par des machines, non discriminatoires, sans licence restrictive, durables.

Le deuxième principe insiste sur le fait que les données doivent être rendues disponibles sous leur **forme primaire**, c'est-à-dire non transformées, directement issues de la source, et sous forme détaillée. Tim Berners-Lee, l'un des promoteurs majeurs du mouvement[64], a particulièrement insisté sur ce point dans une conférence Ted restée célèbre, en février 2009, dans laquelle il fait répéter à son public « *Raw data now* ! »[65].

Mais ce principe a été rapidement battu en brèche, en particulier dans l'ouvrage collectif « Raw Data is an Oxymoron » (Gitelman 2013) : les auteurs y remettent en question l'idée de données brutes existant de manière « pure » ou « objective ». Outre le fait que les données sont toujours contextualisées par les choix humains dans leur collecte, leur structuration et leur interprétation, ils pointent une illusion de neutralité : présenter les données comme « brutes » masque les biais implicites dans les processus de création et d'analyse des données. Enfin, de façon évidente, les données ont besoin de métadonnées et d'une interprétation pour être significatives et utiles.

Dans « La fabrique des données brutes », (Denis, Goëta 2014) remettent également en cause le mythe des données brutes, qui masque les transformations nécessaires pour rendre les données utilisables, en soulignant que les données sont toujours le produit d'un travail humain et technologique. Ils vont même plus loin, en introduisant le concept de « brutification » des données : mise au format standard, nettoyage et désindexicalisation pour les rendre compréhensibles et utilisables par des tiers. Ils montrent que le processus d'ouverture des données ne consiste pas seulement en un transfert technique, mais nécessite une coordination entre acteurs institutionnels, des ajustements techniques, des compromis.

Ainsi, la notion de « donnée brute » n'a rien d'absolu : par exemple, en statistique d'entreprise, on récupère des données comptables, la statistique publique voyant ces données comme « brutes ». Si l'on se place maintenant du point de vue de l'entreprise, c'est tout l'inverse, les données comptables sont issues d'un long processus de construction, de mise au point, elles découlent de multiples calculs. Les données, de façon générale, sont donc au contraire extrêmement transformées – voir aussi le plaidoyer de (Christen, Schnell 2024) à ce sujet. Pour illustrer cela simplement, il suffit de prendre l'exemple des capteurs (de température, de pollution, ...) : la donnée la plus brute possible

63 Voir (Goëta 2024) pour une présentation de l'histoire de l'open data.

64 A l'origine du web, en étant l'inventeur du protocole http, et plus récemment promoteur du web sémantique.

65 Voir https://www.youtube.com/watch?v=OM6XIICm_qo , vers 11' – 11'30. Tim Berners-Lee évoque bien d'autres choses dans son discours, notamment l'importance de la standardisation des formats et l'idée d'un système de données liées (*linked data*).

serait … une impulsion électrique localisée et datée, donc tout à fait inutilisable telle quelle. *In fine*, on va donc paradoxalement *construire* le caractère « brut » de la donnée, en conditionnant les données précisément pour l'usage prévu.

Changer d'univers ne va pas sans frictions

Paul N. Edwards, dans « A Vast Machine » (2010), montre le rôle central des données dans la science climatique, décrivant leur collecte, leur transformation et leur utilisation pour modéliser et comprendre le climat global. Il insiste sur le fait que les données climatiques ne sont pas des "faits bruts" mais des "produits" créés à travers un processus d'interprétation et de transformation … processus qui différeront entre météorologues et climatologues, qui n'ont pas les mêmes objectifs.

Les météorologues vont en effet se concentrer sur la prévision à court terme, avec des données très précises et localisées, alors que les climatologues étudient les tendances à long terme du système climatique, avec des moyennes et agrégats sur longues périodes, en cherchant à comprendre des interactions globales (atmosphère, océans, cryosphère,...).

Il souligne que ces différences conduisent à des attentes divergentes quant à la nature et à l'utilisation des données, ce qui peut générer des incompréhensions. Les météorologues produisent ainsi des données d'observation essentielles pour les climatologues, mais qui posent plusieurs problèmes dans leur utilisation à long terme, notamment des biais instrumentaux (changements dans les capteurs ou leur emplacement) et des discontinuités (réseaux d'observation non uniformes dans le temps et l'espace, laissant des lacunes importantes dans les archives climatiques).

Ceci nécessite que les climatologues modifient, complètent et homogénéisent les données météorologiques, un processus que certains météorologues perçoivent comme une altération ou une manipulation des "faits bruts".

Dans le même ouvrage, Edwards introduit le concept de "**data friction**", largement repris depuis lors (Bates 2017, Lehuede 2024) : cette notion fait référence aux obstacles, tensions et efforts nécessaires pour collecter, harmoniser, partager et interpréter des données dans des environnements complexes, ce qui se complexifie encore au niveau international. Il emprunte l'idée de "friction" aux sciences sociales et à la physique, où elle décrit la résistance au mouvement.

La *data friction* se produit à l'interface[66] de deux « surfaces » de données, et restreint, au même titre qu'un frottement en physique, le mouvement des données, augmentant le coût et l'effort nécessaire pour les déplacer (Lombardo 2023)[67]. Les mouvements sont en effet entravés par les incohérences entre formats et standards, les biais méthodologiques dans la collecte des données, les limites des infrastructures techniques, ou les tensions institutionnelles.

(Courmont 2021) retrouve des idées similaires dans un autre contexte, hors domaine scientifique : celui de l'ouverture des données du Grand Lyon, à des fins citoyennes, par exemple la mise à disposition de données de mobilité à partir de sources administratives. Il constate la très fréquente inadéquation de ces sources administratives (par exemple celles de la voirie), et la nécessité d'organiser un « détachement » de leur environnement initial pour effectuer un « réattachement » pour un futur usage. Donc, à nouveau, tout un travail de conditionnement des données.

66 Une interface certes technique, mais aussi sémantique, ce qui est central dans le propos du présent document.
67 Pages 5 à 7.

- ***Qu'en est-il de l'usage statistique ?***

Le passage de l'univers administratif à l'univers statistique est une illustration parmi d'autres de la transition, du mouvement des données entre différents domaines métier : la nécessité de conditionner les données, l'existence de *data frictions*, le besoin de se détacher de l'écosystème de départ pour se réattacher au nouveau, tout cela reste valable.

On peut cependant ajouter des spécificités, l'une en amont de ce mouvement côté fournisseur, l'autre en aval côté utilisateur.

Côté fournisseur de données, il faut souligner que la notion de *data friction* a été introduite dans le domaine scientifique (Edwards et al 2011), où on peut considérer, certes en simplifiant pour les besoins du propos, qu'on dispose de données figées, car représentant des résultats d'expériences. Or les données administratives, susceptibles d'évoluer régulièrement en fonction d'événements (événements d'état-civil, demandes de l'usager, décisions de l'administration), n'ont pas, dans de nombreux cas, cette propriété. Le détachement de l'univers d'origine va donc nécessiter, en plus du reste, un travail de figeage des données.

Côté utilisateur, l'usage statistique est un usage à des fins d'information, structuré par la grille d'analyse du premier chapitre (unités, variables, champ, temps, catégories) : si les questions relatives aux unités et variables apparaîtront peu ou prou pour n'importe quel usage, la question de la complétude du champ va jouer un rôle majeur, de même que la qualité de catégorisation. Un travail fin sur les doubles temporalités sera également nécessaire. Enfin, l'usage statistique se caractérise par l'idée d'une « marge d'erreur » possible, et même mesurable, à l'opposé par exemple d'un usage comptable (où l'on doit pouvoir vérifier les choses à l'euro près), ou d'usages avec implications opérationnelles (montant d'une retraite, par exemple).

Dans tous les cas, il s'agira au bout du compte de parler un langage commun, de se doter de conventions partagées, et de traquer l'implicite.

3.e. Statistiques fondées sur données administratives : une démarche en trois phases

La statistique n'est pas la finalité première de ces données, et requiert leur déplacement hors de leur système d'information d'origine. Les fondements même des estimations statistiques disparaissent, et un ensemble de transformations sont donc nécessaires pour rendre possible la pertinence des estimations. Par ailleurs une évaluation détaillée de la qualité des données administratives se révèle essentielle : (Statistics New Zealand 2016) propose pour cela une démarche très complète, sous forme de guide, en analysant les sources possibles d'erreur.

Pour passer des données administratives « initiales » à des données équivalentes à ce qui sortirait d'une enquête, on propose ici une démarche en trois phases, chacune cherchant à résoudre prioritairement une ou plusieurs des difficultés rencontrées.

La première phase, la phase **d'acquisition**, vise à constituer ce qui sera notre « vraie » source administrative de référence. Cette phase incorpore la définition et la réalisation du figeage dont on a parlé précédemment. Elle s'achève par une étape de qualification, qui permet un « go-no go » pour la suite. A la fin de cette phase 1 on a affaire à des données compréhensibles par l'entité

administrative, ce qui permet justement des vérifications, des questions. Soulignons que dans la conception même de cette phase d'acquisition, il faut avoir en tête ce qu'on veut en faire derrière, i.e. quel genre de statistiques on veut produire. La phase d'acquisition sera l'objet du chapitre 4.

Dans la 2e phase, dite de **transformation**, on va délibérément se détacher du monde administratif, pour se réattacher (au sens de Courmont) au monde statistique. On se situe toujours dans une branche séparée du Y (cf. partie 2.e), avec un travail spécifique aux données administratives, qui ne va pas de soi : on l'a vu (Borgman, Groth 2025), on ne passe pas impunément d'un créateur de données à un utilisateur, d'un univers métier à un autre.

La phase de transformation, qui sera l'objet du chapitre 5, s'achève naturellement par une étape de qualification, mais qui n'a plus rien à voir avec celle de la phase 1 : on est cette fois passés au monde statistique, avec des objets, des variables, un champ, … compréhensible par les statisticiens. Cette fin de phase 2, c'est donc le nœud du Y, à la suite duquel les traitements seront communs avec ce que l'on fait dans le cas d'une enquête. Si l'on se positionne vis-à-vis du GSBPM, ces deux premières phases, d'acquisition et de transformation, sont parallèles aux phases de conception, élaboration et collecte qui prévalent pour les enquêtes.

Elles permettent de passer d'un monde à l'autre, ce qu'on peut résumer dans le tableau qui suit:

Ce qui existe dans le monde administratif		**Ce dont on a besoin dans le monde statistique**
Système d'information vivant		Jeu de données figé
Système d'information très riche, disparate		Observations et informations utiles pour répondre au besoin statistique
Pas de période de collecte Pas de période de référence	→	Période de référence associée à une donnée
Listes de codes administratifs		Nomenclatures
Objets administratifs		Unité(s) statistique(s)
Concepts administratifs		Variables d'intérêt
Champ réel administratif		Appartenance au champ statistique Représentativité du champ statistique

Etant désormais dans un monde statistique, on pourra passer à la phase 3, dite de **traitement statistique,** qui est très similaire à ce qui se fait dans les enquêtes (c'est donc très exactement la phase de traitement qu'on trouve dans le GSBPM). Elle présente quelques particularités pour les données administratives (y compris des fonctions qui disparaissent et qui étaient présentes pour les enquêtes) qu'il faut préciser, et qui seront présentées au chapitre 6.

B. La démarche proposée

4. La phase d'acquisition : vers une source administrative qualifiée

4.a. Principes

Lors de cette première phase, on va évacuer plusieurs défauts inhérents au système d'information administratif, en vue de faciliter l'exploitation statistique future des données[68]. On vise ainsi à traiter les questions suivantes :

- le caractère vivant du SI d'origine ;
- la très grande taille, et complexité, du SI d'origine ;
- le manque de clarté des temporalités (période d'obtention, période de référence) ;
- l'hétérogénéité des fichiers récupérés.

L'acquisition va donc conduire à figer les données administratives, en s'appuyant sur la temporalité connue des informations, à travers des opérations du type : filtrages sur les objets ou sur les attributs (on ne prend pas tout ce qui est disponible), jointures entre plusieurs tables du SI d'origine, des fusions de fichiers (par exemple des fichiers par département rassemblés en un seul), mais pas de transformation de données.

La source ainsi produite, issue de cette phase 1, restera ainsi dans la logique de l'entité administrative, avec ses propres concepts, objets, domaines, … C'est là une propriété décisive de cette phase : grâce à cela, les statisticiens parleront le même langage que les représentants de l'entité administrative, pourront échanger, alors que ce ne sera pas possible une fois les transformations opérées.

La phase d'acquisition ne peut se limiter à « « récupérer une source figée et bien délimitée ». On veut aussi avoir une bonne compréhension de sa qualité. La phase d'acquisition incorpore donc naturellement tout un travail de qualification, et des itérations avec l'entité administrative. On produit à la fin de cette phase une source administrative bien cadrée et qualifiée… alors que le vrai point de départ, le SI, était particulièrement nébuleux.

Pour cela, la démarche proposée comporte deux temps, s'inspirant de (Cerroni, Bella, Galie 2014) et (Daas 2009). Ces auteurs séparent d'une part l'analyse de la source et celle des métadonnées, d'autre part le travail sur les données elles-mêmes (cf. ***annexe 2***, pour l'application à Istat).

Le premier temps est celui de la **préparation**. Il comporte une découverte de l'univers administratif considéré, de sa signification, de son fonctionnement, de son cadre juridique : les données ne sont pas des abstractions, elles emportent avec elle un domaine métier, des objectifs, un langage, des pratiques, et il faut en saisir les codes. Tout cela nécessite l'accès à des documentations détaillées, et requiert la connaissance des métadonnées (sur les domaines de valeurs, par exemple), qui vont être indispensables pour la suite[69]. La préparation inclut également l'élaboration concertée d'une organisation de la transmission de données : protocole, contrat d'interface, *data contract*, ...

La **réception – qualification** a lieu dans un second temps. Cette fois, on récupère les données elles-mêmes, mais on effectue aussi toute une série de contrôles automatiques (voire manuels). Ceci exige d'avoir formalisé au préalable ce que l'on entend par qualité.

68 Il faut souligner qu'on parle ici des données individuelles, mais qu'il peut exister des cas où certaines données arrivent sous forme agrégée. On ne développe pas cette question dans le présent document.

69 Voir aussi (Bonnans 2019) pour les métadonnées utilisées dans le système statistique public français.

Cette subdivision en deux temps ne veut pas dire qu'on se limite à un enchaînement linéaire : la qualification peut conduire à repérer des problèmes de diverses natures (valeur de code imprévue, incohérences, distribution de valeurs non plausibles, …), ce qui va entraîner des discussions avec l'entité administrative. On introduira ainsi l'idée de « boucles de rétroaction », qui peuvent permettre d'améliorer les fichiers produits. Ceux-ci seront à la fois un *output* de la phase d'acquisition et un *input* de la phase de transformation.

Enfin, le processus ainsi décrit, avec les deux temps et les boucles de rétroaction, se déroule nécessairement de façon différente lorsqu'on utilise plusieurs fois la même source : c'est le cas pour des usages répétés de la même source par le même producteur statistique, ou pour des usages mutualisés entre producteurs. On en expliquera l'impact sur le processus d'acquisition.

4.b. Le temps de préparation : source, écosystème, convention

La littérature sur les données administratives souligne l'importance d'adopter une approche globale, qui s'intéresse non seulement aux données elles-mêmes, mais aussi au processus qui les produit et à l'organisme qui en a la charge, qu'on va appeler « fournisseur ».

- ***Le fournisseur***

Il se caractérise par un statut, une position, et un cadre juridique dans lequel son activité s'inscrit.

> Ainsi, la DGFIP a naturellement en charge les données fiscales, qui sont au centre de son activité. Les données peuvent aussi être détenues par un opérateur, qui n'a pas de responsabilité métier, mais assure la gestion de flux de données : c'est le cas du GIP-MDS, qui a en charge un très vaste ensemble de déclarations sociales, en premier lieu la DSN, mais aussi la déclaration préalable à l'embauche (DPAE) de la MSA et de l'Urssaf, ou le Pasrau.

Face à une source potentielle de données, une fois qu'on a déterminé quel organisme est légitime pour cette fonction de fournisseur, il faut déterminer si cet organisme, *dans ce rôle*, est fiable. Dans le cas des déclarations administratives, cette fiabilité va découler de l'existence d'obligations légales pour les déclarants.

> C'est le cas pour la déclaration sociale nominative, mais aussi pour le répertoire d'établissements de santé FINESS, pour lequel toute modification doit avoir une base légale (Bensoussan, Bizingre, Courvalin 2023).

On se posera également la question de la pérennité de la source, ce qui ne va pas de soi (cf. la disparition de la taxe d'habitation), voire du fournisseur.

- ***La source potentielle, l'écosystème***

Mais on ne « fournit » pas des données comme on le ferait pour des tables et des chaises, par exemple. On l'a vu au chapitre 3 : avant toute obtention de fichier, il est essentiel de prendre connaissance de manière approfondie de l'univers dans lequel naissent les données, de la raison de leur existence, des problématiques rencontrées par l'administration à l'origine des données.

Un minimum de curiosité, d'ouverture de la part du service statistique, s'impose ici. Cet effort est indispensable pour comprendre la façon dont sont gérées les données, les objets concernés, les liens entre eux, les concepts, l'histoire de leur mise en place, etc. Tous ces échanges nécessitent du temps (cf. la *data friction*). Ils peuvent aller jusqu'à une démarche d'immersion d'un statisticien au sein de l'administration détentrice des données.

> La pratique de l'immersion, plus intensive, est notamment envisagée par l'institut de statistique britannique (ONS), sur plusieurs mois. L'ONS peut ainsi participer à la co-construction de la source produite à l'issue de la phase 1 (Fermor-Dunman, Parsons 2022).

Bien comprendre, c'est aussi cerner la dynamique d'élaboration des données. Zhang (2012) propose ainsi de chercher à identifier les différents types d'erreur[70] pouvant survenir dans la naissance et le cycle de vie des données (Shah et al 2021)[71] : il s'agit d'identifier en quoi la mise en œuvre du processus administratif pourrait s'éloigner de son objectif théorique. Cette analyse est précieuse, elle contribue à bien identifier les critères qui seront mis en œuvre dans la qualification des données.

La nécessité de développer la connaissance de l'environnement, de la source, reste une constante. Elle vaut tout au long de la phase de préparation, mais aussi ultérieurement. Attention cependant, il ne s'agit pas de le faire dans l'absolu, mais dans l'optique **de répondre à un besoin statistique défini**. On retrouve ainsi, mais de façon différente, la **définition des besoins**, première phase du GSBPM : simplement, celle-ci est étroitement mêlée à la découverte de la source. En pratique, soit ce besoin existe et on s'assure que la source possède de données pertinentes pour y répondre, au moins partiellement, soit les données elles-mêmes apportent des opportunités nouvelles[72].

Quoi qu'il en soit, l'analyse approfondie des objets administratifs et des concepts associés, ainsi qu'une compréhension (au moins superficielle) des processus métier se révélera très utile pour bien définir le futur fichier figé et sa structure, y compris lorsque celui-ci n'existe pas par défaut dans l'administration d'origine. Notons qu'il existe aussi, heureusement, des situations où les données administratives, leur environnement, leurs usages métier, sont très bien maîtrisés côté statistique : c'est notamment le cas lorsque les sources sont utilisées depuis longtemps (mais rien n'est acquis, et elles peuvent aussi évoluer dans le temps).

- ***Récupérer les métadonnées sur la source***

Ayant bien appréhendé les tenants et aboutissants de la source, on passe à une analyse d'une toute autre nature, plus pratique : on va rechercher toutes les informations, parfois techniques, permettant de mieux comprendre ce que contient la source, comment elle est construite. On veut avoir des données sur les données, i.e. des métadonnées. Dans cette chasse aux documentations, les pépites que l'on recherche renvoient à notre grille d'analyse : définition des objets mobilisés, des variables, formalisation des domaines de valeurs, description du champ, temporalités.

70 Voir page 43. L'auteur se place naturellement du point de vue d'un statisticien. Mais on pourrait gloser sur le sens du mot erreur, qui n'est pas toujours le même selon l'observateur.

71 Il faut être vigilant sur la notion même de « cycle de vie de données », qui ne fait nullement l'objet d'une standardisation ou même d'un consensus. L'article de 2021 cité en référence en dénombre 76 modèles ...

72 Dans le domaine des données privées on peut citer l'utilisation des données de téléphonie mobile pour évaluer les départs d'Ile-de-France lors des confinements, ou dans le cas des données administratives, l'utilisation de la DSN pour évaluer le chômage partiel : https://blog.insee.fr/suivre-la-conjoncture-lorsque-les-entreprises-repondent-moins-aux-enquetes/ .

Le cas de la DSN, à travers son « cahier technique »[73], constitue à cet égard un exemple complet et rigoureux. Le champ déclaratif et les temporalités (de déclaration, de référence) sont ainsi décrits dans la partie introductive de cette documentation. Les concepts dérivent d'un modèle conceptuel de données et donnent lieu à des « blocs » déclaratifs. On y explicite les variables, appelées « rubriques », qui sont des attributs de ces concepts. Et pour chaque variable, le domaine de valeurs est caractérisé, de différentes manières : typage, liste de valeurs, appartenance à un référentiel externe, voire utilisation d'une *expression régulière* formalisant la structure d'une chaîne de caractères (Friedl 2006). L'ensemble de ces éléments constitue une « norme d'échange », versionnée annuellement[74], partagée par les différents partenaires, et associée à un outillage dédié (Dubrulle, Rosec, Sureau 2023). Soulignons simplement qu'en toute rigueur, il s'agit là d'une documentation du flux, et non de la source qu'on peut construire à partir de là. Par ailleurs, la description formelle est importante mais ne prémunit pas de sujets de qualité liés à des pratiques déclaratives (sur la déclaration des heures par exemple).

Séparer chronologiquement compréhension de la source et obtention de métadonnées est quelque peu artificiel : souvent, les travaux portant sur les métadonnées débutent dès la découverte de la source, et se poursuivent avec le contact avec le fournisseur. Le travail de découverte peut s'appuyer sur les métadonnées lorsqu'elles existent, et surtout sur un dialogue fructueux avec le fournisseur (Eurostat, 2017b).

Si l'on compare avec le GSBPM, cette étape a des points communs avec **l'étape de conception**.

- ***Établir les modalités pratiques de l'échange***

In fine, le but de la phase de préparation est d'organiser la transmission d'une « source » entre un service administratif «fournisseur » et un service statistique « récepteur » de celle-ci … et qui sera lui-même producteur de statistiques.

Pour cela, de nombreux échanges vont avoir lieu avec le fournisseur des données : responsables métier, mais aussi responsables informatiques. Il faut en effet mettre en place des processus d'échange sécurisés et robustes, les données étant souvent volumineuses et confidentielles[75].

On va donc organiser le fonctionnement d'une **coopération** entre producteur de statistiques et entité publique (Chaleix, Vanderschelden 2023), dans laquelle l'aspect relationnel ne doit pas être négligé. Cela peut revêtir différentes formes non exclusives, des échanges d'information par mail aux réunions ponctuelles, en passant par des processus de négociation. Considérer l'entité administrative uniquement comme un « fournisseur de données », c'est faire fi d'une chose : on n'a pas vraiment défini ce qui devait être fourni et ce que cela signifiait, et c'est bien là tout le problème. De telles discussions peuvent être facilitées si l'interaction se fait avec une équipe statistique de l'entité administrative[76].

73 https://www.net-entreprises.fr/media/documentation/dsn-cahier-technique-2024.1.pdf

74 Un autre élément remarquable est que l'évolution de la norme se fait dans un calendrier très précis chaque année ce qui permet à la statistique publique d'adapter ses systèmes d'information aux évolutions de la norme dans des délais très confortables.

75 Cela ne va pas toujours de soi, et il peut arriver qu'une transmission par support physique soit encore une solution de repli en 2024…

76 Il arrive aussi que le service statistique soit lui-même client de sa propre entité administrative, et est alors lui-même confronté aux difficultés de figeage de l'information notamment.

En France, certaines entités publiques disposent de leur propre service statistique : c'est le cas en particulier des caisses nationales de sécurité sociale, comme la Cnaf, la Cnam, ou la Cnav[77]. Celui-ci joue alors souvent un rôle utile d'intermédiaire dans les échanges entre producteurs de statistiques et entité administrative fournisseuse : la communication se faisant « entre statisticiens », la compréhension s'en trouve facilitée, et l'on réduit ainsi la *data friction*.

La question du cadre légal de cette transmission de données se posera aussi. En France, avec l'article 7bis de la loi de 1951, un tel cadre existe pour les données administratives, mais tout ne passe pas par lui. De plus, un problème de gestion de la confidentialité va se poser (Redor 2023).

Tout cela conduit petit à petit à mettre en place une **contractualisation** entre l'administration qui détient ou fournit les données et le ou les services utilisateurs, ce qui se matérialise par une convention entre les parties. Elle permet de décrire précisément ce qui est transmis, par qui, quand, comment, selon quelles modalités, dans quel format, …

On peut citer notamment la convention entre l'Insee et la DGFIP pour l'usage des données fiscales (avec une annexe pour chaque source).

D'une certaine façon, la contractualisation est une manière pour la statistique publique de regagner un peu de maîtrise sur la donnée administrative. Pour cela, l'existence de conventions est indispensable : elles doivent être très précises, notamment quant aux dates de livraison (au jour près) lorsque la source est input d'un processus de production statistique lui-même contraint.

Exemple : la DSN est utile pour la production de données conjoncturelles, dont les délais de publication sont imposés par Eurostat. Tout retard de livraison peut se traduire par un retard dans la publication des données et un non-respect des engagements vis à vis d'Eurostat.

