L'intégration des données administratives dans un processus statistique

Industrialiser une phase essentielle

F. Cotton, O. Haag***

La statistique publique a de plus en plus recours à des sources externes, en particulier à des données administratives, pour produire des statistiques. Ceci nécessite d'industrialiser davantage les processus de production et notamment le processus d'intégration de ces données, afin de sécuriser, d'assurer une meilleure traçabilité et de rendre cette intégration la plus reproductible possible.

L'objectif du statisticien public est de mettre en place un cadre général d'intégration de données permettant une démarche automatisée sur des données structurées, livrées par un producteur externe. Plus précisément, il s'agit d'implémenter un pipeline jalonné de points de contrôle qui permettent de s'assurer que la succession des tâches (renommer, restructurer les données, recoder, pseudonymiser, etc.) se déroule correctement et d'arrêter le processus dès qu'un éventuel problème est rencontré. En outre, l'utilisation de standards et de métadonnées actives le long de ce pipeline permettent au concepteur d'être le plus autonome possible et ainsi de pouvoir l'adapter plus facilement aux évolutions des sources externes.

Official statistics increasingly rely on external sources, particularly administrative data, to produce statistics. This requires greater industrialisation of the production and integration processes for these data, in order to make them more secure, more traceable and as reproducible as possible.

The aim is to implement a general data integration framework based on an automated approach to structured data delivered by an external producer. This involves setting up a pipeline with checkpoints to ensure that the succession of tasks (renaming, data restructuring, recoding, pseudonymisation, etc.) is carried out correctly and to stop the process as soon as any problems are encountered. In addition, the use of standards and active metadata upstream of this pipeline allows the designer to be as autonomous as possible, making it easier to adapt to changes in external sources.

* Expert, DSI, Insee,
franck.cotton@insee.fr
** Directeur du programme « Répertoire statistiques d'individus et de logements », DSDS, Insee,
olivier.haag@insee.fr

La statistique publique produit des chiffres et des études à partir d'enquêtes qu'elle gère de façon totalement autonome, de la conception à la diffusion après les étapes de collecte et de traitement de la non-réponse. Toutefois, avec la profusion de données produites et mises à disposition par d'autres organismes, le recours massif à des sources externes, et notamment à des données administratives, pour produire des statistiques augmente fortement *(Lamarche et Lollivier, 2021, Hand 2018)*. Cette pratique se développe également au niveau international (*UNECE Integration, statswiki, Cros 2014*). De ce fait, il y a aujourd'hui une forte volonté d'industrialiser davantage l'intégration de sources administratives afin de sécuriser, d'assurer une meilleure traçabilité et de rendre le plus reproductible possible cette intégration de données.

> ***Une forte volonté d'industrialiser davantage l'intégration de sources administratives afin de sécuriser, d'assurer une meilleure traçabilité et de rendre le plus reproductible possible cette intégration de données.***

Bien souvent, la donnée administrative n'est pas utilisable en tant que telle *(Courmont, 2021)*. L'intégration consiste alors à détacher les données et les métadonnées de leur univers de gestion d'origine pour les attacher au monde statistique (unité statistique, concepts, nomenclature, etc.). Cette opération nécessite une appropriation de la source par ses futurs utilisateurs.

Lorsque la qualité de la donnée administrative est jugée suffisante pour produire des statistiques fiables (« qualification de la source »), la première étape du processus consiste à intégrer les informations externes dans le système d'information statistique. À l'instar du processus de collecte par enquête, l'intégration est la première étape des traitements statistiques (contrôle, validation, redressement) qui permettent de passer du statut de données brutes à celui de données diffusables.

Les objectifs, le périmètre exact et les étapes de ce processus d'intégration des données sont présentés dans cet article. La phase de qualification n'est abordée ici que brièvement bien qu'elle soit essentielle. En effet les sources administratives peuvent présenter des défauts de couverture (exemple de données absentes en Alsace-Lorraine car définies après le régime concordataire) ou utiliser des concepts éloignés de ceux de la statistique publique (par exemple : données collectées au niveau de compteurs électriques pour facturer la consommation d'un logement) : il est nécessaire de bien identifier ces problèmes en amont afin de les traiter au plus tôt.