Soulignons cependant que dans le cas d'un usage ponctuel des données, il peut aussi ne pas y avoir de transmission formalisée, pas de convention du tout : c'est la situation dans laquelle le statisticien (ou le data scientist) essaie de « faire au mieux » avec les jeux de données qu'il découvre au fur et à mesure. C'est une situation courante dans les « pôles data » des inspections, dont le travail est limité dans le temps : l'essentiel des données ne présentent plus d'intérêt une fois la mission achevée (Bolliet et al 2025, Berthe 2025).

- **Résultat du temps de préparation**

Concrètement, le travail de préparation a pour résultat :

- un jeu de métadonnées,
- une documentation,
- une forme de contractualisation,
- … et la décision de poursuivre le processus ou non.

En effet, l'un des choix opérationnels possibles à l'issue de cette évaluation est de décider ... de ne pas poursuivre le projet d'exploitation, parce que la source ne sera pas en mesure de répondre aux besoins. Les raisons peuvent être multiples : il peut s'agir de considérations statistiques (concepts ou objets trop différents des unités d'intérêt, champ théorique de la source trop restreint, délai

77 Cnaf = Caisse nationale des allocations familiales, Cnav = Caisse nationale d'assurance vieillesse, Cnam = Caisse nationale d'assurance maladie.

d'obtention des données trop long), mais également de raisons éthiques (manque de proportionnalité dans l'usage de données personnelles, acceptabilité du projet par la société civile).

> On peut prendre l'exemple du programme Résil qui a décidé de ne pas mobiliser des données de santé (utilisation de la carte vitale) ou du RNCPS (répertoire utilisé à des fins de lutte contre la fraude) ou de données sur les titres de séjour pour une meilleure acceptabilité sociale de son traitement au détriment d'une meilleure qualité statistique qui avait été prouvée par des expérimentations sur des échantillons de données (Dupont et alii, 2024).

Si au contraire on décide la poursuite du projet, le temps de préparation aura rendu possible la suite, i.e. le temps de réception-qualification. Celui-ci vise ainsi à obtenir, pour les phases ultérieures, des données dont on connaît le niveau de qualité.

Mais avant même d'expliciter la qualification d'une source, qu'entend-on vraiment par « qualité » ?

4.c. Qualité des données administratives : une notion contingente

- ***Notion de qualité de données***

La qualité n'est pas une notion absolue : elle s'évalue **par rapport à un usage** : « *Quality is fitness for use.* » (Juran 1951). Dans le cas d'un véhicule par exemple, la qualité ne sera pas appréhendée de la même manière selon qu'on l'utilise pour partir en vacances en famille ou pour faire de la vitesse sur circuit.

Cela s'applique à la qualité des données : « *Quality is not a property of the dataset itself, but of the interaction between the dataset and the use to which it is put.* » (Hand 2018). Ainsi, même si certaines caractéristiques sont générales (ex : la conformité à un format), on ne peut pas parler de qualité de données « en soi ». On peut ainsi aboutir à un paradoxe : avoir des données jugées bonnes pour un usage, et non pour un autre. (Edwards 2010) pointe cette difficulté lorsqu'il évoque les incompréhensions entre météorologues et climatologues, qui n'ont pas du tout la même utilisation des données, et donc des critères de qualité très différents.

Deuxième idée importante : la qualité des données est multidimensionnelle. Dans la littérature académique sur le sujet, on retrouve peu ou prou les mêmes composantes de la qualité.

(Di Ruocco et al 2012) définissent des familles d'indicateurs : pertinence, exactitude, complétude, consistance, précision temporelle, accessibilité, interprétabilité, unicité, cohérence, conformité à un standard. (Bontems, Goulin 2013) se limitent à 6 dimensions : pertinence, exactitude, rapidité de diffusion de l'information[78], accessibilité, possibilité d'interprétation, cohérence. (Batini, Scannapieco 2016) introduisent la notion de qualité du schéma de la base de données, et soulignent la nécessité de faire des compromis entre les différentes dimensions. (Loshin 2011) distingue dimensions intrinsèques (exactitude, *lineage* ou traçabilité, cohérence et homogénéité des formats, cohérence sémantique – existence de définition, unicité du nommage) et des dimensions contextuelles (complétude, cohérence, actualité et ponctualité,…). Au passage, on retrouve des idées similaires pour caractériser la qualité *des données statistiques*, dans le code de bonnes pratiques de la statistique européenne (Eurostat 2017a).

78 Cette rapidité, du point de vue du statisticien, c'est en réalité l'écart entre la temporalité de référence et la temporalité d'acquisition. Pour l'utilisateur, la qualité d'une statistique produite comporte un élément d'actualité, et se réfère donc à la temporalité de diffusion, qui sera ultérieure à l'acquisition.

- ***Quid de la qualité des données administratives ?***

Comme n'importe quelles données, les données administratives ont été conçues pour des usages particuliers, et répondent aux critères de qualité associés, différents de ceux des statisticiens publics. Les déplacer dans l'univers statistique nécessite de vérifier leur adaptation à ces nouveaux usages.

Il s'agit à la fois de s'assurer de la qualité pour ses dimensions intrinsèques puisque le processus est réalisé par un tiers, mais également pour ses dimensions contextuelles puisque les données sont « détachées » de leur usage premier, et réutilisées pour un objectif statistique.

> Au Royaume-Uni, l'autorité statistique (*UK Statistics Authority*) a ainsi mis en place des piliers de *Quality Assurance of Administrative Data* (QAAD) incluant aussi bien le contexte opérationnel du recueil de données dans l'administration que la communication avec les fournisseurs de données ou les principes d'assurance qualité de ceux-ci (Babb 2017). Plus récemment, un cadre général d'assurance qualité a été défini en commission statistique à l'ONU, subdivisé en 10 exigences critiques (United Nations 2025).

Le fait que les données administratives servent à supporter l'activité d'une administration n'est pas une garantie de qualité pour les besoins statistiques. Ainsi, elles peuvent ne pas avoir été mises à jour depuis longtemps, par exemple parce que les intéressés n'ont pas transmis l'information.

> C'est le cas par exemple pour la variable adresse dans les données administratives de santé en Irlande du Nord : (Foley et al 2018), après analyse d'un vaste échantillon, montrent que certaines sous-populations sont moins enclines que d'autres à informer l'administration de leur changement d'adresse (les jeunes, par exemple), ce qui engendre beaucoup d'erreurs.

Les données peuvent parfois comporter un **biais**, un risque d'erreur dû au fait qu'on a « intérêt » à déclarer à l'administration telle valeur, pour obtenir une allocation, une subvention, par exemple : on ne parle pas ici de mensonge délibéré, mais plutôt de choix d'interprétation favorable.

> Exemple : l'activité principale d'une entreprise (APE) peut avoir dans certaines situations des impacts très opérationnels, et jouer un rôle dans l'éligibilité à certaines aides.

Plus généralement, contrairement à ce qui se passe en statistique, ou dans des expériences scientifiques, la donnée n'est nullement une fin en soi pour l'administration. Comme on l'a vu au chapitre 2, elle n'existe le plus souvent que comme déclencheur d'actions, comme intermédiaire, comme co-produit. Dès lors, du point de vue administratif, il n'y a pas de problème de qualité si la donnée permet de déclencher les actions adéquates de manière opportune.

> Exemple : une « carrière » est vue par une caisse de retraite comme déclencheur de l'action « calcul des droits à retraite ». Si elle est incomplète, mais que ses manques ne changent rien aux droits à retraite, la carrière est « de qualité », pour l'administration. Elle ne l'est pas pour un chercheur qui veut avoir la totalité de l'information pour ses analyses longitudinales.

Par conséquent, des données peuvent tout à fait être jugées de qualité suffisante pour un usage administratif, alors que ces mêmes données sont considérées comme de mauvaise qualité pour un usage statistique.

En d’autres termes, « qualité » n'est pas synonyme d'absence ou de faible nombre d'erreurs : sans la moindre erreur, sans le moindre dysfonctionnement de la part de l’administration, les données administratives peuvent présenter, pour un regard statisticien, différents défauts.

- ***Quelques exemples de divergence entre qualités statistique et administrative***

Dans les exemples qui suivent, différentes dimensions de la qualité peuvent être concernées.

Des informations approximatives

Le fait que les données soient précises n’est pas toujours une exigence indispensable pour une administration, en particulier sur des périodes anciennes.

> Exemple : l'année de construction des bâtiments dans le fichier du cadastre présente des pics de distribution pour les années multiples de 10 au XIX^{e} siècle. On peut imaginer que ce soit dû aux modalités d’enregistrement de l’information à l’époque. Il n’y a pas de problème de qualité pour les services du cadastre, en l’absence d’impact opérationnel direct. Mais si l’on voulait effectuer une analyse temporelle fine, avec par exemple l’impact des événements de l’époque (1830, 1848, 1870), la qualité serait jugée très insuffisante.

Un format d'information inadapté à l’usage statistique

L'information présente dans les sources peut être exacte mais avec un format qui rend difficile son exploitation, et qui ne facilite donc pas une utilisation sur des volumes importants.

> Exemples : dans le fichier des cartes grises, l'information sur les noms des propriétaires figure dans un seul champ, entraînant une complexité de traitement lorsqu'un véhicule est détenu par plusieurs personnes.
>
> Il peut aussi arriver que des informations soient récupérées sous forme de document scanné en PDF et qu’on soit obligé d’utiliser des techniques de reconnaissance optique de texte sur des images (OCR) pour retrouver les données. (Denis 2018) illustre cette question avec les données de localisation des pistes cyclables dans les villes.

Des conventions pouvant être mal interprétées

Dans la manière d’attribuer des valeurs à des données, une administration adopte des conventions qui lui sont propres, quitte à ce que, vues de l’extérieur, elles paraissent un peu étranges.

> Exemple : les étages de locaux dans les données du cadastre réservent des surprises. Une analyse de la distribution de cette variable[79] montre un pic à … 81 ! Y aurait-il tant d’immeubles méconnus de plus de 80 étages ? Non : il s’agit simplement d’une convention pour désigner l’étage -1, ce qui est le droit le plus strict du concepteur de la base de données. On peut imaginer que le champ était prévu pour un entier positif, et que pour éviter de tout réécrire, on a choisi des valeurs grandes et rarissimes pour coder les étages négatifs.

79 On verra plus loin qu’il est toujours instructif de regarder les distributions *effectives* des variables dans un fichier qu’on récupère.

Des mises à jour non effectuées

Il existe des situations dans lesquelles il n'y a eu aucune erreur lors de l'enregistrement de l'information, mais où, simplement, celle-ci n'a pas été mise à jour. Et si l'on y réfléchit, c'est même tout à fait normal.

> Exemple : un établissement change de numéro de téléphone, ou d'adresse e-mail de contact. En pratique, rien n'assure que cette information soit prise en compte dans le SI[80] : on ne peut pas demander à l'administration d'être sur le qui-vive pour faire des vérifications auprès de tous les établissements, en envoyant des questionnaires tous les jours.

Il en résulte inévitablement des décalages liés à ces problèmes de temporalités. Ils peuvent être dus à l'administration (qui ne peut pas tout vérifier tout le temps), mais aussi à l'usager.

> Exemple fréquent : les personnes qui envoient leur feuille de soins avec beaucoup de retard. Tant qu'elle n'a pas été reçue, tout se passe comme si le soin n'avait jamais eu lieu, car elle n'existe pas dans le système d'information.

De façon générale, cette situation se présente fréquemment pour l'enregistrement des sorties d'un répertoire. Les mises à jour sont souvent asymétriques, les entrées étant toujours connues, notamment lorsqu'elles sont indispensables pour accéder à certains droits (comme exercer une activité), alors que l'absence d'information sur la sortie n'a pas de conséquence pour la personne.

Les valeurs refuges, ou la gestion d'une information incomplète

Il peut arriver que l'entité administrative soit placée dans une situation où l'information dont elle dispose est incomplète, imprécise, voire absente. Et elle peut gérer cela de différentes manières, qui ne sont pas toujours très cadrées.

> Exemple : une administration qui ne connaît pas la date de naissance d'une personne. En pratique, cela peut la conduire à saisir des valeurs hors du domaine : 00/00, 99/99, ou WW/WW. Elles représentent l'absence (ou imprécision) de l'information, mais souvent sans normalisation, et varient ainsi parfois d'une source à l'autre. On peut aussi se servir de valeurs qui appartiennent au domaine, mais cela complexifie leur repérage : par exemple mettre le 1er janvier (01/01) lorsqu'on ne connaît pas la date. Dans ce cas on ne sait pas distinguer cela d'une « vraie » date au 1er janvier.

4.d. Le temps de réception - qualification

Après le temps de préparation, qui nous a permis de mieux comprendre la source de données, de disposer de documentation et de métadonnées, on entre dans le vif du sujet, les données elles-mêmes : (Cerroni, Bella, Galie 2014) soulignent qu'il faut « évaluer » les données administratives issues de leur univers de naissance, avant toute transformation statistique[81]. Mais nous venons de voir à quel point la qualité des données administratives était difficile à cerner.

80 Sauf si de tels changements nécessitent une déclaration à l'administration.

81 Attention, cela ne veut pas dire que les données ne sont pas transformées, qu'elles sont « brutes » : on a vu plus haut, avec (Gitelman 2013) ou (Denis, Goëta 2014), que c'était un oxymore. De toute façon on aura affaire à des données transformées, d'une façon ou d'une autre, mais *au sein de l'entité administrative*.

On a donc un service administratif « fournisseur » qui transmet à un service statistique « récepteur » un jeu de données, qui se présente techniquement sous forme d'un ou plusieurs fichiers. Selon les cas, ce jeu de données a été construit par l'administration seule ou en collaboration entre le service statistique. En pratique, la transmission de données de la part de l'entité administrative ne se résume pas à un fichier : chaque *envoi*[82] peut contenir plusieurs fichiers, et la livraison de l'ensemble des données prévues peut se subdiviser en plusieurs envois.

> Exemple : les déclarations administratives DSN et Pasrau sont transmises en plusieurs livraisons : deux fois par mois pour la première[83], ou quotidiennement pour la seconde, les périodes de référence de ces déclarations étant mensuelles.

Dans la réception – qualification, on veut valider le bon déroulement technique de la phase 1 dans son ensemble, de la constitution du jeu de données à sa transmission, pour préparer les phases suivantes dans de bonnes conditions. Pour cela, on met en œuvre des mesures quantitatives pour qualifier les données. L'idéal est que celles-ci aient été définies au préalable, en s'appuyant sur la connaissance du processus administratif et l'expression du besoin statistique.

- ***Conformité de la livraison***

On veut s'assurer que les fichiers transmis, dans leur ensemble, correspondent à ce qui est attendu[84], et en premier lieu sur un plan quantitatif. On peut chercher à valider la livraison de plusieurs manières et notamment en s'intéressant à des indicateurs de volume des données :

  - le nombre de fichiers transmis (conformité à la convention, ou encore à une connaissance en extension d'une liste de fichiers attendus),

  - le nombre d'enregistrements : il permet de donner une indication sur la conformité de la livraison mais aussi de vérifier des ordres de grandeur relativement au *champ* prévu.
    Par exemple, on est supposé recevoir des données sur 1,2 million de déclarants, on vérifie qu'à l'issue de tous les envois, on est proche de cette quantité.

En pratique, de nombreuses sources administratives sont en effet transmises sous forme de plusieurs fichiers : il peut s'agir de fichiers correspondant à des unités administratives (une direction d'une administration, une agence régionale) ou à un ensemble de traitements périodiques. Le nombre de fichiers attendus, tel que communiqué au départ par l'administration, est ainsi une information importante : on va vérifier que cela concorde avec le nombre de fichiers effectivement transmis.

> Par exemple, si l'administration transmet un fichier par département, une livraison de 50 fichiers serait partielle. Il arrive aussi que l'administration indique directement le nombre de fichiers attendus, voire, de façon bien plus précise, une liste de noms.

L'intérêt de ces premières vérifications très basiques réside dans leur rapidité de mise en œuvre. On épargne ainsi un temps précieux : il n'est pas souhaitable de dérouler intégralement un processus coûteux de chargement de données lorsque, par exemple, un tiers seulement des données a été correctement transmis. L'objectif est d'être réactif et de retourner rapidement vers le fournisseur en cas de problème dans la transmission des données (corruption d'un fichier pendant l'envoi) ou dans la confection des fichiers par le fournisseur (un fichier a été tronqué au moment de sa création), afin

82 On retrouve cette notion d'*envoi* dans le cahier technique de la DSN.
83 A partir de 2025.
84 C'est là un problème tout à fait classique de définition de protocole d'échange.

que ce dernier puisse procéder rapidement à une nouvelle livraison complémentaire le cas échéant. En outre, cette réactivité est vitale dans le cas où les données administratives sont l'input d'un processus statistique conjoncturel très contraint dans les délais (c'est le cas pour la DSN).

Lors de ce contrôle de conformité de livraison, on contrôlera également un certain nombre de *paradonnées* : on pense ici à des informations portant sur le bon déroulement du processus de transmission, d'accueil et de chargement des données, produites éventuellement par les services en charge de l'accueil des données.

- ***Conformité du contenu : domaine***

Pour chaque envoi, on va procéder à d'autres contrôles automatisés, consistant à :

  - vérifier le format : dessin de fichiers, typage (alphanumérique, texte, entier, date, …),
  - vérifier, pour chaque *variable* jugée d'intérêt pour un usage statistique[85], si le *domaine* indiqué dans la convention / documentation est bien respecté : appartenance à une liste de valeurs[86], à un intervalle, respect d'une certaine structure de la chaîne de caractères[87], ...

Tous ces contrôles sont donc réalisés automatiquement, les sources administratives représentant en général des volumes beaucoup plus conséquents que les enquêtes - parfois quelques millions d'individus / objets, contre quelques milliers. L'automatisation est indispensable.

> Ainsi, le système d'accueil et de mise à disposition des données démographiques et sociales développé dans le cadre du programme Résil, utilise un outil générique d'accueil-réception-contrôle des données administratives nommé ARC (Lefebvre, Soulier, Tortosa 2023), permettant des traitements automatisés.

Avec de gros volumes, on ne peut en effet se payer le luxe d'un processus élaboré de vérification manuelle (*manual editing*) sur des micro-données, comme c'est le cas pour les enquêtes auprès des entreprises, où l'on affine la répartition entre contrôles automatiques et manuels (Hedlin 2003).

Un non-respect de ces règles de contrôle, dans des conditions définies par la convention, conduira à un échec du chargement des données. Suite à cela, le service statistique pourra donc être amené, en tant que de besoin, à demander à l'entité publique des correctifs, des explications, voire une deuxième livraison des données.

- ***Plausibilité des distributions***

On passe ici à une part plus subtile de la qualification, qui cette fois ne sera pas pleinement automatisable, et nécessitera un travail d'interprétation. En étudiant la distribution des valeurs prises par les différentes variables, on va regarder s'il ne se passe rien d'anormal. Notre sujet est donc à nouveau le domaine, et dans les exemples qui suivent on se limitera au cas où le domaine est une liste prédéfinie de valeurs.

85 Ce qui suppose de les avoir définies. On ne va donc pas contrôler toutes les données administratives, ce serait une perte de temps, mais uniquement celles susceptibles d'intéresser la statistique.

86 On considère ici la modalité qui signifie « manquant » (par exemple, 9999) comme faisant partie du domaine.

87 Par exemple, un NIR, ou une date, sont des successions de caractères organisées d'une manière précise. Voir à ce sujet (Dubrulle et al 2023) pour des exemples ou (Friedl 2006) pour une explication des « expressions régulières ».

Valeur non atteinte

On peut isoler un premier cas de figure qui ne requiert pas de connaissance métier. Supposons que le domaine d'une variable X soit {a, b, c, d}, et que dans le fichier transmis on ne trouve en pratique que a et d. Les valeurs b et c ne sont pas atteintes, et cela interroge, même si ce n'est pas la manifestation indubitable d'une erreur.

> Exemple : la première transmission des données GMBI comportait uniquement des locaux dont la variable « Nature de l'occupation » prenait la valeur « Local occupé ». L'absence d'autres valeurs a montré une erreur sur la sélection des données transmises.

Pics de distribution

Il peut arriver que des valeurs du domaine soient non pas absentes, mais que leur fréquence observée soit excessivement haute.

> La date de naissance, déjà évoquée, est un exemple évident : on constate que certaines dates sont plus fréquentes qu'attendues, comme le 01/01, ou le 31/12. L'année 00 est aussi un pic. Tout cela traduit probablement une méconnaissance de la véritable date par l'administration, ou même par l'administré, comme on l'a vu. Ce n'est donc pas vraiment une erreur, mais une information utile à conserver pour les usages ultérieurs.

Vraisemblance des distributions

Plus difficile, et encore moins automatisable, on va vérifier la vraisemblance des distributions des variables « intéressantes ». Dans un contexte de primo-acquisition, il n'est pas toujours simple d'avoir des points de comparaison ou des *a priori* solides sur la distribution des informations. Mais on peut par exemple s'attendre à disposer de données dans tous les départements, à avoir une distribution par âge ou par sexe relativement proche de celles de la population totale lorsque la source couvre toute la population. Il est donc toujours utile de consulter les distributions de variables connues pour s'assurer qu'elles ne sont pas aberrantes.

On verra que les contrôles sur les distributions peuvent être utilisés bien plus tard dans le processus, en phase 3, pour évaluer la nécessité de redressements.

- ***Le champ : sur-couverture et sous-couverture***

S'il est complexe de s'assurer que le champ prévu (on parle ici du champ de la source, tel qu'il est défini par l'administration) est bien couvert par le ou les fichiers transmis, on peut au moins vérifier qu'il n'y a pas de problème manifeste.

Cela commence par de simples contrôles sur les volumes (nombre d'objets), en s'intéressant notamment aux évolutions d'une période sur l'autre : si la source n'est pas nouvelle et qu'on dispose de données sur des périodes antérieures, on peut alors comparer les agrégats entre les livraisons, ce qui permet de détecter d'éventuels problèmes ou changements dans les données.

Exemple : lors du passage à GMBI, l'administration fiscale a décidé d'immatriculer toutes les dépendances qui étaient par le passé rattachées au local dont elles dépendaient. Le nombre de locaux immatriculés avait donc explosé entre deux années de livraison. Cette augmentation était « normale », elle reflétait simplement un artefact de gestion, mais a conduit à l'ajout d'un filtre afin de retrouver la population d'intérêt statistique.

On peut aussi se rendre compte qu'une part du champ théorique est absente du fichier, simplement en observant que des types d'objets sont manquants, ou en quantité anormalement faible. On parle alors de *sous-couverture* du champ. La sous-couverture est toujours difficile à repérer en pratique, car « on ne sait pas ce qu'on ne sait pas ».

Exemple : il est arrivé que dans des données relatives à des équipements (Helfenstein 2022), on trouve des départements dans lesquels il n'existe aucune station d'essence, ce qui révèle une invraisemblance.

A l'inverse, il arrive aussi que le fichier comporte des unités hors du champ théorique administratif : c'est la *sur-couverture*. De façon symétrique au cas précédent, on peut le détecter avec des types d'unités non prévus.

La présence de *doublons* est un cas très particulier de sur-couverture : on va au-delà du champ car on compte la même unité deux fois. Ainsi, dans un système d'information où une opération déclenche une action, on peut rencontrer des comportements conduisant à reproduire deux fois la même opération (pour déclencher l'action), ce qui n'a pas d'importance dans le système d'origine mais crée des doublons, et pose problème pour les usages statistiques ultérieurs.

- ***Outiller la qualification***

L'ensemble des traitements à réaliser sur des données (vérification des formats, contrôles inter-variables, contrôles de distribution, …) est synthétisé dans la notion de *data profiling*, ou profilage, bien décrite dans (Olson 2003). Les techniques de profilage, mais aussi les fonctions de standardisation et de dédoublonnage, sont mises en œuvre dans des *data quality tools* (Van Dromme et al 2007, Boydens 2021), disponibles dans des outils du marché ou en open source. Comme l'indiquent (Boydens, Hamiti, Van Eeckhout 2021), ces outils relèvent d'une approche curative de la qualité des données[88].

On soulignera néanmoins que pour déterminer la qualité effective des fichiers administratifs reçus, le statisticien travaille un peu « sans filet » car il n'a pas grand-chose avec quoi comparer. Intuitivement, cette qualité effective mixerait la conformité des valeurs, la plausibilité de la distribution (pour chaque variable), la plausibilité « macro » (taux de non déclaration, valeurs manquantes), l'absence (ou le faible taux) de doublons, voire la vraisemblance des distributions croisées.

La comparaison de données est fréquente dans le monde statistique, mais à cette étape du processus, on se place encore totalement dans l'univers administratif. L'évaluation de la qualité mobilisant des données individuelles ou agrégées issues d'une autre source sera évoquée dans la phase ultérieure.

88 Elle est complémentaire d'une approche *préventive* de la qualité, que (Boydens et al 2021) développent dans leur article : c'est une approche tout à fait intéressante si on se place du point de vue de l'administration, mais qui est hors du champ du présent papier.

- ***Commentaires***

Dans ce travail de qualification, il faut aussi anticiper que tout est indéfiniment améliorable, et éviter, par conséquent, l'effet « tonneau des Danaïdes ». C'est pourquoi il importe de bien expliciter au départ le besoin statistique, i.e. les attentes en matière de qualité des futures statistiques envisagées : quelle granularité de diffusion souhaitée ? Quelle attente des utilisateurs en matière de fraîcheur des données ? Que prévoit-on en matière d'unités statistiques et de variables d'intérêt ?

Ainsi, et cela peut paraître contre-intuitif, qualifier le jeu de données transmises ne consistera pas à caractériser l'ensemble de ses anomalies, invraisemblances, erreurs, mais plutôt à s'assurer d'un niveau de qualité suffisant *pour répondre aux objectifs des produits statistiques* que l'on imagine. Ainsi, dans l'exemple des dates de construction évoqué plus haut (beaucoup d'années multiples de 10), il existe de nombreux usages pour lesquels cette imprécision n'aura aucune importance.

Ce n'est qu'après avoir défini ces attentes qu'on pourra les décliner, autant que possible, en indicateurs mesurables, en cibles de qualité à atteindre, ce qui permet de se donner un cadre, d'autant plus utile qu'il sera réutilisable en cas d'usages répétés, cadre qu'on déclinera en outils.

4.e. La boucle de rétroaction avec l'entité administrative

Dans la mesure où les données administratives présentent toujours des défauts, d'une manière ou d'une autre, le travail de qualification va le plus souvent révéler un certain nombre d'anomalies. Mais contrairement aux enquêtes, on ne peut pas vérifier auprès de l'unité (individu, entreprise) concernée : la seule rétroaction possible a lieu avec l'entité publique, et se fait non pas au niveau de chaque individu (i.e. chaque enregistrement), *mais à un niveau plus macro, celui de la livraison, ou du fichier.* Cela change fondamentalement la nature de la vérification.

Ce qu'on nomme ici « boucle de rétroaction », c'est tout le processus d'échange entre entité administrative et service statistique, qui va éventuellement conduire à de nouveaux envois (de données, ou de métadonnées) ou à une amélioration de la qualité de la source administrative pour les années futures (Renne, 2018).

- ***Un préalable : qualifier les données administratives dans leur univers propre***

Dans tout ce qu'on vient de voir, la qualification s'appuie encore sur les objets et variables tels qu'ils sont définis et connus par le fournisseur. On se situe entièrement dans un monde administratif et non statistique, et ce n'est pas un détail. Car une fois les données transformées pour des besoins statistiques (ce qu'on verra avec la phase 2), elles deviendront illisibles pour l'entité publique et la discussion deviendra impossible[89]. Adopter le langage de l'autre entité (*contribuable* plutôt qu'individu, *foyer fiscal* plutôt que ménage…) peut sembler limitant voire artificiel pour les statisticiens, mais l'expérience montre que c'est une condition nécessaire pour échanger avec elle de manière constructive[90]. Il s'agit d'un point-clé.

Et pour cela, c'est au monde statistique de s'adapter au monde administratif, non l'inverse.

89 C'est le sujet de la continuité sémantique entre émetteur et récepteur des données, ce qui rejoint le sujet « *data friction* » (quand il y a discontinuité). On retrouve aussi la notion de distance entre producteur et utilisateur, évoquée plus haut à propos de (Borgman, Groth 2025).

90 Ce n'est pas uniquement une question de terminologie, c'est aussi la sémantique portée par le langage.

Une telle démarche permet :

- de mieux distinguer les origines des problèmes dans les données (changement de qualité des données, problèmes de transmission, problèmes de traitement) ;
- d'échanger avec le fournisseur en utilisant un vocabulaire, des concepts, objets, variables communs. Or cette discussion est indispensable pour comprendre la signification des éventuelles bizarreries dans les données transmises.

En effet, il est fréquent que cette étape de qualification comporte son lot de surprises, d'*a priori* formulés sur les données qui ne sont pas confirmés. Comme on l'a vu sur plusieurs exemples, la production de statistiques descriptives simples (distributions, fréquences, totaux etc.) permet d'aller plus loin dans la compréhension des métadonnées, et suscite des questions adressées au fournisseur.

C'est précieux : lors de l'exploitation de données administratives, on dispose en réalité de peu d'occasions de rétroagir sur les données. La phase 1, et en particulier la qualification, est **« le »** moment où l'on peut revenir vers l'administration. Cela permet aussi d'engager des échanges constructifs pour faire évoluer à plus long terme les processus administratifs et les données livrées.

- ***Une boucle de rétroaction « macro »***

Le premier chapitre (en 1.c) a montré l'importance de l'activité de vérification dans le processus d'enquête, en particulier avec le concept de *data editing*. On avait bien affaire à une boucle de rétroaction, associée à des contrôles automatisés préalables, qui avaient détecté des *anomalies*. Mais celle-ci conduisait à des modifications de micro-données, donc au niveau de l'unité statistique : à travers le travail de l'enquêteur, qui échange avec l'enquêté (boucle de rétroaction immédiate, « en direct »), ou *via* l'intervention du gestionnaire d'enquête, qui contacte *a posteriori* l'unité enquêtée, l'itération étant donc différée (*manual editing*).

> L'anomalie, utilisée pour la boucle de rétroaction était par exemple : pour telle entreprise, la valeur de la variable Activité principale, 47.21A, est hors domaine (en l'occurrence celui de la NAF, nomenclature d'activité française). Et l'on vérifiera pourquoi auprès de l'entreprise enquêtée, pour arriver à constater, par exemple, une erreur de saisie. Le déclencheur de la rétroaction porte ainsi sur une micro-donnée, et son résultat aussi.

Dans le cas des données administratives, l'étape de qualification va permettre de la même manière de détecter des situations surprenantes, anormales, d'avoir un regard critique, mais le déclencheur de la boucle de rétroaction[91] et son résultat se situeront cette fois à un niveau macro.

> L'anomalie sera cette fois, par exemple : dans 5 % des cas, les entreprises du fichier ont une activité principale 47.21A. On s'en ouvrira auprès de l'administration fournisseuse. Et on constatera peut-être que l'administration a délibérément introduit un nouveau code d'activité, pour ses propres besoins. Il n'y a donc aucun retour vers l'unité (ici l'entreprise), et le résultat de la rétroaction sera, par exemple, une modification de la documentation.

Les anomalies macro sont de plusieurs types (cf 4.d) : non-conformités fréquentes au domaine, domaine non couvert, distribution des valeurs non plausible, champ non couvert (pour des raisons de volume ou de types d'objet), unités hors-champ obtenues, temporalités non respectées.

91 Rendue possible par l'adoption du langage de l'entité publique.

La rétroaction, donc la discussion avec l'administration, va ainsi porter sur les explications de ces anomalies. On peut les diviser en trois catégories :

1. Incompréhensions de la part du service statistique,
2. Erreur ou insuffisance dans les métadonnées transmises (et la documentation),
3. Erreur ou insuffisance dans les données transmises.

Le premier cas se règle par des explications, et c'est aux statisticiens d'améliorer leur propre documentation. Ainsi, l'anomalie « étage 81 » n'était pas un problème de données transmises : il manquait simplement l'information selon laquelle le code 81 signifiait l'étage -1.

Dans le second cas, on va constater qu'une valeur non prévue dans le domaine apparaît (par exemple un code activité), ou que le champ est un peu plus large que ce qui avait été expliqué. La boucle de rétroaction va donc se matérialiser par l'envoi de nouvelles versions de documentation (parfois de simples mises à jour), sans nouvelle transmission de données. On est assez proche du premier cas de figure.

Reste le 3e cas : la documentation n'est pas en cause, et il y a effectivement un problème avec les données livrées. Ainsi, le fait que le champ soit incomplet peut montrer que la livraison n'est en réalité pas achevée (il manque des départements) ; la présence de doublons peut mettre en évidence des incohérences dans le système d'information et conduire l'entité publique à un travail en profondeur ; idem pour une distribution de valeurs anormales dans un domaine. Cela peut engendrer des envois complémentaires, ou qui « annulent et remplacent » des envois déjà réalisés.

Notons simplement qu'il n'est pas toujours possible d'intégrer toutes les modifications nécessaires dans le délai souhaité[92].

On constate ainsi que les modifications possibles ne vont pas porter sur des données individuelles, et qu'elles auront une dimension plus systémique, vertueuse, car conduisant indirectement à améliorer les données administratives, ce qui sera très utile en cas d'usages répétés. On trouvera ainsi dans (Boydens 1999) une démarche d'utilisation « positive » des anomalies dans les données de sécurité sociale belge, généralisée depuis (Boydens 2018). Le terme utilisé pour désigner la boucle de rétroaction est celui de back tracking, issu du « data tracking » de Redman (1997), et il s'inscrit dans une logique d'amélioration continue de la qualité des données de l'administration, légèrement différente de notre cadre d'analyse, centré sur l'usage statistique.

Pour finir sur cette boucle de rétroaction, soulignons qu'elle peut avoir lieu à plusieurs reprises : la qualification des données est en réalité une opération progressive, itérative, qui se déroule tout au long du processus de découverte et d'acquisition des données administratives, et même après.

4.f. Le cas des usages répétés et multiples

- ***Qualifier une source dans le cas d'un usage répété***

Si la qualification est indispensable lors de la première acquisition d'une source, elle l'est aussi dans un contexte d'acquisition répétée. Le temps de la découverte est révolu, la source est connue des statisticiens qui l'utilisent. Une bonne documentation des premiers travaux est bien sûr indispensable et contribue à la transmission de la connaissance.

92 Dans le contexte d'une acquisition répétée, la détection d'anomalies peut conduire à modifier la livraison ultérieurement, et nécessiter des traitements correctifs de niveau statistique pour la période courante.

Ainsi, dans le cas d'une périodicité annuelle, le processus d'acquisition la 2e année devrait être plus simple : l'acculturation à l'univers administratif ayant eu lieu, les étapes de préparation et de réception-qualification ont été en principe rodées, documentées, outillées. Certes, on devra toujours contrôler le processus d'extraction et de transmission des données, effectuer les vérifications de conformité et plausibilité, les boucles de rétroaction, mais l'investissement initial aura été fait.

Cependant, de nouvelles questions vont apparaître, relatives aux évolutions : y a-t-il eu des changements législatifs ? Des modifications de nomenclatures, de concepts ? Un nouveau logiciel de gestion, de nouveaux formats ?

Répondre à ces interrogations implique des travaux plus approfondis : tout d'abord une **veille législative**, enjeu essentiel ; également des échanges réguliers avec le fournisseur, une attention aux métadonnées fournies avec la source, mais aussi, et c'est un point qu'on oublie souvent, une attention particulière aux évolutions des distributions des variables importantes. La mobilisation de plusieurs occurrences temporelles d'un même jeu de données permettra d'introduire des contrôles de cohérence temporelle dans la qualification, ce qui enrichira … et complexifiera le processus.

L'exigence de satisfaction des critères de qualité peut évoluer avec la maturité de l'exploitation statistique : lors d'une primo-acquisition, l'évaluation fera le bilan de la qualité de la source, sa capacité à répondre au besoin à court terme, mais pourra aussi dessiner les évolutions nécessaires pour répondre à l'ensemble des besoins dans un délai plus long. La qualité n'est pas invariable dans le temps, et il faut pouvoir le repérer : le taux de couverture d'une source peut s'améliorer ; le délai d'obtention d'une source peut se réduire ; ou a contrario, une source dont la finalité est de disparaître va voir sa qualité se dégrader.

> Exemple : la disparition annoncée de la taxe d'habitation (TH) s'est traduite par une baisse de couverture du lien entre les foyers fiscaux et leur logement. Il y avait 37 millions de foyers rattachés à un logement en 2022, 27 millions en 2023 et un peu moins de 7 millions en 2024. Des mesures palliatives ont donc été mises en place pour adapter le système d'information. La dégradation de la qualité de la localisation des individus avait ainsi été anticipée pour que ces millésimes soient exploitables, avec l'utilisation de la source GMBI.

Le caractère répété dans le temps de la qualification a une autre conséquence : il faut s'assurer de la traçabilité des traitements, de leur reproductibilité (Pérignon et al 2019). Cela renvoie à la qualité des développements (lisibilité, versionnage), mais aussi à l'importance accrue de la documentation.

Enfin, dans la mesure où des modifications sont inévitables (changements législatifs, par exemple), il faut veiller à ce qu'on n'aboutisse pas à un processus automatisé (de type pipeline) trop rigide, trop difficile à faire évoluer. Il doit être suffisamment souple pour qu'on puisse en modifier des éléments : nouvelles variables, nouveaux domaines, nouveaux indicateurs de qualité, etc. Ce qui est en jeu ici, c'est la mise en place d'une autre boucle de rétroaction, cette fois intertemporelle, qui va porter non pas sur le processus de livraison, mais sur le processus d'acquisition lui-même.

- ***Qualifier une source pour plusieurs usages ?***

Si la qualité est liée à l'usage, alors qualifier une source pour plusieurs usages semble contradictoire : certains objectifs peuvent s'avérer antinomiques (arbitrage classique entre fraîcheur et complétude des données), ou la prise en compte de différents usages peut amener à des analyses trop poussées. Cependant, lorsqu'une source est utilisée fréquemment dans différents processus de

production, il peut être intéressant de mutualiser une partie des travaux de qualification, et surtout d'en partager les résultats (par exemple, les collaborations entre Drees et Insee sur la source Pasrau) : partager les nouvelles connaissances, la compréhension de la source, ou automatiser la production d'indicateurs quantitatifs. Cette question se pose aussi dans un contexte de centralisation de l'accueil des données, pour différents utilisateurs.

La mutualisation peut prendre différentes formes, comme des échanges dans un groupe de travail ou le partage des travaux de qualification sur des domaines d'intérêt différents. Lorsqu'un nouvel usage potentiel apparaît, il peut être nécessaire d'effectuer des travaux complémentaires sur une source déjà connue et qualifiée.

> Exemple : les données du cadastre déjà utilisées pour la localisation des locaux d'habitation ont fait l'objet de nouvelles analyses lorsqu'elles ont été mobilisées dans le projet Fidelimmo (André et Meslin, 2022). En effet, le projet s'intéressait à la variable de surface, qui, n'étant pas utile pour les usages précédents, n'avait pas été qualifiée.

4.g. Résultat de la phase d'acquisition et passage à la phase suivante

La phase d'acquisition s'avère particulièrement inconfortable pour les statisticiens, car ne faisant pas appel à leurs compétences traditionnelles, à leur cursus académique. Elle comporte une large part d'immersion dans un métier qui leur est étranger, avec la découverte de tout un lexique, d'une sémantique, de processus opérationnels, et d'objectifs qui ne sont pas particulièrement connus.

Elle comporte une part de logistique, en quelque sorte, avec l'organisation des mécanismes de livraison, leur contrôle, leur suivi, et l'élaboration de contrats pour mieux assurer leur bon fonctionnement. Elle inclut également une dimension de négociation, de mise au point, d'itérations, se matérialisant par des boucles de rétroaction.

Le résultat de cette phase peut être caractérisé ainsi :

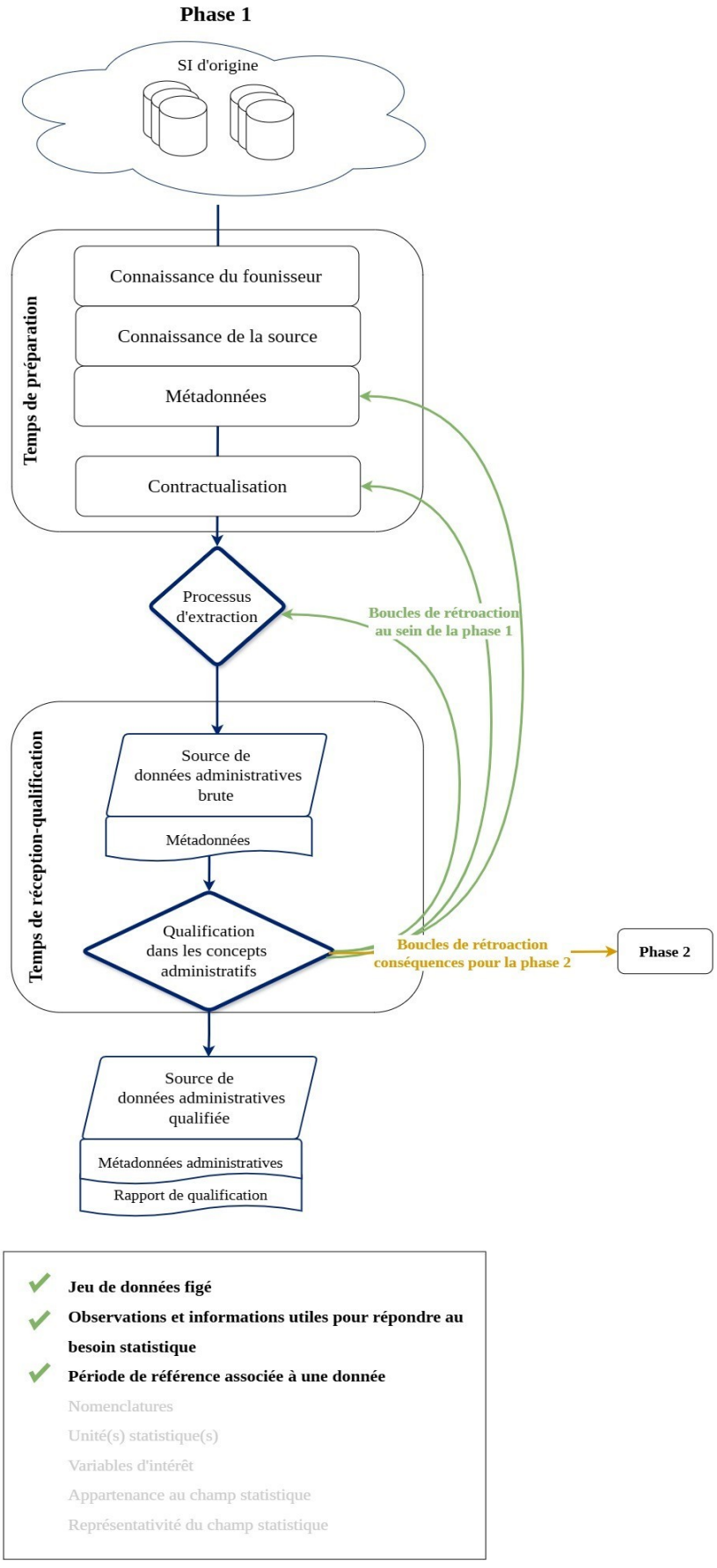
- des fichiers de données, par essence figés
- des métadonnées, une documentation permettant d'en comprendre le sens, facilitant leur usage : définition des objets, domaines, champs, variables, temporalités
- une convention de transmission des données[93]
- une analyse de la qualité des fichiers de données reçus : en effet, on n'a pas nécessairement des fichiers d'excellente qualité, mais des fichiers « dont on connaît les défauts » (incomplétude, non conformité, …)

Tout le travail effectué lors de cette phase s'exprime **dans le langage de l'administration**, avec ses objets, variables, … et **aucune modification de données** ne peut être faite par l'équipe statistique.

Une précision s'impose ici au sujet des boucles de rétroaction avec l'administration : il peut tout à fait arriver, en pratique, qu'il n'y ait aucune rétroaction, i.e. que l'entité administrative ne renvoie ni nouveaux fichiers, ni nouvelles documentations, ni explications, par exemple pour des raisons de manque de temps. On voit alors tout l'intérêt de la qualification : elle aura permis de pointer les défauts de la source, et l'on saura que les phases suivantes devront tenir compte de ces défauts, voire les « rattraper ». Mais pour cela, il faut d'abord passer dans un univers adapté au travail statistique : c'est tout l'objet de la phase 2, phase de transformation.

93 C'est un *output* dans le cas d'une primo-acquisition. Pour les suivantes, elle existe déjà, mais on peut adjoindre des avenants.

***Figure 4* - Phase 1 - Acquisition de données administratives**



5. La phase de transformation : passer du monde administratif au monde statistique

Le fait même que les données issues de la phase 1 relèvent de l'univers administratif constitue une limite. On ne peut les utiliser telles quelles : il faut les **transformer** pour qu'elles s'écrivent désormais dans un langage, un cadre approprié pour les statisticiens. On va donc devoir, pour reprendre la terminologie de (Courmont 2021), après les avoir « détachés » de leur univers d'origine, les « réattacher » à l'univers statistique. C'est un changement de référentiel, une projection de l'univers administratif sur des dimensions statistiques choisies.

L'objectif de **la phase 2, dite de transformation**, c'est d'obtenir un ensemble de données adaptées à un usage statistique, dans le sens où ces données reflètent les unités et les variables identifiées lors de l'expression des besoins.

Cette transformation consiste en une mise en œuvre concrète d'une succession d'opérations sur les fichiers, souvent conçues très en amont[94]. Les données obtenues à l'issue de cette phase auront ainsi un « statut » similaire aux données issues d'enquêtes, ce qui sera essentiel pour la phase suivante.

Dans cette partie, on va commencer par détailler les différentes opérations que l'on peut trouver dans une transformation. On se posera ensuite la question de la réplicabilité de la chaîne ainsi produite, puis on verra qu'un travail de qualification (avec boucle de rétroaction associée) est toujours nécessaire, mais qu'il n'est pas de même nature.

5.a. Principes

Pour passer de l'administratif au statistique, il faut passer d'une grille d'analyse à l'autre : en particulier, on veut un résultat de transformation :

- partant d'une vision claire des objets, champs, variables, domaines, temporalités des fichiers transmis, selon la sémantique administrative ;

- permettant d'appréhender convenablement les unités statistiques, champs, variables, catégories, temporalités choisis pour l'usage statistique. On va donc systématiquement se référer à ces différentes dimensions, en soulignant simplement que la temporalité a été en bonne partie réglée lors de la phase 1.

Il faut donc concevoir ce processus de transformation, en ayant en tête quelques points de vigilance :

1. Les données ainsi produites pourront être utiles à d'autres services statistiques, à d'autres usages. Si c'est le cas, on devra tenir compte de ces enjeux de mutualisation dès le départ (c'est une question importante pour la DSN, par exemple), et prendre en compte les différents usages identifiés ;

2. Selon les choix opérés, la transformation peut conduire à rendre impossibles d'autres usages. En effet, le fait même qu'un tel processus conduise à supprimer de l'information (exclusion d'enregistrements), ou à agréger des informations (unités statistiques fusionnant des

94 Ce qui peut d'ailleurs impacter la conception de la phase 1.

enregistrements), implique une perte d'informations dans les données post-transformation, éventuellement préjudiciable à d'autres utilisations.

3. La transformation n'est pas juste un programme, c'est un objet d'intérêt en soi qui doit être lisible, qu'il faut pouvoir faire évoluer, dont on doit assurer la maintenance.

5.b. Les principales opérations de la phase de transformation

Le processus sera décomposé en étapes, qui vont transformer les données dans leurs différentes caractéristiques : objet, champ, attribut (ou concept), domaine et parfois également valeur. Elles seront le plus souvent déterministes, mais pas toujours réversibles. (Cotton, Haag 2023) proposent par exemple un découpage en six étapes[95].

- ***De l'objet administratif à l'unité statistique***

Restructurer les données par unité statistique

Les données administratives sont relatives aux objets d'intérêt pour l'administration, en lien avec la finalité d'un processus. Ces objets peuvent être complexes, il y a souvent des liens entre différents objets (des individus, des établissements, des évènements …).

La statistique publique s'intéresse souvent à des données portant sur des **individus**, des **ménages**, des **entreprises**, des **unités légales** ou des **établissements**. Elle utilise également des données évènementielles, le plus souvent en lien très fort avec ces unités « habituelles » (naissances, décès, créations, cessations, restructurations etc.).

Le passage d'objets administratifs aux unités statistiques est souvent complexe et requiert une bonne connaissance des données. A partir d'une même source de données, on peut construire différents jeux de données, fondés sur des unités statistiques différentes.

> Les données de la DSN, que l'on peut définir comme des *lignes de fiches de paie*, permettent de construire des données *par poste de travail*, mais aussi *par salarié*.

Il arrive qu'on doive utiliser plusieurs objets, et liens entre eux, pour aboutir à une unité statistique.

> Dans les données fiscales, les liens entre un logement et ses occupants sont utilisés pour définir un ménage. Autre exemple : en statistique d'entreprises, les liens capitalistiques entre unités légales vont permettre de créer des groupes de sociétés.

Le passage des objets aux unités n'est pas nécessairement bijectif : il peut y avoir scission ou à l'inverse fusion d'objets, ou encore sélection de certains liens uniquement. Dans le cas où une unité statistique est constituée de plusieurs objets, cela conduit, de fait, à une sorte de dédoublonnage qui n'est pas « technique », mais lié aux concepts.

95 A savoir : filtrage des enregistrements utiles, renommage, restructuration en unités statistiques, recodage, calcul de variables dérivées, pseudonymisation.

Exemple : l'utilisation des données du DRM[96] par la Drees vise à suivre dans le temps les revenus de certains individus. Cette source comporte par exemple des lignes de revenus ou prestations, *versées par un payeur à un bénéficiaire*. La Drees est intéressée par le caractère pérenne ou nouveau de ces liens payeur-bénéficiaire, afin d'avoir un suivi temporel des revenus ou prestations. Or en pratique, certaines administrations créent pour chaque versement une nouvelle relation, i.e. un nouvel objet « prestation », plutôt que modifier un objet existant. Il y a donc tout un travail pour distinguer les enregistrements relevant de la même unité statistique, i.e. du même couple (bénéficiaire, payeur) donnant lieu à versement.

Le fait que l'unité statistique soit très différente de l'objet administratif peut avoir des conséquences sur l'identification de ces unités.

Dans le cas du cadastre, on trouve des données relatives aux transactions. Dans certaines situations, l'objet administratif est le local concerné par la transaction, auquel seront associés les noms des individus participant à la transaction, mais ceux-ci ne sont pas toujours décrits de façon suffisamment précise pour qu'on puisse bien les identifier. Ceci rend difficile le passage à l'unité statistique « individu » (André, Meslin 2022).

Pseudonymiser l'unité statistique

Passer de l'objet à l'unité statistique, c'est aussi faire en sorte que cette dernière ait les bonnes propriétés pour un usage statistique : dans le cas des données sur les personnes notamment, une propriété visée est d'empêcher l'accès à des informations individuelles, i.e. la possibilité de faire le lien entre une personne et les données à son sujet. Il existe de nombreuses méthodes visant ainsi à préserver la confidentialité des données (Matthews, Harel 2011). On parle désormais de *Privacy-Enhancing Technologie*s (PET), qui abordent la question dans toute sa généralité, avec des développements récents dans le domaine des statistiques officielles (Ricciato 2024).