La qualification de la source : un pré-requis indispensable avant l'intégration

Avant d'intégrer une source, il convient de s'assurer qu'elle dispose des qualités suffisantes pour pouvoir être utilisée à des fins statistiques. Ainsi, il est nécessaire d'échanger avec le producteur de la donnée afin de vérifier que la source est :

- exploitable (les données contenues peuvent être restructurées pour mesurer des concepts statistiques) ;

- complète (aucune sous-couverture évidente qui empêcherait son exploitation) ;
- disponible dans un délai raisonnable ;
- documentée (présence de métadonnées).

À noter que l'office national statistique anglais envoie des statisticiens directement dans les administrations pour réaliser cette phase de qualification *(Fermor-Dunman et Parsons 2022).*

▶ De quoi parle-t-on ?

Les données administratives sont recueillies et structurées par des organismes pour leurs besoins propres. Par exemple, la déclaration sur le revenu est collectée au niveau du foyer fiscal (unité gérée par la DGFiP[1]). Cette notion désigne l'ensemble des personnes inscrites sur une même déclaration de revenus et ne correspond pas directement aux concepts d'individu et de ménage utilisés pour la statistique. En effet un foyer fiscal peut regrouper un ou plusieurs individus et il peut y avoir plusieurs foyers fiscaux dans un seul ménage : par exemple, un couple non marié compte pour deux foyers fiscaux si chacun remplit sa propre déclaration de revenus. Il faut donc restructurer cette information administrative de base pour l'exploiter à des fins statistiques.

> ***La phase d'intégration transforme la donnée administrative individuelle en donnée individuelle brute statistiquement exploitable.***

La phase d'intégration transforme la donnée administrative individuelle en donnée individuelle brute statistiquement exploitable. Elle sélectionne uniquement les informations indispensables des sources externes, respectant ainsi les principes de minimisation et de nécessité auxquels les statisticiens sont tenus. On change de gouvernance en passant du monde administratif au monde statistique. Les données pourront être modifiées par le statisticien sans en informer le producteur initial. Par ailleurs, ces données deviennent alors des données statistiques et sont donc soumises au secret statistique.

Cette phase permet de constituer un ensemble cohérent à partir de plusieurs fichiers différents, voire de plusieurs sources différentes. Un exemple très intéressant à cet égard est celui de Fidéli[2]. Ce fichier intègre plusieurs sources fiscales : une source décrivant les membres des foyers imposables, une source contenant les informations sur l'impôt sur le revenu de ces foyers, une source contenant les informations relatives à la taxe d'habitation et une autre à la taxe foncière, chacune pouvant être constituée de plusieurs fichiers, plus d'une centaine par source pour obtenir une information sur la France entière.

La phase d'intégration produit la version de base de la donnée qui alimentera la suite du processus. À ce stade, il s'agit seulement de réorganiser l'information contenue dans les fichiers administratifs mais en aucun cas de les contrôler et encore moins de les

1 Direction générale des finances publiques : direction de l'administration publique centrale française qui dépend du ministère chargé de l'économie.
2 Fichier démographique sur les logements et les individus.

corriger. Néanmoins, un effort de traçabilité est nécessaire pour garantir la reproductibilité de cette base de départ.

L'intégration, point de départ du processus statistique, voire point de reprise en cas de problème, est également un point de référence. En comparant la donnée brute à la donnée modifiée par des gestionnaires dans la phase de « contrôle et corrections manuelles et automatiques », il est possible de juger de l'efficience des contrôles mis en place. De même, en comparant les données brutes aux données issues de redressements automatiques, on peut mesurer la variance de ces traitements et le gain en qualité, pour différents indicateurs, notamment en matière de réduction de biais.