La *pseudonymisation* est l'une des approches possibles. Elle consiste à enlever, dans les fichiers et pour chaque individu, ses traits d'identité et identifiants directs. Il faut bien la distinguer de l'anonymisation (Esayas 2015). Pour la Commission nationale informatique et libertés (CNIL), « *L'anonymisation est un traitement qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, en pratique, toute identification de la personne par quelque moyen que ce soit et de manière irréversible.*»

En d'autres termes, l'anonymisation est beaucoup plus exigeante, et peut conduire à enlever délibérément une grande quantité d'informations.

La pseudonymisation, en revanche, est une opération « relativement » simple (au moins conceptuellement), et qui peut s'effectuer automatiquement[97] : elle a donc toute sa place dans la succession d'opérations de la phase 2 de transformation. Dans le système statistique public français, le service rendu par le CSNS, ou Code statistique non signifiant (Benichou, Espinasse, Gilles 2023), est une façon de la mettre en œuvre.

96 DRM = Dispositif de Revenus Mensuels. Voir https://net-entreprises.custhelp.com/app/answers/detail/a_id/2362/~/donn%C3%A9es-dsn-et-pasrau-par-le-dispositif-de-ressources-mensuelles-%28drm%29. Pour un développement sur son utilisation dans la sphère sociale, voir (Puche 2022).

97 Voir aussi (Girard-Chanudet 2023), qui apporte un éclairage tout à fait original sur l'automatisation de la pseudonymisation, dans un cadre totalement différent de celui de la statistique publique.

- ***Du champ administratif au champ statistique***

Lorsqu'on évoque le champ, il convient de bien préciser de quel champ on parle : il peut s'agir du champ des données administratives elles-mêmes, ou du champ statistique d'intérêt. Dans la phase de transformation, on effectue le passage de l'un à l'autre, sachant que le champ administratif est imposé par la source. Ainsi, rien ne garantit qu'on pourra couvrir le champ statistique souhaité. Dans la transformation, on va donc surtout veiller à éliminer les enregistrements qui ne relèvent pas du champ d'intérêt. En d'autres termes, dans cette phase, on va traiter la question de la *sur-couverture* et non celle de la *sous-couverture*[98] : il n'est pas question de chercher à compléter le champ (ce sera un sujet de la phase 3).

Le dédoublonnage

Traiter la sur-couverture, c'est en premier lieu éliminer les **doublons** : idéalement, les doublons dans l'univers administratif ont dû être détectés en phase 1. Mais il est possible qu'ils n'aient pas été pris en compte par l'entité administrative dans la boucle de rétroaction, voire qu'il n'y ait pas eu de rétroaction du tout. Il s'agit d'un dédoublonnage « technique », visant à détecter des doublons d'objets administratifs, sur la base de leurs traits d'identité, et pour lequel il existe des outils et méthodes : on en trouvera une large revue dans (Malherbe 2023).

Il est très différent du dédoublonnage « conceptuel » qui peut avoir lieu, postérieurement, dans le passage objets – unités (par exemple, plusieurs liasses donnent une seule entreprise).

Le filtrage du champ statistique

Il s'agit aussi, en s'appuyant sur les catégories disponibles, d'éliminer du champ ce qui n'est pas pertinent pour le champ statistique visé.

> Ainsi, on pourra décider d'exclure du champ des unités de certains secteurs d'activité économiques (secteurs non marchands, par exemple), ou des individus présents dans les données administratives mais n'appartenant pas à la population résidente, ou encore des locaux qui ne sont pas des logements, … tout cela en fonction de l'usage visé.
>
> Autre exemple : chaque mois la CAF envoie à l'Insee un fichier des personnes qui ont eu une allocation au cours des dernières années. Mais on veut se limiter à ceux qui ont réellement reçu une prestation légale le mois considéré : il faut donc filtrer le fichier sur une modalité particulière de la variable qui définit leur « noyau dur ».

Les filtres sont définis à partir de variables disponibles dans la source. Ils sont calculés sur les enregistrements présents dans les données, à partir d'informations existantes et peuvent conduire à sélectionner des objets, des unités ou seulement certaines composantes des objets ou unités.

> Par exemple, les données fiscales comportent des informations sur des locaux d'habitation. Ce concept ne correspond pas exactement à la notion de logement. Il faut donc faire des choix pour passer d'un concept à l'autre. Dans ce cas précis, on pourra exclure les parkings, les dépendances … qui sont des locaux d'habitation mais ne sont pas des logements.

98 Sur ces questions, voir (Wallgren, Wallgren 2016).

Les filtres peuvent être plus ou moins complexes. Il est essentiel de bien expliciter les filtres et leur position relative dans le processus : l'enchaînement des traitements n'est pas neutre. Soulignons ici qu'une telle descriptionfait l'hypothèse a priori de l'exactitude de la variable de filtre.

La qualité des variables de filtrage doit être très soigneusement évaluée : les erreurs sur le champ peuvent ultérieurement avoir un très gros impact sur la qualité des statistiques, et c'est d'ailleurs une explication fréquente de problèmes rencontrés. Les évolutions de ces variables doivent également faire l'objet d'un suivi rigoureux : la connaissance du processus administratif et sa législation, de modifications dans la mise en œuvre.

> On peut citer une évolution administrative récente : l'inscription obligatoire à France Travail « des personnes qui demandent le bénéfice du RSA, des jeunes accompagnés par les missions locales et des personnes qui, en situation de handicap, sollicitent un accompagnement spécialisé ». Celle-ci modifie la définition du champ des demandeurs d'emploi et a donc un impact majeur sur les statistiques correspondantes, conduisant mécaniquement à une augmentation significative difficilement interprétable. La situation est à ce point problématique qu'un groupe de travail (sous l'égide du Cnis) y a été dédié. Il a proposé de nouvelles catégories, « en miroir » des nouvelles catégories administratives : les personnes inscrites « en parcours social » (catégorie F) et les demandeurs du RSA en instance d'orientation (catégorie G). Disposant de ces catégories il devient envisageable de reconstruire le champ de façon appropriée, mais tout un suivi est nécessaire[99].

L'évolution du champ en lien avec une décision administrative peut avoir été repérée en amont, comme dans l'exemple précédent, ce qui permet d'en anticiper les conséquences. Il arrive aussi, symétriquement, que ce soient de fortes évolutions de volume qui permettent de détecter la modification du champ de la source, et donc la mesure administrative sous-jacente.

> Exemple : la mise en œuvre du prélèvement à la source a entraîné un changement dans la population présente dans les sources fiscales. Ce changement a été détecté, lorsqu'on a observé ses conséquences sur le nombre d'individus présents dans les bases de sondage construites par l'Insee. La complexité des traitements effectués sur les données fiscales a rendu délicate l'analyse des évolutions des données en volume.
>
> Autre exemple : on observe des bonds spectaculaires dans les nombres de licences sportives, qui sont simplement dus à des changements dans les pratiques de gestion des fédérations. Ainsi, on va subordonner la pratique de telle activité non compétitive à la prise d'une licence. La fédération française de canoë-kayak a ainsi quasiment doublé son nombre de licences de 2016 à 2019, et on peut l'expliquer par de tels changements (Vicard 2023).

- ***Des variables administratives aux variables statistiques***

Il arrive souvent que les variables disponibles dans la source administrative soient différentes de ce qui est souhaité, mais qu'il soit possible d'approximer les variables d'intérêt à partir d'elles. Il y a donc en premier lieu tout un travail sur le sens des données (administratives et statistiques), avant d'arriver opérationnellement à une formule de calcul et de nommer les choses.

99 Cf. la synthèse du rapport du GT du Cnis "Conséquences de la mise en place de la loi « Pour le plein emploi » sur les statistiques de demandeurs d'emploi" : https://www.cnis.fr/wp-content/uploads/2024/09/rapport-cnis-166.pdf

Le calcul de variables dérivées

On nomme ainsi les variables (ici, statistiques) calculées à partir d'autres variables (ici, administratives), sur la base d'une formule de calcul[100].

> Exemples : des variables qui reconstituent le salaire à partir de la DSN. On peut avoir une règle du type : on prend la rémunération brute déplafonnée, si on ne l'a pas on prend l'assiette brute déplafonnée, et si on ne l'a pas non plus on prend l'assiette CSG.
>
> Autre exemple : (Van Delden et al 2016) construisent un chiffre d'affaires à partir de plusieurs variables de la déclaration de TVA néerlandaise, et font valider leur formule de calcul par des experts du domaine.

A travers cet exemple on voit que la formule de calcul n'est pas toujours du genre z = x + y, mais qu'elle doit aussi tenir compte de l'absence ou de la présence d'informations, en construisant une variable qui fait fi de ces manques. On cherche donc à établir le meilleur proxy de la variable statistique à partir des variables administratives disponibles. Dans tous les cas il est fondamental de conserver la formule de calcul précis (qui est en réalité une métadonnée).

Attention, les variables ainsi dérivées peuvent demeurer différentes de la variable statistique cible.

> Dans l'exemple hollandais cité plus haut : calculé sur des secteurs d'activité construits à un niveau assez fin (324 groupes), le chiffre d'affaires total issu des déclarations de TVA hollandaises présente en réalité des différences avec le chiffre d'affaires cible fondé sur des enquêtes, et ce sont des différences conceptuelles parfois subtiles.

La constitution des unités statistiques a également un impact sur le calcul des variables liées aux objets. Lors de l'utilisation de déclarations administratives, il est possible de choisir comme unité statistique le déclarant et d'agréger plusieurs enregistrements initiaux en une seule unité, en utilisant des fonctions d'agrégation (selon les variables, une somme, un maximum, l'élément le plus récent).

> Exemple relatif aux statistiques structurelles d'entreprise : le passage des objets « liasses fiscales » aux unités statistiques que sont les unités légales entraîne des agrégations de liasses. Les modalités de calcul diffèrent selon la nature des variables : on peut sommer des variables de flux (chiffre d'affaires) mais on conservera la valeur la plus récente pour les variables de stock de fin. Par ailleurs, il faut tenir compte du fait que les données fiscales portent sur l'exercice comptable, qui peut différer de l'année civile : les formules de calcul peuvent se révéler complexes.

Le renommage

Le changement de nom d'une variable est une opération réversible, qui n'a pas de conséquence sur les données, mais sur les métadonnées associées aux données produites. Comme elle est simple, peu coûteuse, on peut avoir tendance à la sous-estimer, voire à l'oublier. Or elle est importante pour assurer une bonne maîtrise / pérennité des concepts, et que ceux-ci soient partagés.

Le renommage est une opération dont il faut garder trace afin de documenter les données pour les utilisateurs finaux. Elle influe sur une dimension importante de la qualité : la lisibilité.

100 Voir par exemple l'index des variables du recensement 2021 en Irlande du Nord, dans lequel cette distinction est bien explicitée : https://niopa.qub.ac.uk/bitstream/NIOPA/15777/1/census-2021-variables-index.pdf

Par exemple, les données fiscales ont des noms tout à fait illisibles pour les non-spécialistes, comme on peut le voir dans les exemples de (Cotton, Haag 2023). Ainsi, les nom, prénom, sexe, et date de naissance sont notés respectivement LNCON, LNCOPF, LCCOT et DNCO.

Le faire de façon rigoureuse, claire, permettra de gagner du temps dans la suite des opérations, car on saura de quoi on parle. Le renommage s'applique aux variables, ou aux valeurs d'un domaine.

- ***Des domaines administratifs aux catégories (domaines) statistiques***

Pour illustrer cette situation de transformation de domaine, on peut considérer une variable initiale dont le domaine est {a, b, c, d, e, f}, qu'on veut ramener à un domaine plus simple : par exemple, les valeurs 1, 2, 3, où 1 correspond à {a, b, f}, 2 à {c}, et 3 à {d, e}.

D'une certaine manière, cette situation est proche de celles des variables dérivées : derrière le changement de domaine, il y a le calcul d'une variable dérivée qui prend ses valeurs dans une liste. On parle ainsi de codage, ou de recodage si l'output est une liste de codes. L'input peut avoir pour domaine une liste, comme dans l'exemple initial, mais ce peut être une donnée quantitative (ex : passage d'un âge à une tranche d'âge). Il faut aussi souligner que le calcul de la valeur dans le domaine cible peut s'appuyer sur plusieurs inputs, avec des règles de calcul complexes.

Ces règles doivent aussi tenir compte des valeurs manquantes et de leur dénomination, en amont comme en aval. En amont, on dispose, suite à la phase 1, de connaissances sur la façon de désigner « manquant » dans les données administratives (par exemple, 999, 9999, -1, …). En aval, côté statistique, on se donnera une convention de notation (par exemple « NA »), et des règles de décision. Mais attention, dans la phase de transformation, on ne traite pas les données manquantes, on n'effectue pas d'imputation (ce sera le rôle de la phase 3) : on établit et applique simplement des conventions de notation et des règles d'affectation de cette modalité à partir de l'input administratif.

Les transformations de domaine doivent aussi tenir compte des évolutions dans le temps des pratiques de gestion : on en a vu plus haut l'illustration avec les nouvelles règles d'inscription à France Travail.

5.c. Un processus qui est aussi technique, et doit être réplicable

La transformation comporte également des enjeux techniques qui peuvent être liés au volume des données, à leur fréquence de transmission, à leur format… La complexité du système d'information d'origine, et notamment les liens inter-objets, se reflète dans les données statistiques disponibles pour le statisticien.

Dans les choix techniques de mise en œuvre, il est nécessaire de tenir compte des usages :

- il faut arbitrer entre fréquence de chargement des données et stabilité de la base,

  Exemple : la livraison quotidienne de données du Pasrau ne peut être répercutée directement dans les données mises à disposition des statisticiens. Afin de garantir une certaine stabilité des traitements, les mises à jour des bases visibles pour les statisticiens sont mensuelles ;

- la structure des données, l'architecture des bases doit répondre aux usages et permettre facilement de filtrer les informations ou les relier entre elles.

Pour des sources régulières, on a besoin de mettre en place un processus de production automatisé, une succession de transformations qu'on appelle souvent *pipeline*[101]. Comme on l'a vu plus haut au sujet de la phase 1, le pipeline doit offrir des qualités de souplesse, pour mettre en œuvre facilement des évolutions. Il doit aussi garantir, jusqu'à un certain point, la réplicabilité (National Academies 2019[102]), avec des conséquences sur la conservation des données avant transformation, la traçabilité du cycle de vie de la donnée (*data lineage*), et, souvent, la capacité à traiter de gros volumes. Idéalement on devrait viser la réplicabilité forte sur des processus de production avec un enjeu fort pour le système d'information.

Mais il faut surtout retenir ici que la transformation, en tant que succession automatisée d'opérations, va produire une nouvelle source, qui relève cette fois de l'univers statistique : elle contiendra des données individuelles relatives aux unités statistiques d'intérêt, avec le champ, les variables et les catégories d'intérêt choisies. Simplement, comme en fin de phase 1, la source obtenue comportera des imperfections, et une étape de qualification va s'imposer.

5.d. Qualifier les données transformées

Comme pour la qualification de fin de phase 1, on va qualifier les données reçues (issues de la transformation) pour détecter les anomalies éventuelles, puis réaliser une boucle de rétroaction afin de les prendre en compte. Si le principe est le même, il n'en demeure pas moins que qualification et rétroaction seront de nature très différente de ce qu'on a vu en phase 1.

- ***Principes***

Tout comme en fin de phase 1, il va s'agir de détecter d'éventuels problèmes, et de repérer ceux qui paraissent suspects. Mais cela s'effectuera dans un univers statistique, un univers connu au sujet duquel on dispose d'informations.

La qualification de fin de phase 2 est une qualification intermédiaire, hybride, puisqu'elle qualifie le passage d'un univers à l'autre ; si on est dans l'univers statistique, on n'a pas encore complètement quitté l'univers administratif. Elle est hybride au sens où elle qualifie les données individuelles *et* les agrégats. Ainsi, son fonctionnement est particulier, engendrant deux types de conséquences différentes : une boucle de rétroaction lorsqu'on détecte des anomalies statistiques, et la facilitation de la phase 3 via le repérage systématique d'anomalies individuelles.

Détection d'anomalies statistiques

La véritable qualification statistique se fera en phase 3, lorsqu'on aura « toutes » les données, imputées ou non. Ici, en phase 2, on se bornera à quelques **vérifications statistiques à gros traits.** En effet, les données et unités manquantes n'ont pas été traitées, et on ne peut construire de statistiques valables à ce stade. On se limitera donc à contrôler des ordres de grandeur, en vérifiant :

101 Voir à ce sujet (Cotton, Haag 2023), déjà cité.

102 A noter que dans les comités constitués pour la rédaction de ce livre, on trouve bien un « Committee on national statistics ».

- la cohérence avec des totaux déjà connus, par exemple le nombre de ménages dans Fideli (Lamarche et Lollivier 2021) ;
- des évolutions de totaux, ce qui permettra de repérer des évolutions de champ (cf. exemple de France Travail) ;
- la complétude du champ, notamment quand champs administratif et statistique sont comparables : estimation de défaut de couverture ;
- éventuellement, la plausibilité de quelques distributions, en utilisant pour cela d'autres distributions connues (recensements, enquêtes).

Pour ces comparaisons, on se situe désormais totalement en dehors de l'univers administratif (on pourra utiliser toute autre source jugée fiable et comparable, au moins sur certaines variables). *Elles se feront cette fois sur les variables et unités statistiques*.

> Ainsi on comptera le nombre de ménages dans Fidéli plutôt que le nombre de foyers fiscaux dans la source FIP, on s'intéressera au salaire net en ETP plutôt qu'à des lignes de déclaration DSN.

Ces travaux sont menés de façon systématique à des fins d'évaluation, mais peuvent aussi être réalisés ponctuellement pour de nouvelles utilisations d'une source administrative.

> C'est le cas des données de la DSN, utilisées depuis plusieurs années pour produire des statistiques annuelles sur les salaires, et notamment la base tous salariés. Elles ont fait l'objet d'une qualification pour un nouvel usage, et ont permis en 2024 de réduire d'un tiers les questionnaires de l'enquête Ecmoss. Cette évaluation spécifique pour ce nouvel usage a été menée en dehors du processus de production de la base tous salariés.

Repérage d'anomalies individuelles

Il s'agira ici de réaliser un « scan » du fichier pour repérer automatiquement les valeurs manquantes (non-réponses partielles) et les valeurs aberrantes, avec des démarches de type *data profiling* déjà évoquées (Olson 2003). Le repérage des valeurs manquantes se fera simplement sur la base de la connaissance de la notation « manquant ». Trouver les données aberrantes est plus complexe, et présuppose d'avoir défini des règles : par exemple, si le ratio chiffre d'affaires / effectif est supérieur à tel seuil, dans tel secteur d'activité, on considère que c'est aberrant.

Cette étape permettra de produire un fichier complet dans lequel on aura les données (micro-données) + des métadonnées micro indiquant pour chaque valeur si elle est aberrante ou manquante.

- ***La qualification de phase 2 ne se substitue pas à celle de phase 1***

Il convient de souligner ici que certains problèmes dans la source ne sont pas détectables en phase 1 (mais seulement par le biais de comparaisons statistiques) alors que d'autres pourraient l'être, mais ont échappé à la vigilance du statisticien, par lacune de la qualification de phase 1, ou par manque de temps : dans de nombreux processus, la réception et l'intégration de données se déroulent dans un temps très court.

Lors de la primo acquisition d'une source administrative, les données sont encore peu connues, et le processus de transformation est encore à bâtir. Il est naturel dans ce contexte de réaliser une

qualification de fin de phase 1, i.e. une qualification des données administratives dans leur univers, concepts, objets etc...

En revanche, en cas d'acquisition répétée, on pourrait avoir tendance à se contenter d'une qualification de phase 2, éliminant ainsi celle de la phase 1, et ce d'autant plus que le processus d'intégration est volumineux, élaboré, ancien, etc.

Ce fonctionnement a priori rationnel peut toutefois être pénalisant à plusieurs titres :

- il complexifie l'identification de l'origine des problèmes : problème dans les données, problème de transmission, problème de transformation ? Au minimum, à défaut d'une étape isolée de qualification de fin de phase 1, il est important de tracer les volumes de données (livrées, intégrées, transformer) ou de disposer de statistiques descriptives sommaires sur les données avant transformation ;
- il rend plus coûteux les échanges avec l'administration d'origine ;
- il peut « retarder » la détection de changements dans les données administratives, ou de problèmes de qualité nouveaux, et ce d'autant que la qualification a lieu tard, parfois même après les traitements statistiques de la phase 3. Ce retard est rédhibitoire pour les productions conjoncturelles fortement contraintes par les délais.

5.e. Une boucle de rétroaction à deux niveaux

Si l'on détecte des problèmes dans les données grâce à la qualification, il ne peut exister que deux coupables : la source, issue de la phase 1, ou la transformation. Les deux sujets sont très difficiles à démêler. Si vraiment il y a un problème avec la source, on doit dans l'idéal pouvoir le repérer en fin de phase d'acquisition, et réaliser la boucle de rétroaction avec l'administration à ce niveau-là, pré-transformation, ce qui est loin d'être simple[103].

- ***Une rétroaction privilégiée : sur la transformation elle-même***

On va donc dans un premier temps décrire une rétroaction sur la transformation (en supposant implicitement que la source est bonne) : dès lors, s'il y a des anomalies, elles doivent résulter de bugs dans cette opération, qu'il faut trouver. Il faut que la transformation soit réplicable : si elle contient des étapes manuelles, ou des traitements automatiques non documentés, il devient très difficile de la modifier.

Les anomalies peuvent avoir plusieurs origines :

- erreur de programmation ;
- erreur d'interprétation des données administratives (définition de variable ou de valeur dans un domaine) ;
- formules de calcul incorrectes : ces formules, pour le calcul des filtres ou des variables dérivées, doivent être très précises, et prendre en compte tous les cas ;
- subtilités dans la constitution des unités statistiques ;
- et surtout : changements d'outils de gestion administratifs, de législation, de processus administratif, entraînant des changements dans les concepts, les domaines, ...

103 Cela signifie en effet qu'il faut avoir conservé l'état des données « fin de phase 1 », ce qui peut être très lourd. Cela veut dire aussi qu'il faut relancer une qualification de fin de phase 1 … et détecter des problèmes.

Exemple : dans sa démarche de qualification des données DRM, la Drees a constitué un sous-échantillon lui permettant de diagnostiquer de la manière la plus complète possible les problèmes de qualité des données. Le but est également de définir les critères pour les détecter ainsi que les corrections à apporter, et de préciser les critères de constitution des unités statistiques d'intérêt. Cet échantillon sert également pour la validation des traitements. En effet, une fois le processus de transformation des données écrit, il est appliqué à l'échantillon de test, ce qui permet de valider que le résultat est conforme à l'attendu.

Ces erreurs sont souvent liées à des modifications fines, parfois peu visibles, dans les procédures administratives. On tombe là sur un écueil très important de l'usage des données administratives en statistique, qui met en lumière le besoin d'une compréhension de l'univers administratif dans lequel gravitent ces données.

Au total, on se rend compte que la mise au point du *data pipeline* est une opération très exigeante, une mécanique de haute précision qui tolère peu les erreurs, même minimes, et requiert de solides compétences de *data engineer* (Schweinfest, Jansen 2021).

- ***Une rétroaction vers la phase 1 plus délicate à mener***

Si la détection de problèmes avec la source n'est pas idéale en phase 2, elle peut cependant se produire pour différentes raisons, et notamment parce que la comparaison avec d'autres sources statistiques permet d'évaluer de façon plus précise les problèmes de sous-couverture, mais également de valeurs lorsqu'il est possible de comparer des données individuelles.

Il est alors utile d'échanger avec le fournisseur, mais avec une difficulté supplémentaire : les données ayant été transformées, le constat sera souvent exprimé dans le langage des statisticiens, avec des concepts qui ne sont pas ceux de la source administrative. *C'est là un sujet majeur d'incompréhension entre les mondes administratif et statistique*. Un effort est alors indispensable pour parvenir à reformuler les anomalies détectées, dans un langage et des concepts compréhensibles par le fournisseur.

Attention cependant, les modifications diverses pouvant intervenir sur une source administrative peuvent rendre certaines parties du processus caduques.

Exemple : les conséquences de la mise en œuvre du prélèvement à la source de l'impôt sur les personnes. Ici, l'entité à l'origine de l'information sur le revenu n'est plus l'individu, mais son employeur. Ce changement du processus administratif a entraîné des évolutions de la couverture de la source, et notamment une meilleure couverture de revenus perçus, qui étaient précédemment non déclarés. Les données statistiques produites par le processus Fideli comportaient cette année-là un million d'individus supplémentaires (!). En effet, le processus de transformation de « contribuables » en « individus résidant en France » n'avait pas été modifié de prime abord.

Le diagnostic de cet écart a été difficile. Le changement du processus administratif conduisait du point de vue statistique à une correction de la sous-couverture (certains revenus précédemment non déclarés par des personnes résidentes en France), mais également à une augmentation de la sur-couverture (revenus de personnes non résidentes en France, comme les saisonniers par exemple).

5.f. Résultat de la phase de transformation

Moins inconfortable que la phase 1, la phase 2 de transformation est plus proche des compétences standard des statisticiens publics, notamment en informatique. Le but de l'opération aura été de mettre au point une succession d'opérations permettant de ramener les données à un cadre connu.

Le résultat de cette phase se décline ainsi :

- les données « brutes[104] statistiques »,
- les métadonnées associées,
- le code de la transformation,
- le résultat de la qualification.

Tout cela peut paraître très exigeant et appelle donc quelques commentaires sur chaque point.

Pourquoi conserver les **données** « statistiques brutes » ? Si l'on se rend compte d'un problème dans les traitements statistiques ultérieurs, il sera toujours possible de modifier ces traitements et de les rejouer sur les données statistiques brutes. Leur conservation est donc décisive pour la réplicabilité des traitements. Par ailleurs, contrairement aux fichiers envoyés par le fournisseur, l'archivage de ces données statistiques brutes est du ressort du service statistique : elles ont le même statut que des données brutes d'enquête.

Les **métadonnées** (définition, caractérisation des domaines, …) auront été progressivement construites dans la mise au point de la transformation. On rappelle juste ici qu'il faut les avoir enregistrées proprement, pour une bonne maîtrise des résultats, mais aussi en vue d'usages ultérieurs, ou dans une perspective de mutualisation.

Autant que possible, la **transformation** devra avoir été codée de façon à ce qu'elle soit réplicable (≈ reproductible)[105], ce qui renvoie à la qualité des développements et à la documentation du code (cf 5.b). C'est d'autant plus sensible que plusieurs choix opérés lors des transformations ne seront pas réversibles. Une transformation sous forme de *pipeline*, entièrement automatisée, est la solution la plus pratique : c'est encore plus évident sur les données privées (Joubert 2025). Mais dans la pratique il arrive que certains passages de la transformation requièrent une intervention humaine et qu'on ait ainsi plusieurs codes, correspondant à des étapes automatisées successives.

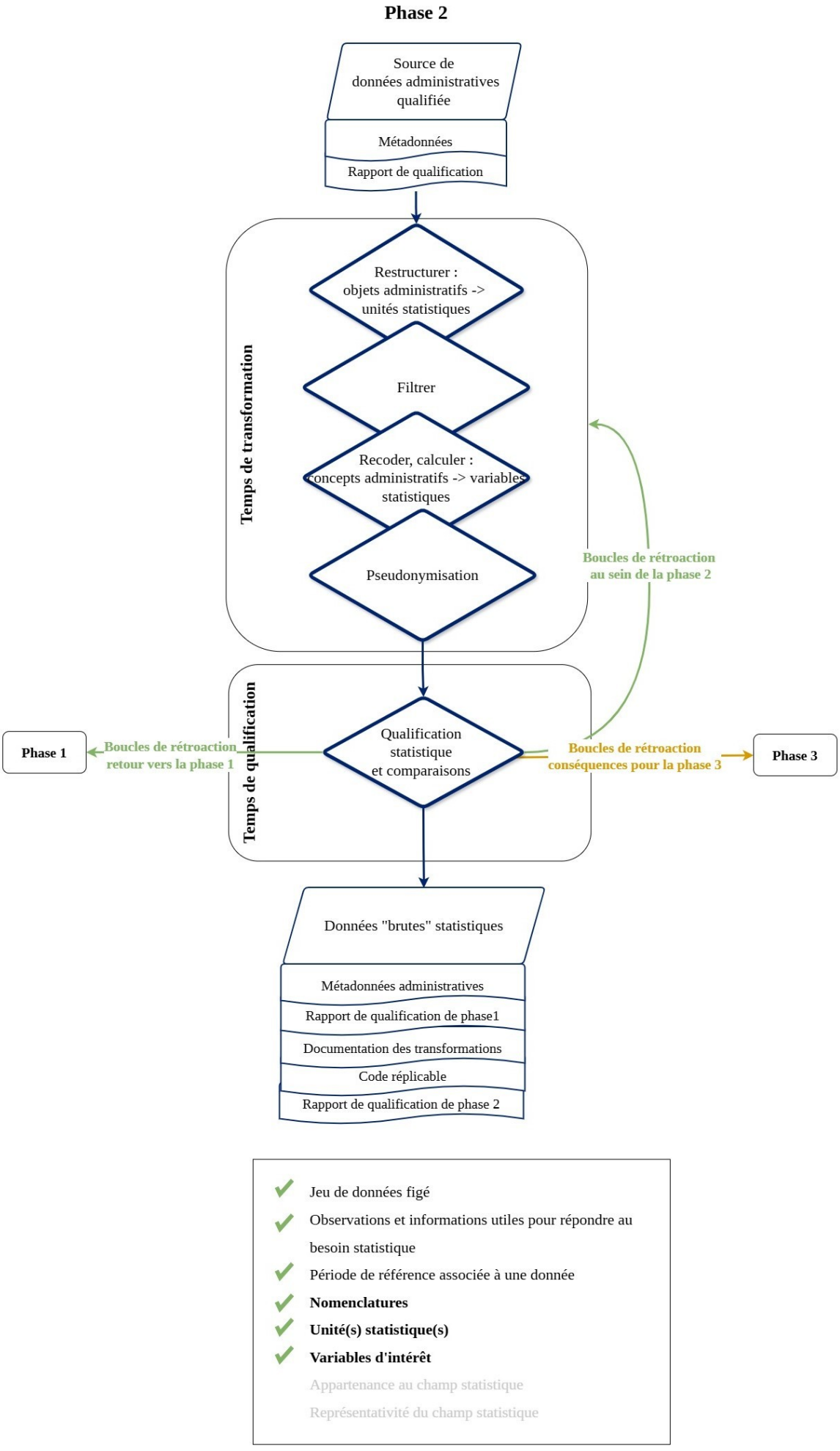
Enfin, les résultats de la **qualification** ne requièrent pas un formalisme excessif, et sont surtout importants pour les enseignements qu'on en tire. Ils permettent de mettre en évidence la réalité des données en fin de phase 2 : incomplétude (sous-couverture), données douteuses, données dont la valeur est « manquant » (avec des règles claires sur la manière de le noter). Ces informations seront décisives pour orienter la phase suivante de traitement statistique, car on saura où l'attention devra être portée.

---

104 On a insisté dans le chapitre 3 sur le problème posé par le terme « données brutes » (Gitelman 2013), en expliquant que ce qui est brut pour un utilisateur peut être très élaboré pour un autre. La signification de l'adjectif est donc contingente. D'où l'importance, ici de parler de données brutes *statistiques*, au sens d'input de traitements statistiques, comme on peut parler de données « brutes d'enquête ». Mais même avec cette précision, il faut reconnaître que cela reste fragile sémantiquement.

105 Les 2 concepts sont distincts. La reproductibilité (= rejouabilité) est la capacité à obtenir les mêmes résultats qu'une étude originale en utilisant mêmes données et même code. C'est très exigeant, adapté à la recherche scientifique. La réplicabilité est la capacité à obtenir des résultats similaires à ceux d'une étude originale en utilisant des données indépendantes et/ou une nouvelle mise en œuvre des méthodes. Dans l'esprit, c'est plus adapté aux usages statistiques : on n'est pas dans une démarche de preuve. Voir à ce sujet (National Academies 2019), cité plus haut.

***Figure 5* - Phase 2 - Transformation de données administratives**



6. La phase de traitement statistique

Cette troisième phase 3 ne diffère pas fondamentalement des traitements effectués sur les enquêtes, i.e. la 5e étape du GSBPM, appelée justement traitement. Donc cette phase sera beaucoup plus courte à décrire. Comme on l'a vu, les deux premières phases conduisent à un point de jonction avec les enquêtes à partir duquel le processus est commun.

6.a. Contexte

Dans cette phase de traitement, on part donc d'un fichier avec données « brutes statistiques » qui est susceptible de comporter des défauts : sous-couverture, valeurs manquantes, valeurs aberrantes ou douteuses, incohérences avec d'autres productions. On va donc appliquer des traitements pour éliminer ou atténuer ces défauts (imputation par exemple), puis produire des statistiques, avec d'éventuelles boucles de rétroaction, comme pour les données d'enquête. C'est tout un travail de mise au point du fichier de données, y compris un « nettoyage » (Chu et al 2016) qui est à réaliser, et sera essentiel pour de futurs usages.

Vis-à-vis des précédentes phases, un élément nouveau apparaît : on ne se situe plus du tout dans un contexte administratif, et en passant à l'univers statistique on bénéficie d'une *infrastructure de connaissances* (Borgman 2015) associée, qui n'est pas nécessairement complète.

Dans le contexte de la statistique publique, cette infrastructure se matérialise entre autres par :

- un référentiel de métadonnées partagé (Bonnans 2019) avec par exemple des définitions de concepts ou des domaines mutualisés ;
- des nomenclatures partagées ;
- des référentiels d'individus, de ménages, de logements avec Resil (Lefebvre 2024 ; Espinasse, Roux 2022), ou d'établissements (Hachid, Leclair 2022 ; Bensoussan et al 2023)
- des méthodes reconnues, documentées, diffusées ; des outils mutualisés ;
- d'autres statistiques de référence déjà publiées, auxquelles on pourra se comparer ;
- des séminaires, colloques, comités, réunissant des experts de la statistique publique ;
- etc.

L'existence de cette infrastructure facilite les choses, fournit un environnement de travail aux statisticiens, mais ne résout pas tout. Un point-clé de la phase 3, c'est **l'existence ou non d'un référentiel de la population considérée**. L'impact sera très important sur toute cette phase.

Si un tel référentiel existe, on pourra en permanence s'y raccorder, comme on le fait pour les enquêtes, où on peut s'appuyer sur des bases de sondage. Cela apporte des garanties de qualité et rend possible des comparaisons, notamment pour mesurer la sous-couverture, ce qui n'était pas faisable (ou alors seulement à gros traits) en phases 1 et 2.

> Ainsi, le système Resil (Lefebvre 2024), qui est justement constitué de données administratives, fonde tout son dispositif sur une « colonne vertébrale » de référentiels : les répertoires d'individus, de logements et de ménages, qui apportent beaucoup de garanties de sécurité. On peut arrimer à ces référentiels des sources importantes élaborées par la statistique publique comme Fideli (Lamarche, Lollivier 2021) pour les individus, ou Filosofi[106] pour les ménages, qui offriront des points de comparaison

106 https://www.insee.fr/fr/metadonnees/source/serie/s1172

Dans le cas des statistiques d'entreprise, c'est le répertoire Sirus (Hachid, Leclair 2022) qui sera cette colonne vertébrale, auxquelles se raccorderont des sources comme la DSN, ce qui permettra entre autres de former une « Base tous salariés ».

Mais on n'a pas toujours le référentiel en question, et ce n'est pas sans conséquences sur la suite.

6.b. Les traitements proprement dits

Les opérations standard à effectuer sont bien connues et documentées par des papiers de méthodologie statistique depuis fort longtemps. Font ainsi partie de cette phase le traitement des non-réponses, partielles ou totales (Caron 2002), le calage sur marges (Deville, Särndal 1992), les démarches visant à gérer la sur-couverture ou la sous-couverture (Wallgren, Wallgren 2007), et le *data editing* (De Waal et al 2011).

Outre l'environnement, la phase 3 a une autre caractéristique distinctive : lors de cette phase, on a la possibilité de **modifier les données pour des raisons statistiques** :

- en s'appuyant sur les modèles qu'on juge pertinents (pour l'imputation notamment),
- mais aussi éventuellement à travers des modifications manuelles effectuées par des gestionnaires.

Ce n'était en effet pas possible en phase 1 (on ne touche pas aux données administratives), ni en phase 2 (les données étaient des résultats de transformation, avant analyse statistique justement).

Dans les traitements, si les méthodes statistiques habituelles s'appliquent, elles vont nécessiter des adaptations dans leur mise en œuvre, qu'il s'agit ici de préciser. On va se limiter au champ, aux données manquantes et à la tabulation, alors que le *data editing* sera vu dans la partie suivante.

- ***Champ***

Il faut distinguer champ statistique et champ administratif. Lors de la phase 2, à travers les étapes de filtrage, on a éliminé une bonne partie de la sur-couverture du champ statistique, mais la sous-couverture vis-à-vis du champ administratif n'est pas observable, car on n'a pas d'éléments de comparaison. Traiter la sous-couverture est donc du ressort de la phase 3.

De manière générale, il est possible de traiter les problèmes de champ[107] (champ statistique différent du champ administratif) et de sous-couverture du champ statistique grâce à des imputations, des compléments apportés par une autre source, voire des pondérations, bien que le recours à la repondération et au calage lors de l'exploitation de données administratives soit encore peu répandu dans la statistique publique.

La façon de procéder va dépendre de l'existence ou non d'un référentiel de population.

107 Ce ne sont pas les mêmes problèmes de champ que dans la phase précédente. En effet, en phase 2, on traitait surtout les questions de sur-couverture, qui conduisent à éliminer potentiellement des informations (ou au moins à calculer une indicatrice d'appartenance au champ). Ici on va plutôt traiter la sous-couverture ou les problèmes d'exhaustivité, en mobilisant d'autres informations.

- *Cas où on dispose d’un référentiel de la population-cible*

Dans ce contexte on n’a pas à se poser la question de la distinction entre champ statistique et administratif. On va pouvoir observer la sous-couverture du champ statistique, en regardant la différence entre population « de référence » de ce champ et la population qu’on trouve dans les fichiers issus de la phase 2. Et on pourra compléter le champ soit avec de l’imputation de non-réponses totales, soit en utilisant d’autres sources s’il en existe.

- *Cas où on ne dispose pas d’un référentiel de la population-cible*

N’ayant pas de l’équivalent d’une base de sondage permettant de connaître la population d'intérêt, on a besoin d'informations sur le champ provenant d’autres sources de données pour évaluer l'appartenance au champ des unités présentes et des unités attendues et absentes. Tout dépend ici de la nature des informations supplémentaires dont on dispose.

Il peut arriver qu’on ait connaissance de marges permettant de faire du calage : on peut alors imaginer de traiter les unités manquantes par repondération, mais c’est une approche à utiliser avec précaution.

Une autre possibilité est de disposer de sources partielles sur le champ d'intérêt, qui peuvent servir d'échantillon de contrôle. Celui-ci peut être conçu pour évaluer la couverture (tout comme les enquêtes post censitaires pour évaluer la couverture des recensements), ou provenir d'une source de données déjà existante. L'utilisation d'échantillons de contrôle diffère du calage évoqué plus haut, et aboutit aux méthodes de capture recapture (Ardilly et Koumarianos 2025).

De façon générale, les exploitations multisources permettent l'utilisation de méthodes pour évaluer la couverture (méthode d'estimation par système multiple ou capture-recapture (Chiperfield et al 2024)) ou pour caractériser l'appartenance au champ des unités, notamment avec la méthode des signes de vie (Óvári, Kočiš 2022, Tiit 2017).

Enfin, on peut n’avoir aucune information sur le champ : on pourra alors croiser des informations pour traiter la sur-couverture.

> Par exemple, on pourra utiliser des informations sur les personnes décédées dans le RNIPP pour nettoyer le fichier des permis de conduire, mais il s'agit d'un usage administratif.
>
> Autre exemple, plus compliqué qui a trait aux deux difficultés (sur couverture et sous couverture) : la détermination d'une population résidente à partir de données administratives repose sur l'exhaustivité de ces sources, propriété souvent recherchée, mais rarement atteinte. Afin de construire des données fiables, les statisticiens vont chercher des signes de présence, positifs, ou négatifs. Certains pays souhaitent mobiliser des informations telles que l'inscription d'enfants dans le système scolaire, le non-renouvellement d'un permis de conduire, ou encore la consommation de soins médicaux. C'est la conjonction de plusieurs informations administratives qui définit ainsi la probabilité d'appartenance au champ d'intérêt.

Une difficulté courante est de caractériser les unités absentes des données administratives : il n'y a pas d'équivalence entre la présence de données relatives à une unité, et son existence réelle. La disparition réelle d'une unité (ex : cessation d'activité d'une entreprise) peut être enregistrée

tardivement dans les SI administratifs, mais l'absence d'éléments comptables peut constituer un signal plus précoce de sa disparition. A l'inverse, la présence d'une unité dans un fichier administratif de type répertoire peut être ancienne, et ne signifie pas que l'unité soit toujours active.

- ***Données manquantes***

La partie précédente sur la sous-couverture renvoyait de fait à des questions de traitement de données manquantes, en l'occurrence d'unités manquantes (on peut les rapprocher des non-réponses totales, bien que l'absence de base de sondage rende les deux difficultés - non réponses versus unités manquantes - très différentes). Dans ce qui suit on évoque surtout les données manquantes au sein d'unités existantes, donc l'équivalent des non-réponses partielles, mais plusieurs réflexions s'appliquent à l'absence d'unités.

Pour les traiter efficacement, il faut d'abord savoir de quelle manière se répartissent les données manquantes, et c'est là un des résultats de la qualification de fin de phase 2.

A ce sujet, (Little et Rubin 2002) décrivent trois types de répartition des données manquantes (univariée, monotone, sans structure). Ils formalisent les différents mécanismes qui conduisent à des données manquantes, qu'ils distinguent en trois catégories : les données manquantes complètement aléatoirement (MCAR), les données manquantes aléatoirement (MAR), et les données manquantes non aléatoirement (NMAR)[108]. Leur typologie n'est pas liée aux modalités d'obtention des données et peut donc s'appliquer aussi bien aux enquêtes qu'aux données administratives (qui seront plus concernées par MAR et NMAR). Dans le cas des enquêtes, ces mécanismes font appel à la *propension à répondre* … qui n'a pas de sens dans le cas des données administratives, puisqu'il n'y a pas de réponse à une question[109].

Pour les données administratives, la question du biais lié à l'absence de données, ou d'unités statistiques, va se poser : cette fois, comprendre l'origine des absences (d'unités ou de données) nécessite de revenir à la logique des processus de gestion administratifs. Pour cela, il sera utile de produire quelques statistiques exploratoires sur la répartition des données manquantes dans la population, selon les différentes variables.

On traitera alors les données manquantes par des méthodes d'imputation adaptées à la configuration rencontrée : (Imbert, Vialaneix 2018) offrent un large tour d'horizon des configurations en matière de données manquantes et de méthodes appropriées, qui peut tout à fait s'appliquer aux données administratives. Bien entendu si on dispose d'une autre source statistique permettant d'avoir un proxy pour ces données, c'est encore mieux.

- ***Données douteuses***

La phase 2 a permis, sur la base d'une batterie de contrôles automatiques de cohérence / vraisemblance, de repérer des données douteuses : en l'occurrence, non pas des statistiques sujettes à caution, mais bien des micro-données. La phase 3 elle-même, via le *data editing*, produit aussi ce type d'information.

108 MAR = Missing At Random, MCAR = Missing Completely At Random, NMAR = Not Missing At Random. L'existence d'une corrélation entre absence des informations et variables d'intérêt (NMAR) peut conduire à des estimations biaisées, quel que soit l'origine des données (enquête ou données administratives).

109 Sauf peut-être dans le cas des déclarations administratives, et encore.

Mais les données ainsi jugées anormales peuvent se révéler tout à fait exactes.

> Par exemple, il arrive qu'une entreprise ait un chiffre d'affaires très élevé avec un effectif salarié nul ou quasi. C'est le cas notamment lorsqu'elle s'appuie essentiellement sur un effectif prêté.

Comment les traiter ? Pour chaque donnée douteuse, on ainsi le choix entre plusieurs possibilités :

- laisser la donnée telle quelle,
- remplacer la donnée par une donnée imputée,
- remplacer la donnée par son équivalent dans une autre source, si cela existe.

Le choix de laisser la donnée en l'état peut être fait sur la base d'un critère automatique, avec une sorte de degré de gravité de l'incohérence au-delà duquel on considère qu'on ne peut la garder. Ou alors, si le nombre de cas n'est pas trop élevé, on peut confier à des gestionnaires le soin d'effectuer les vérifications à la main, pour décider de l'imputation ou non de la donnée. La grande différence avec les enquêtes, c'est qu'on ne pourra pas aller vérifier à la source.

Pour l'imputation elle-même, la situation est similaire à celle des données manquantes, et ne diffère pas de ce qui se fait sur une enquête. On pourra avec profit utiliser des méthodes de redressement qui permettent d'assurer le respect des différents contrôles : ce sont les méthodes issues de l'article fondateur de (Fellegi, Holt 1976)[110].

- ***Tabulation***

Il s'agit du dernier « traitement » : on n'est plus en train de mettre au point un fichier de données individuelles, mais on calcule les différents agrégats statistiques à partir de là. La tabulation ne présente pas de spécificités par rapport aux enquêtes. Simplement, on n'a pas de poids de sondage, et si l'on ne dispose pas d'une population de référence le risque d'erreur est très important. La boucle de rétroaction jouera donc un rôle majeur.

6.c. Qualification et rétroaction : l'importance de l'*output editing*

Le point essentiel de la qualification en phase 3, c'est la vérification de vraisemblance des agrégats statistiques. Techniquement, c'est assez proche de ce qu'on a vu en phase 2 : cohérence avec des totaux déjà connus, cohérence des évolutions de totaux, vérification de complétude du champ, vérification de la plausibilité de quelques distributions. Mais alors qu'en phase 2, il s'agissait essentiellement de faire quelques vérifications statistiques élémentaires pour voir s'il n'y avait pas d'erreur de champ, notamment, en phase 3 l'objet va être de vérifier les statistiques (agrégats, distributions) sur les variables d'intérêt, dans une optique de validation.

Il s'agit donc ici des vérifications finales de statistiques (et non de micro-données) : dans la démarche de *data editing*, déjà évoquée pour les enquêtes, il s'agit de la partie *output editing* (Arbués et al 2013). Cela va consister en premier lieu à explorer les différentes statistiques existantes sur des sujets proches, à regarder les évolutions, pour voir si les ordres de grandeur sont

110 Méthode qui, presque 50 ans après, reste une référence majeure lorsqu'il s'agit de réaliser des redressements *de sorte que les règles de contrôle soient vérifiées*. La méthode minimise le nombre de redressements pour cela.

bons. Si des incohérences subsistent, on devra comprendre pourquoi : dans ce cas, la boucle de rétroaction devra remonter aux phases précédentes, à la phase 2 voire à la phase 1, souvent pour constater des incompréhensions sur le champ administratif. Et ce qui rend les choses très difficiles par rapport aux enquêtes, c'est que ces statistiques ne sont déjà plus dans le langage de l'entité administrative : les statisticiens devront donc « se débrouiller » tout seuls avec les éléments dont ils disposent.

Si l'on analyse les étapes de qualification, en phases 1 2 et 3 respectivement, il apparaît que l'*output editing* reste la vérification ultime, la plus importante : on peut imaginer d'alléger le travail de vérification dans des étapes précédentes, mais celle-là est incontournable. C'est en effet « le » filet de sécurité dont dispose la statistique publique avant de diffuser quoi que ce soit.

On constate aussi que la qualification de fin de phase 2 a beaucoup de points communs avec celle de la phase 3. L'avantage de qualifier en fin de phase 2, c'est que cela permet de valider la transformation. Mais il faut reconnaître que sur les trois qualifications, s'il y en a une qu'on peut à la rigueur omettre, c'est celle de la phase 2.

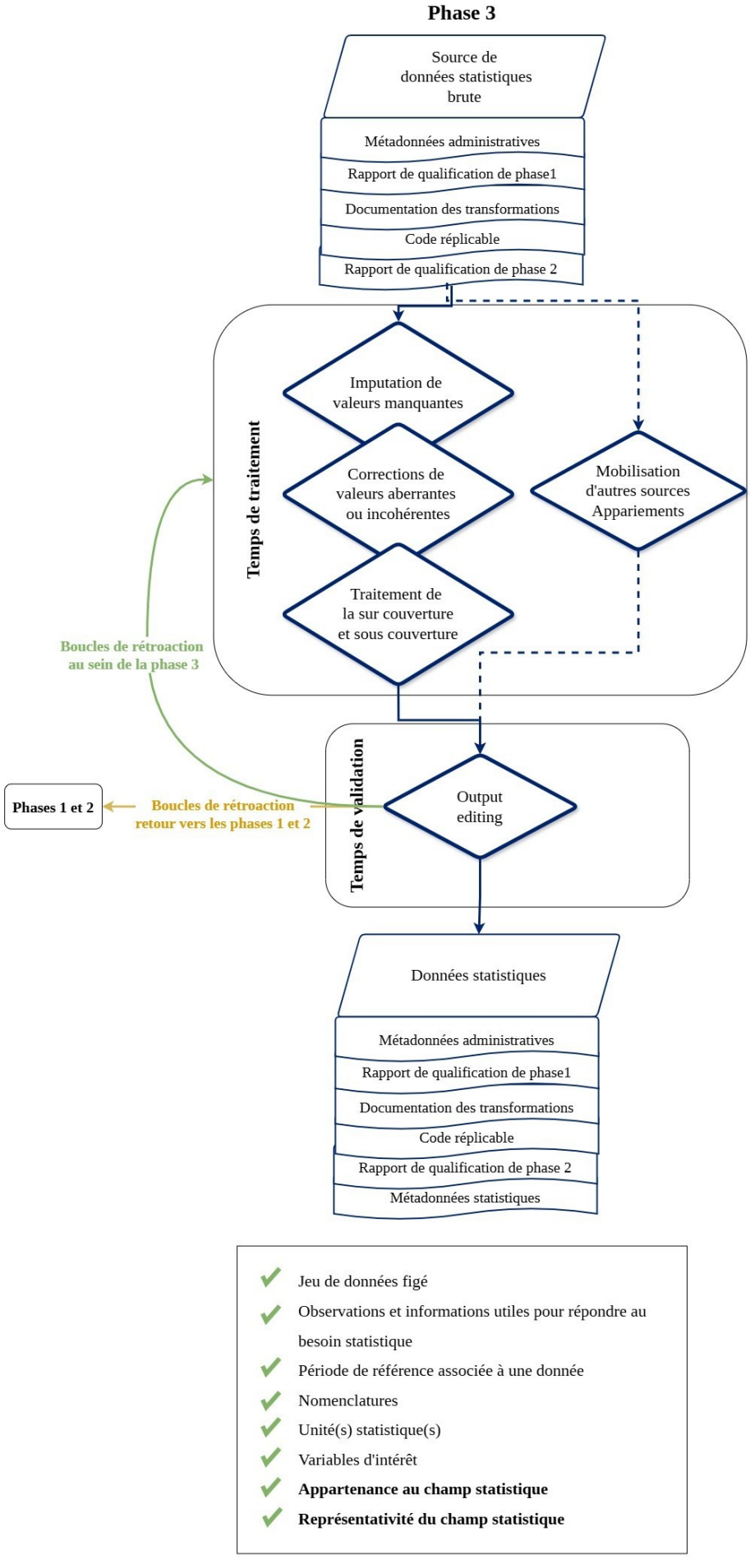
6.d. Résultat de la phase de traitement statistique

A l'issue de la phase de traitement statistique, on a donc :

- Un fichier de données « plein », i.e. qui ne contient plus de données manquantes et qui est cohérent par rapport au champ statistique
- des tableaux statistiques
- des résultats de qualifications

C'est en général sur ce troisième aspect que les choses pêchent un peu : on ne garde pas nécessairement trace des vérifications faites, or ce peut être une information utile si la source doit être qualifiée plusieurs fois, ou s'il y a plusieurs types d'utilisateurs.

Figure 6 **- Phase 3 - Traitement des données statistiques**



C. Synthèse

7. Synthèse et mise en perspective

En janvier 2024, Peter Christen et Rainer Schnell ont publié un court article pointant les multiples incompréhensions autour de l'acquisition, du traitement et de l'appariement des données. Les auteurs vont jusqu'à présenter des listes d'idées reçues sur chacune de ces opérations. On sent affleurer, au fil des pages, une forme d'agacement vis-à-vis de la persistance des erreurs commises par les *data scientists*, dues à une vision très simpliste de ce que sont réellement les données. Et si leur travail porte sur la *data science* en général, il s'applique pleinement aux statistiques.

L'utilisation de plus en plus systématique de nouvelles sources de données est désormais un horizon naturel pour la statistique officielle : *datafication* du monde (Rey 2016) et multiplicité des sources de données, capacités accrues d'accès et de stockage, de puissance de calcul et d'analyse, tout plaide pour une statistique de plus en plus *data-driven* (Salemink et al 2020). Mais comme le soulignent ces auteurs, les défis sont nombreux.

Dans le présent document, nous nous plaçons délibérément dans un contexte restreint : l'utilisation des données administratives « seulement », pour reprendre la terminologie de (Rancourt 2018), et en se focalisant sur la méthodologie de production des statistiques, incluant l'évaluation de leur qualité (Laitila et al 2011 ; Eurostat 2016a), mais omettant les usages indirects[111], ou dans un contexte multisources. Et l'on se retrouve confrontés, comme Christen et Schnell, à l'imperfection fondamentale inhérente à ces données, observée avec la focale de l'utilisateur statisticien. Il faut donc imaginer des dispositifs d'acquisition de données administratives à des fins statistiques (FAO 2018), capables de supporter cette imperfection endémique, ainsi que la distance entre les producteurs de données administratives et les utilisateurs (Borgman et al 2025)... qui sont eux-mêmes producteurs de données statistiques.

7.a. Résumé de la démarche proposée

Pour réaliser de tels dispositifs, le présent document propose un cadre général qui se présente sous forme d'une grille d'analyse, et d'une démarche en trois phases, chaque phase ayant des enjeux qui lui sont propres.

On y développe une vision conceptuelle que l'on pourrait imager de la manière suivante.

La phase 1 peut être vue comme une projection : on part d'un espace « administratif » ayant de très nombreuses dimensions (le système d'information), l'une de ces dimensions étant le temps. Et le but de cette phase est de projeter les données sur un sous-espace « administratif » plus petit (la source administrative), en figeant la dimension temps. L'objet de l'opération serait ainsi de prendre en compte l'immensité des possibles en se ramenant à ce qui est nécessaire, tout en homogénéisant et en fixant les temporalités.

Mais on ne peut se limiter à cette image : l'idée est aussi et surtout de tenir compte de la méconnaissance de tout l'écosystème sous-jacent aux données administratives. En réalité, le sujet principal de la première phase, c'est de gérer la *data friction*, au sens de (Edwards et al 2011). Cela passe par une capacité à s'adapter à une culture « étrangère », par une immersion dans cet univers,

111 De Waal et alii (Eurostat 2016 b) listent 6 usages, regroupés en usages directs : tabulation directe, ou utilisation en substitution ou complément d'une autre source de données ; et usages indirects (création et maintenance de référentiels, *data editing* et imputation, estimation indirecte, validation de données). Toutefois, quelque soit l'usage envisagé, le statisticien est confronté à différents problèmes d'intégration.

par une compréhension fine des tenants et aboutissants des processus administratifs, préalable à la mise en place progressive d'une convention avec l'organisme fournisseur de données. Et l'on découvre peu à peu que rien ne va de soi, qu'il faut se défaire d'une vision naïve et simpliste, que l'on pourrait résumer à : « transmettez-nous les données, on saura se débrouiller. »

La phase 1 fait donc la part belle à un large temps de préparation, avant de passer à la réception-acquisition proprement dite. On lui associe une grille d'analyse pour faciliter le travail. Car les données ne sont pas « en l'air », elles s'écrivent selon une certaine grammaire : leur champ, leurs objets, leurs variables, leurs domaines, leur temporalité.

En poursuivant une analogie mathématique, **la phase 2 est en quelque sorte un morphisme** : on part du sous-espace administratif E, à temporalité définie, et dont on a réduit les dimensions. Et on le transporte dans un autre espace, en faisant transiter la structure associée aux données (objet, champ, variable, domaine, temporalité). Cela revient, dans l'esprit, à construire une fonction f telle que F = f(E), une fonction capable aussi de transporter la grille d'analyse.

Ce que nous dit cette 2e phase, en filigrane, c'est qu'en réalité on ne pourra pas travailler directement avec les données administratives telles quelles, car leur grammaire et leur sémantique sont trop éloignées des besoins d'analyse de la statistique publique. Il faut donc absolument, une fois acquises et qualifiées, les sortir du monde administratif (détachement), puis les conditionner, les transformer afin de les arrimer solidement au monde statistique. Avec là aussi une grille d'analyse propre à cet usage spécifique. La transformation, i.e. notre fonction f, est ici un objet en soi, à documenter, maintenir, versionner.

Il s'agit d'une phase intermédiaire, hybride, entre mondes administratif et statistique, une phase pivot pour passer à la suite.

La phase 3, c'est l'activité statistique proprement dite, phase dans laquelle la profession statistique retrouve tous ses repères. Elle se situe intégralement dans un environnement statistique, en adopte les codes, les méthodes, le langage.

C'est la première phase pour laquelle on s'autorise à modifier des données, aussi bien manuellement que via des modèles, pour qu'elles respectent un certain nombre de propriétés (par exemple la complétude, ou l'absence de valeurs manquantes). On y retrouve un ensemble de traitements statistiques tout à fait classiques sur les microdonnées, jusqu'à la tabulation, i.e. la production des statistiques proprement dites, qui marque la fin du processus.

La 3e phase ne se soucie donc plus de *data friction* … ou presque : car elle est l'occasion de constater des incohérences, qui obligent à revenir en arrière.

Ceci nous conduit à souligner un autre aspect de la démarche proposée : la présence de « **boucles de rétroaction** » dans chaque phase. Rien de théorique et d'abstrait ici : dans la phase d'acquisition, par exemple, cela veut simplement dire que si le service statistique n'est pas satisfait des données transmises, il doit être capable de l'objectiver à travers une qualification automatisée, pour légitimer la demande d'une nouvelle livraison de données ou métadonnées, ou tout simplement d'explications supplémentaires. Dans la phase 2 de transformation, la rétroaction portera sur le *pipeline* de transformation, alors que dans la phase 3, il s'agira du processus classique de *data editing*, avec ses itérations finales, et d'éventuels retours en arrière.

Les 3 phases ont inévitablement des **liens entre elles**, notamment à travers les boucles de rétroaction, mais chacune donne lieu à des « états » différents des données, dans l'univers administratif pour la phase 1, dans l'univers statistique pour les phases 2 et 3.

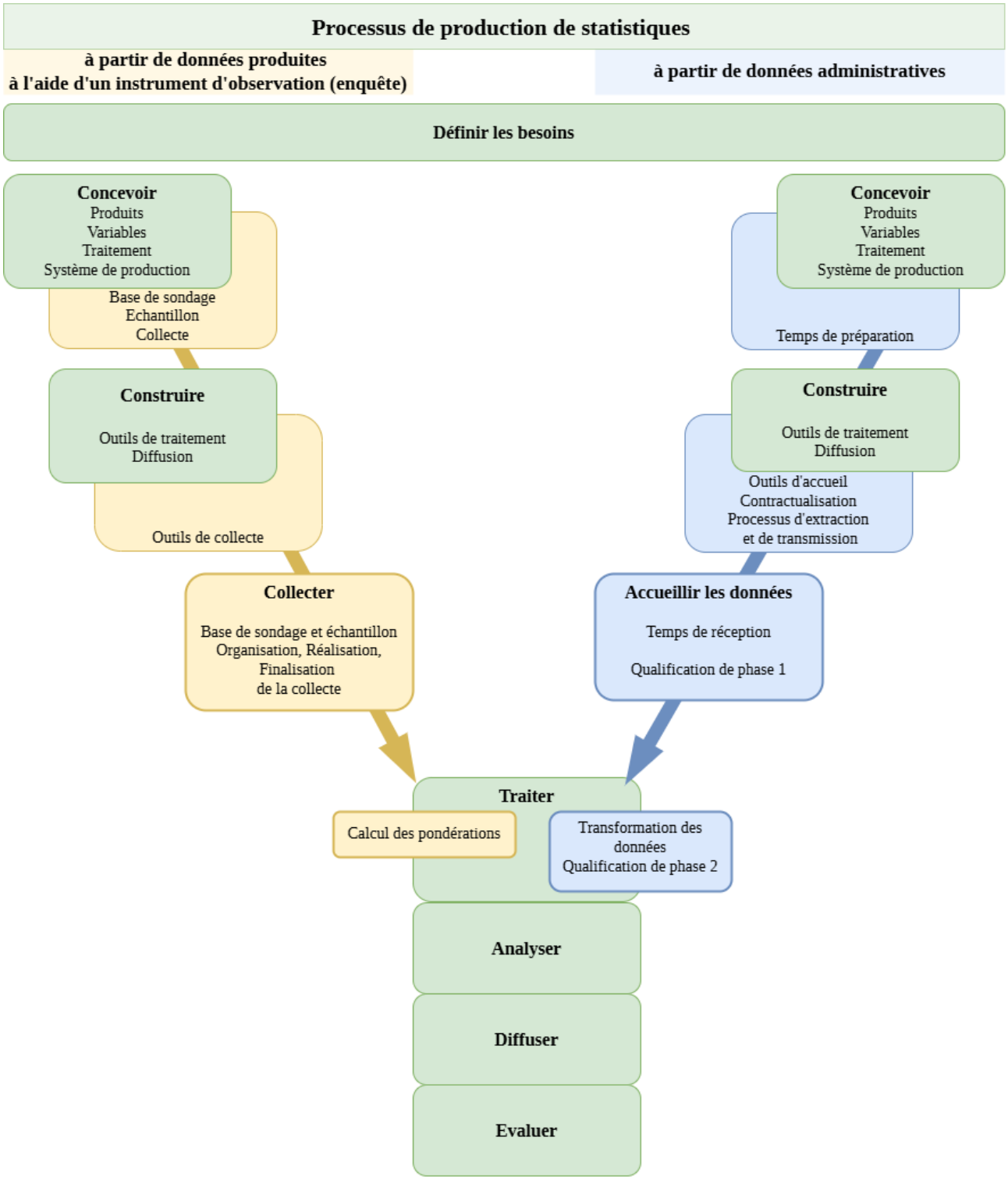
Et chaque phase fait avancer les choses, contribue à résoudre des problèmes, avec ses propres apports, ce que l'on peut résumer ainsi :

***Figure 7* - Passage de la grille d'analyse administrative à la grille d'analyse statistique**

Ce qui existe dans le monde administratif		Ce dont on a besoin dans le monde statistique	Phase 1	Phase 2	Phase 3
Système d'information vivant		Jeu de données figé	✓		
Système d'information très riche		Observations et informations utiles pour répondre au besoin statistique	✓		
Pas de période de collecte Pas de période de référence	→	Période de référence associée à une donnée	✓		
Listes de codes administratifs		Nomenclatures		✓	
Objets administratifs		Unité(s) statistique(s)		✓	
Concepts administratifs		Variables d'intérêt		✓	
Champ réel administratif		Appartenance au champ statistique		✓	
		Représentativité du champ statistique			✓

Les 3 phases permettent aussi de compléter le schéma du « Y », présenté à la fin du chapitre 2 et qui était délibérément vierge dans sa branche liée aux données administratives. On trouvera ce schéma page suivante.

Le schéma du « Y » complété



Ce schéma reprend celui de la page 24, en mettant cette fois ci en évidence les sous-phases liées à l'utilisation de données issues de sources administratives.

La phase Traiter contient différentes étapes évoquées dans les parties 5 et 6 : certaines présentent des spécificités lorsque les données sont issues de sources administratives (par exemple la qualification de fin de phase 2), alors que d'autres sont similaires (calcul des agrégats...). Les sous-phases sont détaillées dans le GSBPM (Unece, 2019).

7.b. Mise en œuvre

Oui, mais tout cela n'est-il pas un peu trop théorique, idéaliste ? En pratique, est-il toujours possible d'effectuer ces décompositions proprement ?

Ce qui peut se révéler impossible, dans certains cas, c'est le principe de boucle de rétroaction avec l'administration : il se peut que celle-ci ne veuille pas d'échanges supplémentaires, ou ne soit pas en mesure de prendre en compte les demandes statistiques. L'administration poursuit en effet un but opérationnel, et la production de statistiques est un usage secondaire par un tiers des données existantes.

La démarche en trois phases, avec véritable séparation entre elles, peut aussi s'avérer très difficile à atteindre opérationnellement : difficultés de stockage de très gros volume, temps de travail requis, ... Cela conduit souvent en pratique à fusionner des étapes. Les phases 1 2 et 3 donnent un cadre, et en pratique **l'idée est surtout de savoir, lorsqu'on s'écarte du cadre, pourquoi on le fait et quelles en sont les conséquences**. Essayons notamment d'analyser l'impact d'une non séparation entre phase d'acquisition et phase de transformation, ou entre phase de transformation et phase de traitement.

Il faut distinguer pour cela deux situations : la primo acquisition, et le cas où la source administrative est utilisée à plusieurs reprises. C'est dans la première situation que le découpage effectif des phases est plus important.

Primo acquisition

En phase 1 le temps de préparation sous diverses formes est incontournable : l'acculturation à un certain univers administratif, la documentation, les métadonnées, une première forme de contractualisation. La séparation entre qualification de phase 1 (et boucle de rétroaction) et phase 2 est essentielle : si on ne le fait pas là, on le paye après, car avec des données transformées il sera beaucoup plus difficile d'interagir avec l'entité administrative, et donc de demander au fournisseur des fichiers de meilleure qualité[112]. D'autre part, si la qualification se fait en fin de phase 2, c'est plus difficile de démêler les problèmes de la source et les problèmes liés à la transformation. Globalement cette qualification est très importante, avec une attention particulière portée aux unités (objets) et variables manquantes, et aux variables aberrantes : ainsi, pour la nouvelle source GMBI, la comparaison avec les données issues du cadastre s'est révélée indispensable.

Si on isole la phase 2, on peut imaginer de la fusionner avec la phase 3 mais en faisant cela on mélange dans un seul et même *pipeline* les traitements "non statistiques" (la transformation)[113] et les traitements statistiques. On ne fait donc pas de vérification élémentaire qui permettrait éventuellement de reboucler, et on ne fait pas d'« arrêt sur image » pour bien comprendre ce dont on dispose avant d'effectuer les traitements statistiques.

Cas d'acquisition régulière de données

En phase 1, à partir de la deuxième acquisition, on pourrait considérer que le temps de préparation est fait... sauf que le contexte peut avoir changé : nouvelles variables, nouveaux domaines,

112 A la rigueur, c'est moins problématique lorsque la « distance » entre mondes administratif et statistique est faible.

113 Non statistiques : au sens où ces traitements n'appartiennent pas aux méthodes statistiques traditionnelles, et où l'idée de transformation pourrait s'appliquer à d'autres domaines.

évolution de champ. Or de telles évolutions sont majeures. On peut certes dire que l'acculturation est faite suite à la première acquisition, mais il reste la mise au point et les éventuels avenants de de convention, plus les changements dans les données et métadonnées, que l'on peut décrire selon la grille (champ / variable / domaine). En primo acquisition, la réception-qualification aura idéalement été outillée, documentée. Le travail de mise à jour ne doit pas être négligé, mais il est vrai que la charge est probablement moindre.

On peut donc imaginer d'intégrer la phase 2 à la phase 1 en ayant une réception de données qualifiées puis transformées, ou transformées puis qualifiées. Mais attention, il faut absolument qu'on conserve la possibilité d'une rétroaction avec l'administration, ce qui est plus complexe une fois les données transformées. Par ailleurs, on peut considérer qu'en qualification, ce sera un peu moins difficile de démêler des problèmes dus à la source et ceux qui sont dus à la transformation.
L'évaluation des volumes reçus reste importante, avant transformation.

Si l'on fusionne phase 1 et phase 2 on aboutira ainsi à une large phase de réception – qualification – transformation, puis dans un deuxième temps une phase de traitement statistique.

Une autre possibilité est d'intégrer la phase 2 à la phase 3, ce qui conduit à reporter toute la qualification de nature statistique à l'*output editing*. Attention, la qualification va être essentielle en matière d'évolution, l'évolution du champ d'une période à l'autre étant un sujet majeur sur lequel se concentrent la majorité des risques.

Si on fusionne phase 2 et 3, on aura donc d'une part une phase de réception qualification, d'autre part une phase de transformation – traitement, aussi automatisée que possible. La clé reste de pouvoir repérer les grosses erreurs, les énormités, et d'effectuer ensuite le bouclage. De manière générale, si on décide de minimiser les opérations de qualification, on risque de tout renvoyer à l'étape finale et indispensable d'*output editing*, qui est le « filet » restant, sans avoir suffisamment de leviers pour intervenir.

Cas d'acquisitions mutualisées

Si l'on essaie maintenant d'aller un peu au-delà du contexte initial (une source administrative[114] et des statistiques fondées sur celle-ci), le premier sujet qui vient à l'esprit est celui de la **mutualisation** : pour une source donnée, si on a défini un cadre propre, rigoureux pour un usage, il serait dommage de ne pas le réutiliser, au moins en partie, pour d'autres usages, au minimum via le partage de la documentation, des métadonnées, de la compréhension du sujet. En fonction des usages, les mutualisations peuvent intervenir à différents niveaux (en fin de phase 1, de phase 2 ou de phase 3), ou conduire à l'élaboration d'une phase 2 la plus ouverte possible. Celle-ci doit être également bien documentée afin de transmettre à d'autres utilisateurs potentiels une bonne connaissance des transformations effectuées.

114 Ou un ensemble cohérent de plusieurs sources administratives (cf. les données du cadastre, par exemple).

7.c. Prolongements

Bien connaître une source, c'est aussi se donner la possibilité de faire du multi-sources, or on a limité le propos, par souci de simplicité, à du mono-source. Or il est fréquent d'utiliser plusieurs sources, qu'elles soient des sources de données administratives ou des enquêtes, au sein d'un même processus. Les statistiques multi-sources soulèvent différents enjeux non exposés ici. Ainsi, pour (De Waal et al, 2020), les problèmes d'intégration de données se ramènent souvent à trois difficultés principales : harmonisation des unités et des concepts, sélection des valeurs pertinentes, et appariements. Il faudrait parler aussi des méthodes d'estimation utilisant simultanément différentes sources (estimations composites).

La question de la précision (erreur quadratique moyenne, variance) n'a pas du tout été abordée, or c'est un sujet qu'on ne peut laisser de côté. Il existe relativement peu de travaux autour de l'estimation de *total survey error*, hors d'un contexte d'enquête, mais on peut en trouver quelques-uns (Groves, Lyberg 2010). Cela suppose un travail de modélisation des différents types d'erreur en n'ayant désormais plus, au centre, l'erreur d'échantillonnage.

Au-delà des pures questions de méthodes de calcul et d'outils, l'utilisation de données administratives ne peut faire l'impasse sur les questions de confidentialité, d'éthique, d'acceptabilité : ces données ne sont pas uniquement un *input* qu'un processus industrialisé doit traiter. Elles ont aussi une sensibilité particulière parce qu'elles concernent des individus, qui ne souhaitent pas nécessairement qu'elles soient ré-utilisées, et à qui on n'a pas nécessairement demandé leur avis, contrairement à ce qui se passe, mécaniquement, pour une enquête.

Enfin, un prolongement possible de ces travaux serait de les appliquer aux données privées. Plusieurs exemples ont été donnés tout au long du document, et à vrai dire les différences entre production de statistiques à partir de données administratives ou de données privées ne sont pas considérables. Mais il en existe : (Lesur 2025) pointe notamment les problèmes d'exhaustivité, de représentativité, les possibilités de comptage multiple (entre opérateurs, par exemple), la complexité accrue du cadre juridique et le fait que les données ne soient pas gratuites[115]. Dans la mesure où il est en général difficile ou impossible d'accéder au système d'information, on met souvent en place des *pipelines* qui effectuent des transformations et même une première agrégation : les statisticiens n'accèdent alors qu'au résultat de ce *pipeline* (Joubert 2025). Dans ce cas on mêle phase 1 et phase 2, et la boucle de rétroaction avec le fournisseur ne va pas de soi.

115 Depuis 2016, la loi n° 2016-1321 du 7 octobre 2016 pour une République numérique pose le principe de transmission obligatoire et gratuite de données publiques entre administrations.

Bibliographie

AMOSSE, Thomas, 2020. La nomenclature socioprofessionnelle 2020 - Continuité et innovation, pour des usages renforcés. *Courrier des statistiques* N° N4, juin 2020.

ANXIONNAZ, Isabelle et MAUREL, Françoise, 2021. Le Conseil national de l'information statistique : la qualité des statistiques publiques passe aussi par la concertation. *Courrier des statistiques* N° N6, juillet 2021.

ANDRE Mathias, ARNOLD, Céline, et MESLIN Olivier, 2021. 24 % des ménages détiennent 68 % des logements possédés par des particuliers. France Portrait Social 2021, Insee.

ANDRE Mathias et MESLIN Olivier, 2022. Patrimoine immobilier des ménages : Enseignements d'une exploitation de sources administratives exhaustives. In : Courrier des statistiques N° N7, janvier 2022.

ARBUÉS, Ignacio, REVILLA, Pedro, et SALGADO, David, 2013. An Optimization Approach to Selective Editing. *Journal of official statistics* Vol. 29, No. 4, 2013, pp. 489–510.

ARDILLY, Pascal, 2006. Les techniques de sondage. Nouvelle éd. actualisée et augmentée. Paris : *Éd. Technip*. ISBN 978-2-7108-0847-3.

ARDILLY, Pascal, KOUMARIANOS, Heidi, 2025. Les estimations par capture-recapture ou par système multiple : quelques éléments théoriques. Insee, Document de travail N° M2025-01, Janvier 2025.

d'AURIZIO Leandro et PAPADIA Giuseppina, 2019. Using Administrative Data to Evaluate Sampling Bias in a Business Panel Survey. *Journal of Official Statistics*, Vol. 35, No. 1, 2019, pp. 67–92.

BABB, Penny, 2017. The assurance of administrative data: A proportionate approach. Statistical Journal of the IAOS 33 (2017) 435–440. DOI 10.3233/SJI-160321

BACH, Laurent, BOZIO, Antoine, DUTRONC-POSTEL, Paul, FIZE, Étienne, GUILLOUZOUIC, Arthur et MALGOUYRES, Clément, 2023. Évaluation de la réforme de la taxe d'habitation. Rapport IPP n°48 – Décembre 2023

BAKKER B., van ROOIJEN J., et van TOOR L., 2014. The System of social statistical datasets of Statistics Netherlands : An integral approach to the production of register-based social statistics, *Statistical Journal of the IAOS* 30 (2014) 411–424

BATES, J., 2017. The politics of data friction. *Journal of Documentation*. ISSN 0022-0418

BATINI, Carlo et SCANNAPIECO, Monica, 2016. Data and Information Quality – Dimensions, Principles and Techniques. *Springer.* ISBN 978-3-319-24104-3.

BECK, François, CASTELL, Laura, LEGLEYE, Stéphane et SCHREIBER, Amandine, 2022. Le multimode dans les enquêtes auprès des ménages : une collecte modernisée, un processus complexifié, *Courrier des statistiques* N° N7, janvier 2022.

BENICHOU, Yves-Laurent, ESPINASSE, Lionel, et GILLES, Séverine, 2023. Le code statistique non signifiant (CSNS) : un service pour faciliter les appariements de fichiers, *Courrier des statistiques* N° N9, juin 2023.

BENSOUSSAN, Johanna, BIZINGRE, Joël, et COURVALIN, Nathalie, 2023. FINESS, le répertoire des établissements de santé, *Courrier des statistiques* N° N10, décembre 2023.

BERTHE, Juliette, 2025. Le défi des données pour l'inspection générale des affaires sociales, *Courrier des statistiques* N° N13, juin 2025

BERTHELOT, Gregory, 2020. Premier bilan du prélèvement à la source après un semestre de mise en œuvre : une entrée en réforme réussie, *Revue française de finances publiques* 2020/1 N°149.

BERZOFSKY Marcus E., LIAO Dan, BARNETT-RYAN Cynthia, SMITH Erica L., 2025. A total data quality paradigm for official statistics based on administrative data. *Statistical Journal of the IAOS*. 2025;. doi:10.1177/18747655251337671

BIEMER, Paul P., de LEEUW, Edith, ECKMAN, Stephanie, EDWARDS, Brad, KREUTER, Frauke, LYBERG, Lars E., TUCKER, N. Clyde, WEST, Brady T. (Eds), 2017. Total survey error in practice. Wiley, 2017.

BOLLIET, Quentin, FLOYRAC, Aymeric, MAILLARD, Sophie, et ROSENZWEIG, Agathe, 2025. L'inspection générale des finances et la science des données - Quelles méthodes pour quels usages ?, *Courrier des statistiques* N° N13, juin 2025

BONNANS, Dominique, 2019. RMéS, le référentiel de métadonnées statistiques de l'Insee, *Courrier des statistiques* N° N2, 27 juin 2019.

BONNET Odran, et LOISEL Tristan, 2024. L'économie racontée par les données bancaires, *Courrier des statistiques* N° N12, décembre 2024

BONTEMS Thierry, et GOULIN, Sabine, 2010. De la Qualité de l'Information Administrative et Scientifique. 8. *Ecole Inter-organismes "Qualité en Recherche et en Enseignement Supérieur", Montpellier,* Sep 2010.

BORGMAN, Christine L., 2015. Big data, little data, no data : scholarship in the networked world. *The MIT Press*, 2015.

BORGMAN, C. L., DARCH, P. T., SANDS, A. E., et GOLSHAN, M. S. , 2016. The Durability and Fragility of Knowledge Infrastructures: Lessons Learned from Astronomy. In *Proceedings of the 79th ASIS&T Annual Meeting* (Vol. 53). Copenhagen: ASIS&T. https://www.asist.org/files/meetings/am16/proceedings/submissions/papers/31paper.pdf

BORGMAN, Christine L., GROTH Paul T., 2025. From data creator to data reuser : distance matters. *Harvard Data science Review.*

BOWKER, Geoffrey C. et STAR, Susan Leigh, 2000. Sorting things out. Classification and its consequences. 25 août 2000. *The MIT Press*. ISBN 978-0262522953

BOYDENS, Isabelle, 1999. Informatique, normes et temps – Évaluer et améliorer la qualité de l'information : les enseignements d'une approche herméneutique appliquée à la base de données «LATG» de l'O.N.S.S. *Éditions E. Bruylant.* ISBN 2-8027-1268-3.

BOYDENS, Isabelle, 2018. Data Quality & « back tracking » : depuis les premières expérimentations à la parution d'un Arrêté Royal, *Smals Research, 2018* https://www.smalsresearch.be/data-quality-back-tracking-depuis-les-premieres-experimentations-a-la-parution-dun-arrete-royal/

BOYDENS, Isabelle, 2021. Data Quality Tools : retours d'expérience et nouveautés. *Smals Research, 2021* https://www.smalsresearch.be/data-quality-tools-retours-dexperience-et-nouveautes/

BOYDENS Isabelle, HAMITI Gani G., et VAN EECKHOUT Rudy, 2021, Un service au cœur de la qualité des données : Présentation d'un prototype d'ATMS, *Courrier des statistiques* N° N6, juillet 2021.

BRACKSTONE, G.J. 1987. Issues in the Use of Administrative Records for Statistical Purposes. *Survey Methodology,* vol. 13, pp. 29 – 43

BRION Philippe, Esane, le dispositif rénové de production des statistiques structurelles d'entreprises. Courrier des statistiques n° 130, mai 2011

BRUMBERG H.L., DOZOR D.,et GOLOMBEK S.G., 2012. History of the birth certificate: from inception to the future of electronic data. *Journal of Perinatology* (2012) 32,407–411

De BROE Sofie, STRUIJS Peter, DAAS Piet, Van DELDEN Arnout, BURGER Joep, Van Den BRAKEL Jan, Ten BOSCH Olav, ZEELENBERG Kees, et YPMA Winfried, 2020. Updating the Paradigm of Official Statistics : New Quality Criteria for Integrating New Data and Methods in Official Statistics, *Working paper* 02-20, *Center for Big Data statistics*, CBS

BYCROFT C., et MATESON-DUNNING N., 2020. Use of administrative records for non-response in the New Zealand 2018 Census. *Journal of the IAOS*, 36 (2020) 107–116

CAMUS, Benjamin, 2022. Le défi de l'élaboration d'une nomenclature statistique des infractions. *Courrier des statistiques* N° N7, janvier 2022.

CARON, Nathalie, 2002. la correction de la non-réponse par repondération et par imputation. Document de travail M0502, Insee, Série des documents de travail « Méthodologie statistique ».

CARON, Daniel J., NICOLINI, Vincent, et BERNARDI, Sara, 2020. Réflexion sur les stratégies de données gouvernementales : rapport de recherche. Ecole Nationale d'Administration Publique, Chaire de recherche en exploitation des ressources informationnelles, Bibliothèque et Archives nationales du Québec, mars 2020.

CERRONI, Fulvia, BELLA, Grazia Di et GALIE, Lorena, 2014. Evaluating administrative data quality as input of the statistical production process. *Rivista di statistica ufficiale*. 2014. N° 1, pp. 30.

CHALEIX, Mylène, et VANDERSCHELDEN, Mélanie, 2023. Lignes directrices sur la coopération entre acteurs pour l'exploitation statistique de données administratives, *Document interne Insee*, septembre 2023

CHALEIX, Mylène, et MIKOL, Fanny, 2024. Accès du service statistique public aux données privées : les étapes-clé, les acteurs, les points de vigilance, *Document interne Insee*, septembre 2024

CHAMBAZ, Christine, 2018. De l'activité de la justice au suivi du justiciable - Faire parler les données de gestion. *Courrier des statistiques* N° N1, décembre 2018.

CHIPERFIELD, James, CHU, Randall, ZHANG Li-Chun, et BAFFOUR Bernard, 2024. Robust Statistical Estimation for Capture-Recapture Using Administrative Data. *Journal of Official Statistics* 2024, Vol. 40(2) 215–237

CHRISTEN, Peter et SCHNELL, Rainer, 2024. When Data Science Goes Wrong: How Misconceptions About Data Capture and Processing Causes Wrong Conclusions. *Harvard Data Science Review*, 31 janvier 2024. Vol. 6, n° 1. DOI 10.1162/99608f92.34f8e75b.

CHU, Xu, ILYAS, Ihab, KRISHNAN, Sanjay, WANG Jiannan, 2016. Data cleaning : overview and emerging challenges. SIGMOD '16: Proceedings of the 2016 International Conference on Management of Data, pp. 2201 – 2206, https://doi.org/10.1145/2882903.29125

COTTON Franck, et HAAG Olivier, 2023. L'intégration des données administratives dans un processus statistique : industrialiser une phase essentielle. *Courrier des statistiques* N° N9, juin 2023.

COUDIN, Elise, et ROBERT, Aude, 2024. Les statistiques sur les causes de décès - Classer et coder... dans la classification internationale des maladies. *Courrier des statistiques* N° N12, décembre 2024.

COURMONT Antoine, 2021. Quand la donnée arrive en ville. Open data et gouvernance urbaine. *Presses univ. de Grenoble, « Libres cours Politique »*, ISBN : 9782706147357.
URL : https://www.cairn.info/quand-la-donnee-arrive-en-ville--9782706147357.htm

DAAS, Piet, 2009. Checklist for the Quality evaluation of administrative data sources. *Discussion paper Statistics netherlands*. 2009. pp. 36.

De WAAL T., PANNEKOEK, et J., SCHOLTUS S., 2011, Handbook of statistical data editing and imputation, *Wiley.*

De WAAL, Ton, Van DELDEN, Arnout et SCHOLTUS, Sander, 2020. Multi-source Statistics: Basic Situations and Methods. *International Statistical Review* (2020), 88, 1, 203–228

DENIS, Jérôme, 2018. Le travail invisible des données – Éléments pour une sociologie des infrastructures scripturales. Août 2018. *Presses des Mines, Collection Sciences Sociales*.

DENIS, Jérôme, et GOETA, Samuel, 2017. « La fabrique des données brutes ». Ouvrir, partager, réutiliser, édité par Clément Mabi et al., *Éditions de la Maison des sciences de l'homme*, 2017, https://doi.org/10.4000/books.editionsmsh.9050.

DESFORGES, Corinne, 2021. Réforme de l'état et gestion publique. Revue française d'administration publique n° 180, 2021, p. 1105-1158.

DESROSIERES Alain, 1993. La politique des grands nombres : histoire de la raison statistique. *La découverte*.

DESROSIERES Alain, 2004. Enquêtes versus registres administratifs : réflexions sur la dualité des sources statistiques , *Courrier des statistiques* N° 111, septembre 2004.

DEVILLE, Jean-Claude, 1997. Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement ?, Paris : Insee, 1997, *Document de travail INSEE. Unité méthodologie statistique*, 9701, 8 p

DEVILLE, Jean-Claude, et SÄRNDAL, Carl-Erik, 1992. Calibration estimators in survey sampling. *Journal of the American statistical Association* 87 (418), 376-382

DI RUOCCO, Nunzio, SCHEIWILER, Jean-Michel et SOTNYKOVA, Anastasiya, 2012. La qualité des données : concepts de base et techniques d'amélioration. In : BERTI-ÉQUILLE, Laure, 2012. La qualité et la gouvernance des données au service de la performance des entreprises. *Hermes Science Publications*. pp. 25-54. ISBN 978-2-7462-2510-7.

DUBRULLE, Bertrand, ROSEC, Olivier, SUREAU, Christian, 2023. Une norme d'échange pour alimenter des référentiels et en assurer la qualité. *Courrier des statistiques* N° N9, juin 2023.

DUPONT Françoise, DUSSART Josy, GUILLAUMAT-TAILLIET François. La concertation : une étape essentielle pour le projet Résil. Courrier des statistiques N° N11, juin 2024.

EDWARDS, Paul N., 2010. A vast machine: computer models, climate data, and the politics of global warming. First paperback edition. Cambridge, Massachusetts London, England : The MIT Press. Infrastructures series. ISBN 978-0-262-51863-5.

EDWARDS, P., MAYERNICK, M., BATCHELLER, A., BOWKER, G., et BORGMAN, C., 2011. Science friction: Data, metadata, and collaboration. *Social Studies of Science*, 41(5), 667-690.

ELBAUM Mireille, 2018. Les enjeux des nouvelles sources de données , *Chroniques*, n° 16, *CNIS*, septembre 2018.

ELMASRI, Ramez, et NAVATHE, Shamkant, 2016. Fundamentals of Database Systems, *Global Edition, Pearson*.

ERIKSON, Johan, 2020. Le modèle de processus statistique en Suède Mise en œuvre, expériences et enseignements. Courrier des statistiques N° N4, juin 2020.

ESAYAS, Samson Yoseph, 2015. The role of anonymisation and pseudonymisation under the EU data privacy rules, *European Journal of Law and Technology* Vol 6, No 2 (2015).

ESPINASSE, Lionel, ROUX, Valérie, 2022. Le Répertoire national d'identification des personnes physiques (RNIPP) au cœur de la vie administrative française. *Courrier des statistiques* N° N8, Insee, novembre 2022.

EUROSTAT, 2016a. Methodology for data validation 1.0. Essnet Validat Foundation, juin 2016.

EUROSTAT, 2016b. Estimation methods for the integration of administrative sources. Deliverable D1. Décembre 2016 (version 3).

EUROSTAT, 2017a. Code des bonnes pratiques de la statistique européenne. 2017.

EUROSTAT, 2017b. Good practices in accessing, using and contributing to the management of administrative data. Eurostat. WP1

EUROSTAT, 2017c. Estimation methods for the integration of administrative sources. Deliverable D2. Janvier 2017 (version 3).

EUROSTAT, 2021. European Statistical System Handbook for quality and metadata reports. *Eurostat manuals and guidelines*, 2021, re-edition.

FAO, 2018. Guidelines on improving and using administrative data in agricultural statistics. Food and Agriculture Organization of the United Nations, Février 2018.

FELLEGI, I.P., HOLT, D., 1976. A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association, Vol. 71, No. 353 (Mar., 1976), pp. 17-35

FERMOR-DUNMAN, Verena et PARSONS, Laura, 2022. Data Acquisition processes improving quality of microdata at the Office for National Statistics. *Q2022*, Vilnius.

FOLEY Brian, SHUTTLEWORTH, Ian, and MARTIN, David, 2018. Administrative Data Quality: Investigating Record-Level Address Accuracy in the Northern Ireland Health Register. *Journal of Official Statistics*, Vol. 34, No. 1, 2018, pp. 55–81, http://dx.doi.org/10.1515/JOS-2018-0004

FRIEDL, Jeffrey E.F., 2006. Mastering regular expressions - Third Edition, O'Reilly.

FROESCHL, Karl A. et GROSSMANN, Wilfried, 2000. The Role of Metadata in Using Administrative Sources. *Research in Official Statistics*, 2000.

GARTNER, Richard, 2016. Metadata – Shaping knowledge from Antiquity to the Semantic Web. Springer.

GIRARD-CHANUDET, Camille, 2023. Le travail de l'Intelligence Artificielle : concevoir et entraîner un outil de pseudonymisation automatique à la Cour de Cassation - *RESET (Recherches en sciences sociales sur Internet)*, Technologies numériques et apprentissage.

GITELMAN, Lisa (éd.), 2013. « Raw data » is an oxymoron. Cambridge, Mass. : *MIT Press. Infrastructures series*. ISBN 978-0-262-31232-5.

GOETA, Samuel, 2024. Les données de la démocratie. *C&F éditions, Collection Société numérique*, février 2024.

GRANQUIST, Leopold, 1997. The new view on editing. *International Statistical Review* (1997), 65, 3, 381-387.

GROVES, Robert, et LYBERG, Lars, 2010. Total survey error – Past, present and future. *Public Opinion Quarterly*, Vol. 74, No. 5, 2010, pp. 849–879.

GUIBERT, Bernard, LAGANIER, Jean et VOLLE, Michel, 1971. Essai sur les nomenclatures industrielles. *Économie et statistique*. Février 1971. Insee. N° 20, pp. 23-36.

HACHID, Ali, et LECLAIR, Marie, 2022. Sirus, le répertoire d'entreprises au service du statisticien. *Courrier des statistiques* N° N8, novembre 2022.

HAND, David J., 2018. Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society: Series A* (Statistics in Society) [en ligne]. juin 2018. Vol. 181, n° 3, pp. 555-605. DOI 10.1111/rssa.12315.

HEDLIN, Dan, 2003. Score functions to reduce business survey editing at the UK office for national statistics, *Journal of Official Statistics*, 2003.

HELFENSTEIN, Xavier, 2022. La base permanente des équipements (BPE) - Une source statistique singulière et constamment en mouvement, *Courrier des statistiques* N° N8, Insee, novembre 2022.

HORVITZ, D.G., et THOMPSON, D.J., 1952. A Generalization of Sampling Without Replacement From a Finite Universe, *Journal of the American Statistical Association*, Vol. 47, No. 260 (Dec., 1952), pp. 663 – 685.

HUMBERT-BOTTIN, Élisabeth, 2018. La déclaration sociale nominative. Nouvelle référence pour les échanges de données sociales des entreprises vers les administrations. *Courrier des statistiques* N° N1, décembre 2018.

IMBERT, Alyssa, et VIALANEIX, Nathalie, 2018. Décrire, prendre en compte, imputer et évaluer les valeurs manquantes dans les études statistiques : une revue des approches existantes. *Journal de la Société Française de Statistique*, Vol. 159 No. 2 (2018)

JOUBERT, Marie-Pierre, 2025. Les données de téléphonie mobile - Une source de connaissance sur la population et ses déplacements. *Courrier des statistiques* N° N13, juin 2025

JURAN, Joseph M., 1951. Quality control handbook. *Mc Graw Hill*

KALIKA Michel, ROWE Frantz, FALLERY Bernard, REIX Robert, et RICHET Jean-Loup, 2023. Systèmes d'information et management - Le manuel de référence sur les SI, 8e édition, *Vuibert*.

KITCHIN, Rob, 2014. The Data Revolution – Big Data, Open Data, Data Infrastructures and Their Consequences. *SAGE Publications*. ISBN 978-1-4462-8747-7.

KOUMARIANOS, Heidi, LEFEBVRE, Olivier, et MALHERBE, Lucas, 2024. Les appariements : finalités, pratiques et enjeux de qualité. Courrier des statistiques N° N11, juillet 2024.

LAITILA, Thomas, WALLGREN, Anders, et WALLGREN, Britt, 2011. Quality Assessment of Administrative Data. *Methodology reports from Statistics Sweden* 2011:2.

LAMARCHE, Pierre, et LOLLIVIER, Stéfan, 2021. Fidéli, l'intégration des sources fiscales dans les données sociales. *Courrier des statistiques* N° N6, juillet 2021.

LAWRENCE, David, et McKENZIE, Richard, 2000. The general application of significance editing. *Journal of Official Statistics*, Vol. 16, No 3, 2000, pp. 243-253.

LECLAIR, Marie, 2019. Utiliser les données de caisse pour le calcul de l'indice des prix à la consommation, *Courrier des statistiques* N° N3, décembre 2019.

LEFEBVRE, Olivier, 2024. Le Répertoire Statistique des Individus et des Logements (Résil). *Courrier des statistiques* N° N11, juillet 2024.

LEFEBVRE, Olivier, SOULIER Manuel, et TORTOSA Thomas, 2024. L'accueil des données administratives : un processus structurant. *Courrier des statistiques* N° N11, juillet 2024.

LEHUEDE, Sebastian, 2024. When friction becomes the norm: Antagonism, discourse and planetary data turbulence. *New Media & Society* 26 (7), 3951-3966.

LESSLER, J.T., KASLBEEK W., 1992. Nonsampling error in surveys. New York: *Wiley.*

LESUR, Romain, 2025. Sources de données privées : panorama et perspectives. *Courrier des statistiques* N° N13, juin 2025

LITTLE, Roderick J.E., 1982. Models for Nonresponse in Sample Surveys. *Journal of the* American Statistical Association, Vol. 77, No. 378 (Jun., 1982), pp.237-250

LITTLE, Roderick J.E., et RUBIN, Donald B., 2002. Statistical Analysis with Missing Data. Wiley.

LOMBARDO, Giacomo, 2023. Data Friction in Data Sharing: a Physics Inspired Model, *Thèse en ingénierie informatique, institut polytechnique de Milan.*

LOSHIN, David, 2011. The practitioner's guide to data quality improvement. Burlington, MA : *Morgan Kaufmann*. ISBN 978-0-12-373717-5.

LOTHIAN, Jack, HOLMBERG, Anders, et SEYB, Allyson, 2019. An Evolutionary Schema for Using "it-is-what-it-is" Data in Official Statistics, *Journal of Official Statistics*, Vol. 35, No. 1, 2019, pp. 137–165.

LYBERG, Lars J., WEISBERG, Herbert E., 2016. Total survey error : a paradigm for survey methodology. In The SAGE handbook of survey methodology, SAGE Publications.

MALHERBE, Lucas, 2023. Appariements de données individuelles : concepts, méthodes, conseils. Insee, document de travail M2023/03, juin 2023.

MATTHEWS, Gregory J., HAREL, Ofer, 2011. Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. Statistics Surveys Vol. 5 (2011) 1–29. ISSN: 1935-7516.

MEIERHOFER Jürg, STADELMANN, Thilo, et CIELIBAK, Mark, 2019. Data products. In *Braschler, Stadelmann, Stockinger (Eds.): "Applied Data Science - Lessons Learned for the Data-Driven Business"*, Springer

MEYER, Bruce D. , MOK, Wallace K. C., et SULLIVAN, James X., 2015. Household Surveys in Crisis, J*ournal of Economic Perspectives*, Volume 29, Number 4, pp. 199–226

MOREAU, Sylvain, 2024. La statistique annuelle d'entreprises : sa nature, son histoire, ses enjeux, *Courrier des statistiques* N° N12, Décembre 2024.

NATIONAL ACADEMIES of Science - Engineering – Medicine, 2019. Reproducibility and replicability in science. *The National Academies Press ;* Illustrated edition (October 2019)

NEYMAN, J. , 1934. « On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection », *Journal of the Royal Statistical Society*, 97, p. 558 à 625.

NORDBOTTEN, Svein, 2010. The Use of Administrative Data in Official Statistics – Past, Present, and Future – With Special Reference to the Nordic Countries, *Official Statistics in Honour of Daniel Thorburn*, pp. 205–223

OECD, 2008. OECD Glossary of Statistical Terms. *OECD Publishing*.

OLSON, Jack E., 2003. Data Quality – The Accuracy Dimension. [en ligne]. Janvier 2003. Morgan Kaufmann. ISBN 1-55860-891-5.

ÓVÁRI, Kristián, KOČIŠ, Martin, 2022. Population Census 2021 in the Slovak Republic: The "Signs of Life" method. Q2022, Vilnius.

PERIGNON, Christophe, GADOUCHE, Kamel, HURLIN, Christophe, SILBERMAN, Roxane et DEBONNEL, Éric, 2019. Certify reproducibility with confidential data. *Science*. Juillet 2019.

PUCHE, Véronique, 2022. Nouveau regard sur « La branche Famille et la protection des libertés individuelles ». *Regards* 2022/1 N°60, pp. 76 à 82.

RANCOURT, Eric, 2018. Les données administratives d'abord comme paradigme statistique pour les statistiques officielles canadiennes : signification, défis et possibilités.
Actes du Symposium de 2018 de Statistique Canada.

REDMAN, Thomas C., 1997. Data Quality for the Information Age. Janvier 1997. *Artech House Computer Science Library.* pp. 227-232. ISBN 978-0-89006-883-0.

REDOR, Patrick, 2023. Confidentialité des données statistiques : un enjeu majeur pour le service statistique public, *Courrier des statistiques* N° N9, juin 2023

REID Giles, ZABALA Felipa, HOLMBERG Anders, 2017. Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ. *Journal of Official Statistics*, Vol. 33, No. 2, 2017, pp. 477–511

RENNE Catherine, 2018, Bien comprendre la déclaration sociale nominative pour mieux mesurer, *Courrier des statistiques* N° N1, décembre 2018.

REY, Olivier, 2016. Quand le monde s'est fait nombre. *Stock*. Octobre 2016

RICCIATO, Fabio, 2024. Steps Toward a Shared Infrastructure for Multi-Party Secure Private Computing in Official Statistics, *Journal of Official Statistics* 2024, Vol. 40(1) 3–15

RIVIÈRE Pascal, 2018. Utiliser des déclarations administratives à des fins statistiques, *Courrier des statistiques* N° N1, décembre 2018.

RIVIÈRE, Pascal, 2020. Qu'est-ce qu'une donnée ? Impact des données externes sur la statistique publique. *Courrier des statistiques* N° N5, 31 décembre 2020.

ROUPPERT Benoît, 2005. Modélisation du processus de traitement d'une source administrative à des fins statistiques. *Document de travail SGI, Insee*, 2005

SALEMINK, Irene, DUFOUR, Stéphane et VAN DER STEEN, Marcel 2020. A vision on future advanced data collection, *Statistical Journal of the IAOS* 36 (2020) 685–699

SALGADO, David, et OANCEA, Bogdan, 2020. On new data sources for the production of official statistics. *ArXiv preprint*, 15 mars 2020.

SCHWEINFEST, Stefan, et JANSEN, Ronald, 2021. Data Science and Official Statistics: Toward a New Data Culture. *Harvard Data Science Review* • Issue 3.4, Fall 2021

SHAH Syed Iftikhar Hussain , PERISTERAS Vassilios et MAGNISALIS Ioannis, 2021. DaLiF: a data lifecycle framework for data-driven governments. Journal of Big Data (2021) 8:89, https://doi.org/10.1186/s40537-021-00481-3

STATISTICS NEW ZEALAND, 2016. Guide to reporting on administrative data quality. Retrieved from www.stats.govt.nz

SUNDGREN Bo, 1993. Statistical Metainformation Systems – Pragmatics, Semantics, Syntactics. Journal of the IAOS, Vol. 10, Issue 2. https://doi.org/10.3233/SJU-1993-10205

TAVERNIER, Jean-Luc, 2020. Fonctionnement de l'Insee dans la période de confinement, *Courrier des statistiques* N° N5, décembre 2020

TIIT Ene-Margit, 2017. Residency testing. estimating the true population size of Estonia. *Statistics in transition new series*, June 2017, Vol. 18, No. 2, pp. 211–226

UNECE, 2011. Using Administrative and Secondary Sources for Official Statistics. A Handbook of Principles and Practices. United Nations Economic Commission for Europe, https://digitallibrary.un.org/record/719971?v=pdf#files

UNECE, 2019. GSBPM. In : site statswiki de l'UNECE. Disponible à l'adresse : https://statswiki.unece.org/display/GSBPM.

UNITED NATIONS, 2025. Module for Quality Assurance when using Administrative and Other Data Sources to produce Official Statistics. Statistical Commission, Expert Group on National Quality Assurance Frameworks, mars 2025

VAN DELDEN, Arnout, PANNEKOEK, Jeroen, BANNING, Reinder et DE BOER, Arjen, 2016. Analysing correspondence between administrative and survey data. Statistical Journal of the IAOS 32 (2016) 569–584 569. DOI 10.3233/SJI-160972

VAN DROMME Dries, BOYDENS, Isabelle, BONTEMPS, Yves, 2007. Data quality : tools. *Rapport Smals Research 2007/TRIM3/02*. Septembre 2007.

VICARD, Augustin, 2023. Quantifier la pratique sportive : une approche sociologique et sanitaire. *Courrier des statistiques* N° N10, décembre 2023.

VOLLE, Michel, 1980. Le métier de statisticien. [en ligne]. *Éditions Hachette Littérature*. ISBN 978-2-010045295.

WALLGREN, Anders et WALLGREN, Britt, 2007. Register-based Statistics : Administrative Data for Statistical Purposes. New York : *Wiley.*

WALLGREN, Anders et WALLGREN, Britt, 2016. Frames and Populations in a Registerbased National Statistical System. *Journal of Mathematics and Statistical Science*, Volume 2016, pp. 208-216.

WIENER, Norbert, 1948. Cybernetics or control and communication in the human and machine, *MIT,* 1948

ZHANG, Li-Chun, 2012. Topics of statistical theory for register-based statistics and data integration. In : *Statistica Neerlandica* (2012) Vol. 66, nr. 1, pp. 41–63

ZHANG, Li-Chun, 2021. On provision of UK neighbourhood population statistics beyond 2021. *University of Southampton. ArXiV,* 4 nov 2021

Annexe 1 : Typologies possibles pour les données administratives

Le terme « sources administratives » englobe un ensemble de données d'origine, de nature qui peuvent être très différentes. On trouve peu de tentatives de typologie dans la littérature (Unece 2011; G. J. Brackstone 1987). On va présenter trois typologies existantes, et en proposer ici une quatrième.

- **Typologie proposée par l'Unece en 2011**

L'Unece propose une approche thématique, en partie liée à l'organisation administrative. Dans le handbook de 2011 sont listées les thématiques suivantes :

- Données fiscales
  - Impôt sur le revenu des personnes physiques
  - Taxe sur la valeur ajoutée (TVA)
  - Impôt sur les entreprises et les bénéfices
  - Impôts fonciers
  - Droits d'importation / d'exportation

- Données relatives à la sécurité sociale
  - Cotisations
  - Prestations
  - Pensions

- Dossiers de santé / d'éducation

- Systèmes d'enregistrement des personnes / entreprises / biens / véhicules

- Cartes d'identité / passeports / permis de conduire

- Registres électoraux

- Registre des exploitations agricoles

- Registres des conseils locaux

- Permis de construire

- Systèmes de licences, par exemple pour la télévision, la vente de produits soumis à des restrictions

- Comptes publiés des entreprises

- Données comptables internes détenues par les entreprises

- Entreprises privées détenant des données :
  - Agences de crédit
  - Analystes d'entreprise
  - Sociétés de services publics
  - Annuaires téléphoniques
  - Détaillants avec cartes de magasin, etc.

- **Typologie proposée par l'Unece en 2018**

En 2018, l'Unece présente dans une conférence un niveau plus agrégé de distinction des sources, en fonction d'un objectif général du processus administratif :

> - Données communiquées aux autorités administratives par des personnes physiques ou morales pour se conformer à la loi ou pour accéder à des services publics
>
> - Données enregistrant les décisions prises par les autorités administratives
>
> - Données générées par les autorités administratives pour soutenir la planification, la mise en œuvre, le suivi et l'évaluation des programmes administratifs

- **Typologie proposée par Statistique Canada en 1987**

En remontant un peu plus loin dans le passé, Brackstone a également tenté l'exercice à Statistique Canada. Il s'appuie sur les objectifs des sources, ceux-ci ayant un impact sur la couverture et la qualité des données :

1. Registres tenus pour réguler les flux de biens et de personnes à travers les frontières : Il s'agit des registres des importations, des exportations, de l'immigration et de l'émigration.

2. Documents résultant de l'obligation légale d'enregistrer des événements particuliers.
   Les exemples incluent les naissances, les décès, les mariages, les divorces, les incorporations ou fusions d'entreprises, les licences, etc.

3. Documents nécessaires à l'administration des prestations ou des obligations.
   Il s'agit par exemple des impôts, de l'assurance chômage, des pensions, de l'assurance maladie et des allocations familiales.

4. Documents nécessaires à l'administration des institutions publiques.
   Exemples : les documents relatifs aux écoles, aux universités, aux établissements de santé, aux tribunaux et aux prisons.

5. Documents découlant de la réglementation gouvernementale de l'industrie. Il s'agit par exemple des documents relatifs aux transports, à la banque, à la radiodiffusion et aux télécommunications.
   Il s'agit également des documents résultant de la gestion de l'offre ou du prix de certains produits, en particulier dans le domaine de l'agriculture.

6. Documents relatifs à la fourniture de services d'utilité publique
   Il s'agit des services d'électricité, de téléphone et d'eau.

- **Proposition de typologie**

Dans les typologies qui précèdent, on voit apparaître différents critères de regroupement : il peut s'agir de prendre en compte la thématique, ou bien de distinguer les données en fonction de la finalité du processus administratif ou de l'entité à l'origine de la donnée.

Certaines données sont liées à un évènement réel : elles peuvent constituer un enregistrement de cet évènement (naissance, décès) ou au contraire, c'est la démarche administrative qui entérine l'évènement

(création d'entreprise, transaction immobilière, mariage). Enfin, certaines données sont des décisions de gestion administratives, sans lien formel avec un évènement de la vie réelle.

Deux critères semblent plus discriminants pour décrire les forces et faiblesses des différents types de sources, comparativement au cadre plus familier des enquêtes : l'entité à l'origine de la donnée (l'administration ou non) et le lien avec un évènement de la vie réelle. On fait l'hypothèse sous-jacente que la manière dont naissent les données influe de manière substantielle sur leurs caractéristiques.

On propose ci-dessous de distinguer six types de données administratives, en cherchant à mettre en évidence certaines caractéristiques fortement liées au processus dans lequel elles naissent :

1. Les déclarations régulières ;
2. Les déclarations ponctuelles ;
3. Les demandes d'accès à certains droits, services ;
4. Les données de gestion de l'activité administrative ;
5. Les répertoires ;
6. Les données issues de capteurs ou traceurs.

Il est important de préciser à ce stade qu'un jeu de données fourni par une administration peut être un ensemble composite de ces différents types de données.

1) Les déclarations régulières

Elles sont à l'initiative de l'administration, qui définit et a une connaissance de la population assujettie. Elles se déroulent de façon répétée, et pendant une période définie. Les données sont généralement recueillies via un formulaire établi par l'administration. Les déclarations font souvent l'objet d'une réglementation, et peuvent revêtir un caractère obligatoire. L'administration définit les concepts et les listes de valeurs. Les données issues des processus concernés sont donc bien définies et maîtrisées par l'administration, en terme de champ, de concepts, de dates de référence.
C'est le cas des déclarations fiscales et sociales notamment.

Dans le cas de déclarations, l'administration peut parfois pré-remplir un certain nombre d'informations, sur la base d'éléments connus par ailleurs (les revenus salariaux pré-remplis dans la déclaration d'impôts par exemple), ou par le biais de déclarations antérieures (le portail GMBI ne demandera qu'une mise à jour des informations existantes pour sa deuxième année de collecte en 2024). Les processus de collecte organisés souffrent alors de satisficing, pouvant conduire à renseigner des informations approximatives, ou à ne pas mettre à jour des informations.

2) Les déclarations ponctuelles, liées à un évènement

Elles sont à l 'initiative d'un usager (déclarant) mais peuvent porter sur un autre objet (naissance, décès d'un proche, création d'une entreprise, déclaration d'un accident, changement d'adresse). L'évènement, en général un évènement du monde physique, peut précéder la déclaration (naissance, décès), ou bien c'est l'enregistrement de l'évènement qui fait exister « quelque chose » (transaction immobilière, mariage, pacs, création d'entreprise). L'administration reste à l'initiative du formulaire, et donc de la structure des données collectées.

Elles se produisent dans une temporalité non contrôlée par l'administration, la survenue de l'évènement n'étant pas prévisible et son enregistrement pouvant être éloigné temporellement de l'évènement lui-même.
La population d'intérêt n'est a priori pas connue, et il peut exister de la non déclaration. On peut expliciter la population « autorisée à déclarer » (habilitation à déclarer, par exemple on sait que les naissances seront déclarées par un individu), mais il peut y avoir une différence entre l'entité déclarante et l'objet concerné par la déclaration. On peut relever également que certains évènements sont uniques (théoriquement), lorsqu'ils concernent la création, l'existence d'un nouvel objet, ou bien sa suppression, sa disparition, alors que d'autres, concernant un changement d'état, pourront se produire plusieurs fois pour un même objet.

3) Les demandes de l’usager auprès de l'administration

Ces demandes pourront lui accorder des droits, prestations… Il peut s'agir d'une demande de retraite, de prestations sociales (allocations, subvention ou aide). La population est définie comme celle des demandeurs. Les ayant droit qui ne demandent pas ne font pas partie du champ du processus administratif, bien qu'ils puissent être des unités d'intérêt pour la statistique. Il n'y a donc pas de non déclaration, comme dans le cas précédent, mais du non recours.

Pour ces trois premiers types, la donnée administrative est la conséquence d'une interaction entre l'administration et un usager, et il y a un processus de recueil d'informations qui se rapproche d'un processus d'observation.

4) Les données de gestion de l’activité administrative

D'autres données sont le reflet du fonctionnement interne de l'administration. Il peut s'agir par exemple de retracer les activités de l'administration judiciaire (Chambaz, 2018). Les objets d'intérêt pour l'administration ne sont pas nécessairement ceux qui intéressent le statisticien. Les objets administratifs peuvent être difficiles à cerner et à suivre dans le temps. L'objet administratif est défini a priori, mais il peut être complexe (cf partie sur les objets dans le chapitre 2).

Lorsqu'on a affaire à des données de gestion, c'est l'administration gestionnaire qui décide des mises à jour des informations : leur évolution ne procède pas forcément d'une interaction avec l'usager (mise à jour de carrière pour les retraites, fermeture et réouverture d'offres d'emploi). La donnée n'est pas nécessairement une observation, dans ce cas il s'agit d'une décision de gestion.

Pour un même objet administratif, il est possible de disposer d'un mélange de données d'observation et de données de gestion. Selon l'organisation de l'administration, les pratiques de gestion peuvent différer d'une entité à l'autre.

On pourrait inclure dans cette catégorie des données très particulières : les paramètres du fonctionnement d'une administration (tarification, point d’indice, ...), qui sont en réalité des données de « haut niveau » permettant d’en calculer beaucoup d’autres.

5) Les répertoires

Ils ont une position un peu spécifique par rapport aux types de sources cités précédemment, car à certains égards, ils ne forment pas une catégorie distincte des précédentes. Leur objectif est de mettre à disposition des données de référence, pour des utilisateurs variés. Elles peuvent provenir à l'origine de déclarations, organisées par l'administration pour constituer un ensemble à partir d'informations de temporalité différente.
Les répertoires, en tant que *liste d’instances d’une même entité* (Rivière, 2022) cherchent à décrire le plus exhaustivement possible l'existence des objets d'intérêt pour l'administration, en enregistrant autant que faire se peut les changements d'état ou d'attribut de ces objets.

Tout comme les jeux de données peuvent mélanger données issues de déclarations et données de gestion, certains jeux de données peuvent mélanger les informations d'un répertoire constitué par l'administration, et des données déclaratives récentes (exemple données fiscales, PAC).

6) Les données produites par des systèmes de mesure automatiques

Ce dernier type de source est moins fréquemment utilisé. Il peut s’agir de capteurs physiques (puces GPS de navires, mesures de trafic routier, conditions météorologiques) ou des traceurs numériques (données de cartes bancaires, utilisation de la carte vitale).

On constate à travers cette tentative de typologie que les notions d'objets, d'entité à l'origine de la donnée, et de temporalité sont importantes dans la construction des données, et par conséquent dans leur exploitation.

Quel que soit le type de donnée administrative, l'exploitation statistique se heurte à des difficultés : les concepts et nomenclatures administratifs divergent des unités d'intérêt statistique, les mécanismes de non réponse ou les problèmes de couverture du champ théorique ne sont pas connus, et de manière générale, la connaissance des processus de collecte et de traitement des données est souvent limitée.
En outre, les données sont liées à l'existence des processus qu'elles soutiennent et ceux-ci peuvent évoluer fortement voire disparaître, comme la TH récemment.

Cependant, l'origine des données conduit également à des différences : l'existence d'un protocole de collecte formalisé contribue à une proximité plus importante des données administratives et données d'enquête.

Tableau : Attributs des différents types de données administratives

	Initiative	Processus de collecte			Connaissance a priori de la population d'intérêt
		Support de collecte défini	Période de collecte délimitée	Période de référence	
Déclaration régulière	Administration	Oui	Oui	Oui	Oui
Déclaration ponctuelle	Usager	Oui	Non	Oui	Non
Demande d'un usager	Usager	Oui	Non	Oui	Non
Fonctionnement interne	Administration	Non	Non	Non	Non
Systèmes de mesure automatisés	Administration	Oui	Non	Oui	Oui

Annexe 2 : Dimensions pour caractériser la qualité des données administratives selon Istat

Source	Métadonnées	Données			
			Jeu de données	**Unités**	**Variables**
Fournisseur Pertinence de la source Sécurité et protection des données Transmission des données Processus de transmission des données, Relations avec le fournisseur	Clarté des concepts et définitions Comparabilité des concepts et définitions Clés primaires Traitement des données effectué par le fournisseur	**Contrôles techniques**	Accessibilité		Conformité aux définitions
		Intégration		Comparabilité des unités (ou capacité à les transformer)	Présence de variables d'appariement comparables
		Précision		Unités présentant des incohérences, des anomalies	Valeurs incohérentes, valeurs en anomalie
		Complétude		Sur couverture, sous couverture, doublons	Valeurs manquantes, valeurs imputées
		Dimension temporelle	Délai par rapport à la période de référence	Evolution des unités dans le temps (sorties et entrées de champ)	Stabilité des définitions

Annexe 3 : Liste des sources de données administratives citées en exemple

La liste des sources administratives présentée ci-dessous se limite aux principales sources évoquées dans le document, en ayant en tête la difficulté que pose la notion même de source, vue au chapitre 3c notamment. On va les décliner par thème.

Données fiscales :

TH : taxe d'habitation. Jusqu'en 2022, tous les logements étaient soumis à la taxe d'habitation, les déclarants étant les occupants. A partir de 2023, seules les résidences secondaires et vacances y sont soumises.

GMBI : Gérer mes biens immobiliers est une source fiscale sur les locaux d'habitation. Les propriétaires sont soumis à déclaration de leurs locaux, ainsi que des occupants éventuels.

POTE : déclaration d'impôts sur le revenu des personnes.

Liasses : déclarations fiscales des entreprises. Elles portent sur les comptes de résultat, mais également le bilan, les immobilisations et amortissements.

Cadastre (ou Majic) : Majic signifie mise à jour des informations cadastrales. Cette source répertorie les parcelles non bâties, bâties. Elle apporte de l'information également sur le bâti, les locaux, et leurs propriétaires.

Données sociales :

DSN (déclaration sociale nominative) :

> « La Déclaration sociale nominative (DSN) – est un fichier mensuel produit à partir de la paie destinées à communiquer les informations nécessaires à la gestion de la protection sociale des salariés aux organismes et administrations concernées. »

Pasrau (passage des revenus autres) :

> « Le dispositif PASRAU (Prélèvement à la source pour les Revenus Autres) résulte de travaux de simplification et de rationalisation des déclarations sociales. Il est le prolongement logique de la DSN, qui a constitué ces dernières années une simplification majeure des procédures déclaratives concernant les salaires et les revenus versés par un employeur. Ce dispositif, fondé sur la norme NEORAU, complète donc la DSN (fondée sur la norme NEODeS) pour les « revenus de remplacement ». »

DRM (Dispositif de revenus mensuels) :

> « Le DRM est constitué de trois traitements de données : une base de données alimentée mensuellement par les données issues de la « déclaration sociale nominative » (DSN), une base de données alimentée mensuellement par des données issues du flux du « prélèvement à la source mis en œuvre par les collecteurs n'entrant pas dans le champ de la déclaration

sociale nominative ou versant des revenus de remplacement » (PASRAU) et un service de restitution. Ce dispositif centralise la quasi-totalité des données relatives aux revenus de la population connus de l'administration fiscale. »

CNAF : caisse nationale des allocations familiales

CNAV : caisse nationale de l'assurance vieillesse

RGCU : répertoire général des carrières unique

Autres données mentionnées :

Cartes grises : il s'agit d'un fichier repértoriant les véhicules et leurs propriétaires.

PAC : politique agricole commune : ces données comportent les demandes d'aide des exploitations agricoles. Elles sont fournies par l'Agence de services et de paiement (ASP).

Série des Documents de Travail « Méthodologie Statistique »

9601 : Une méthode synthétique, robuste et efficace pour réaliser des estimations locales de population.
G. DECAUDIN, J.-C. LABAT

9602 : Estimation de la précision d'un solde dans les enquêtes de conjoncture auprès des entreprises.
N. CARON, P. RAVALET, O. SAUTORY

9603 : La procédure FREQ de SAS - Tests d'indépendance et mesures d'association dans un tableau de contingence.
J. CONFAIS, Y. GRELET, M. LE GUEN

9604 : Les principales techniques de correction de la non-réponse et les modèles associés.
N. CARON

9605 : L'estimation du taux d'évolution des dépenses d'équipement dans l'enquête de conjoncture : analyse et voies d'amélioration.
P. RAVALET

9606 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT).
S. LOLLIVIER, M. MARPSAT, D. VERGER

9607 : Enquêtes régionales sur les déplacements des ménages : l'expérience de Rhône-Alpes.
N. CARON, D. LE BLANC

9701 : Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement ?
J.-C. DEVILLE

9702 : Modèles univariés et modèles de durée sur données individuelles.
S. LOLLIVIER

9703 : Comparaison de deux estimateurs par le ratio stratifiés et application aux enquêtes auprès des entreprises.
N. CARON, J.-C. DEVILLE

9704 : La faisabilité d'une enquête auprès des ménages.
1. au mois d'août.
2. à un rythme hebdomadaire

C. LAGARENNE, C. THIESSET

9705 : Méthodologie de l'enquête sur les déplacements dans l'agglomération toulousaine.
P. GIRARD.

9801 : Les logiciels de désaisonnalisation TRAMO & SEATS : philosophie, principes et mise en œuvre sous SAS.
K. ATTAL-TOUBERT, D. LADIRAY

9802 : Estimation de variance pour des statistiques complexes : technique des résidus et de linéarisation.
J.-C. DEVILLE

9803 : Pour essayer d'en finir avec l'individu Kish.
J.-C. DEVILLE

9804 : Une nouvelle (encore une !) méthode de tirage à probabilités inégales.
J.-C. DEVILLE

9805 : Variance et estimation de variance en cas d'erreurs de mesure non corrélées ou de l'intrusion d'un individu Kish.
J.-C. DEVILLE

9806 : Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE.
N. CARON, J.-C. DEVILLE, O. SAUTORY

9807 : Estimation de données régionales à l'aide de techniques d'analyse multidimentionnelle.
K. ATTAL-TOUBERT, O. SAUTORY

9808 : Matrices de mobilité et calcul de la précision associée.
N. CARON, C. CHAMBAZ

9809 : Échantillonnage et stratification : une étude empirique des gains de précision.
J. LE GUENNEC

9810 : Le Kish : les problèmes de réalisation du tirage et de son extrapolation.
C. BERTHIER, N. CARON, B. NEROS

9901 : Perte de précision liée au tirage d'un ou plusieurs individus Kish.
N. CARON

9902 : Estimation de variance en présence de données imputées : un exemple à partir de l'enquête Panel Européen.
N. CARON

0001 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT) (version actualisée).
S. LOLLIVIER, M. MARPSAT, D. VERGER

0002 : Modèles structurels et variables explicatives endogènes.
J.-M. ROBIN

0003 : L'enquête 1997-1998 sur le devenir des personnes sorties du RMI - Une présentation de son déroulement.
D. ENEAU, D. GUILLEMOT

0004 : Plus d'amis, plus proches ? Essai de comparaison de deux enquêtes peu comparables.
O. GODECHOT

0005 : Estimation dans les enquêtes répétées : application à l'Enquête Emploi en Continu.
N. CARON, P. RAVALET

0006 : Non-parametric approach to the cost-of-living index.
F. MAGNIEN, J. POUGNARD

0101 : Diverses macros SAS : Analyse exploratoire des données, Analyse des séries temporelles.
D. LADIRAY

0102 : Économétrie linéaire des panels : une introduction.
T. MAGNAC

0201 : Application des méthodes de calages à l'enquête EAE-Commerce.
N. CARON

C 0201 : Comportement face au risque et à l'avenir et accumulation patrimoniale - Bilan d'une expérimentation.
L. ARRONDEL, A. MASSON, D. VERGER

C 0202 : Enquête Méthodologique Information et Vie Quotidienne - Tome 1 : bilan du test 1, novembre 2002.
J.-A. VALLET, G. BONNET, J.-C. EMIN, J. LEVASSEUR, T. ROCHER, P. VRIGNAUD, X. D'HAULTFOEUILLE, F. MURAT, D. VERGER, P. ZAMORA

0203 : General principles for data editing in business surveys and how to optimise it.
P. RIVIERE

0301 : Les modèles logit polytomiques non ordonnés : théories et applications.
C. AFSA ESSAFI

0401 : Enquête sur le patrimoine des ménages - Synthèse des entretiens monographiques.
V. COHEN, C. DEMMER

0402 : La macro SAS CUBE d'échantillonnage équilibré
S. ROUSSEAU, F. TARDIEU

0501 : Correction de la non-réponse et calage de l'enquêtes Santé 2002
N. CARON, S. ROUSSEAU

0502 : Correction de la non-réponse par répondération et par imputation
N. CARON

0503 : Introduction à la pratique des indices statistiques - notes de cours
J-P BERTHIER

0601 : La difficile mesure des pratiques dans le domaine du sport et de la culture - bilan d'une opération méthodologique
C. LANDRE, D. VERGER

0801 : Rapport du groupe de réflexion sur la qualité des enquêtes auprès des ménages
D. VERGER

M2013/01 : La régression quantile en pratique
P. GIVORD,
X. D'HAULTFOEUILLE

M2014/01 : La microsimulation dynamique : principes généraux et exemples en langage R
D. BLANCHET

M2015/01 : la collecte multimode et le paradigme de l'erreur d'enquête totale
T. RAZAFINDROVONA

M2015/02 : Les méthodes de Pseudo-Panel
M. GUILLERM

M2015/03 : Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse
E. GROS
K. MOUSSALAM

M2016/01 : Le modèle Logit Théorie et application.
C. AFSA

M2016/02 : Les méthodes d'estimation de la précision de l'Enquête Emploi en Continu
E. GROS
K.MOUSSALAM

M2016/03 : Exploitation de l'enquête expérimentale Vols, violence et sécurité.
T. RAZAFINDROVONA

M2016/04 : Savoir compter, savoir coder. Bonnes pratiques du statisticien en programmation.
E. L'HOUR
R. LE SAOUT
B. ROUPPERT

M2016/05 : Les modèles multiniveaux
P. GIVORD
M. GUILLERM

M2016/06 : Econométrie spatiale : une introduction pratique
P. GIVORD
R. LE SAOUT

M2016/07 : La gestion de la confidentialité pour les données individuelles
M. BERGEAT

M2016/08 : Exploitation de l'enquête expérimentale Logement internet-papier
T. RAZAFINDROVONA

M2017/01: Exploitation de l'enquête expérimentale Qualité de vie au travail
T. RAZAFINDROVONA

M2018/01: Estimation avec le score de propension sous R
S. QUANTIN

M2018/02: Modèles semi-paramétriques de survie en temps continu sous R
S. QUANTIN

M2019/01 : Les méthodes de décomposition appliquées à l'analyse des inégalités
B. BOUTCHENIK
E. COUDIN
S. MAILLARD

M2020/01 : L'économétrie en grande dimension
J. L'HOUR

M2021/01 : R Tools for JDemetra+ - Seasonal adjustment made easier
A. SMYK
A. TCHANG

M2021/02 : Le traitement du biais de sélection endogène dans les enquêtes auprès des ménages par modèle de Heckman
L. CASTELL
P. SILLARD

M2021/03 : Conception de questionnaires auto-administrés
H. KOUMARIANOS
A. SCHREIBER

M2022/01 : Introduction à la géomatique pour le statisticien : quelques concepts et outils innovants de gestion, traitement et diffusion de l'information spatiale
F. SEMECURBE
E. COUDIN

M2022/02 : Le zonage en unites urbaines 2020
V. COSTEMALLE
S. OUJIA
C. GUILLO
A. CHAUVET

M2023/01 : Les réseaux de neurones appliqués à la statistique publique : méthodes et cas d'usages
D. BABET
Q. DELTOUR
T. FARIA
S. HIMPENS

M2023/02 : Redressements de la première vague de l'enquête epicov : un exemple de correction des effets de sélection dans les enquêtes multimodes
L.CASTELL
C. FAVRE-MARTINOZ
N. PALIOD
P. SILLARD

M2023/03 : Appariements de données individuelles : concepts, méthodes, conseils
L.MALHERBE

M2023/04 : Victimations déclarées et effets de mode : enseignements de l'expérimentation panel multimode de l'enquête cadre de vie et sécurité
L. CASTELL
M. CLERC
D. CROZE
S. LEGLEYE
A. NOUGARET

M2024/01 : Estimation en temps réel de la tendance-cycle : apport de l'utilisation des filtres asymétriques dans la détection des points de retournement
A.QUARTIER-LA-TENTE

M2024/02 : La disponibilité des coordonnées de contact dans fidéli-nautile - quels enseignements pour les protocoles de collecte ?
G. CHARRANCE (INED)

M2024/03 : Discuter l'existence d'un effet de sélection dans un cadre multimode grâce à une analyse de sensibilité - Application aux enquêtes annuelles de recensement
L. COURT
S. QUANTIN

M2024/04 : Vers une désaisonnalisation des séries temporelles infra-mensuelles avec JDemetra+
A. SMYK
K. WEBEL

M2025/01 : Les estimations par capture-recapture ou par système multiple : quelques éléments théoriques
P. ARDILLY
H. KOUMARIANOS

M2025/02 : Tests cognitifs pour les enquêtes auto-administrées : quelques éléments de méthode
D. GUILLEMOT
J. DIRAND
C. FLUXA

M2025/03 : statistiques fondées sur des données administratives - esquisse d'un cadre général
H. KOUMARIANOS
P. RIVIÈRE