L'intégration, une étape fondamentale à réaliser pour produire des statistiques

Si l'on compare les cycles de vie de la donnée d'une production statistique réalisée à partir d'une enquête et celle réalisée à partir d'une source externe, on constate qu'ils diffèrent sur la phase amont mais qu'ils sont identiques à partir des traitements de transformation des données brutes statistiques en données diffusables.

En référence au GSBPM[3] *(UNECE, GSPBM)*, l'intégration de données précède les phases aval de traitement des données et d'analyse conduisant aux données diffusables (contrôle, correction manuelles et automatiques, validation des données individuelles, agrégation et validation des données agrégées, diffusion des données et archivage des données). Ces phases ne sont pas évoquées dans cet article.

> ***L'intégration est la première étape du processus.***

Ainsi, si le fichier administratif contient des valeurs manifestement erronées (30 février par exemple), ces valeurs ne seront pas modifiées durant la phase d'intégration des données mais corrigées par une valeur possible lors de la phase suivante de « contrôle, correction manuelles et automatiques et validation des données individuelles ».

L'intégration est la première étape du processus. C'est aussi le moment d'initialiser les métadonnées statistiques qui seront utiles tout au long du cycle de vie de la donnée (concepts, description des variables, listes de codes et nomenclatures). Ce point s'inscrit dans une logique de pilotage des processus par les métadonnées dites « actives[4] ». ***(cf. infra)***

3 GSBPM pour *Generic Statistical Business Process Model*, modèle générique de description des processus de production statistique.

4 Les métadonnées peuvent aller au-delà de leur rôle de description, *via* une interface dédiée ou des applications clientes. Ces outils permettent de tirer parti du caractère exhaustif et normalisé des métadonnées pour produire automatiquement des composants du processus statistique. Les métadonnées acquièrent ainsi un statut nouveau, passant du stade d'« informations facilitant la compréhension des statistiques », au stade de « données participant au processus de production » ; d'où l'idée de métadonnées actives (*Bonnans, 2019*).

L'intégration : phase qui distingue les enquêtes des statistiques produites à partir de données administratives

L'enquête statistique est prise en charge par les statisticiens dès le début et permet de s'assurer, lors de la conception du questionnaire, que les données collectées répondent à des concepts et unités statistiques. Elles n'ont donc pas besoin d'être transformées après leur collecte et sont immédiatement des données statistiques brutes exploitables.

A contrario, les données externes doivent être transformées (constituer des unités statistiques, renommer, etc.) avant de pouvoir être considérées comme des données brutes statistiques. Certaines métadonnées sont initialisées durant cette phase d'intégration, telles que la définition des variables, leurs règles de calcul, leurs liens avec les concepts statistiques, etc.

À l'Insee, il existe des processus statistiques qui associent des données d'enquêtes et des données administratives afin de produire des statistiques robustes tout en limitant la charge statistique, à savoir le temps que les répondants doivent consacrer à l'enquête. On peut citer l'enquête « Revenus Fiscaux et Sociaux » qui fusionne des données de l'enquête « Emploi » avec des données fiscales et sociales, ou encore le système d'élaboration des statistiques annuelles d'entreprises (ESANE) fondé sur les résultats des enquêtes sectorielles annuelles et sur les liasses fiscales et les données de la déclaration sociale nominative (DSN).

Description des différentes étapes du processus d'intégration

Les données administratives sont souvent liées à des politiques publiques, dont les contours juridiques peuvent évoluer rapidement au cours du temps (ajout ou suppression d'un impôt ou d'une prestation sociale, etc.). Il est donc indispensable que les outils et méthodes développés dans le cadre de cette phase d'accueil des données soient les plus transparents et les plus adaptables possibles.

> ***La transformation de la donnée administrative en donnée statistique brute est composée d'une succession de tâches élémentaires.***

L'intégration est une succession d'opérations élémentaires écrites de façon déclarative, c'est-à-dire en spécifiant le quoi et non le comment, qu'il est aisé de « rejouer » permettant d'obtenir une chaîne de traitement pérenne et réutilisable. Pour mettre en œuvre cette reproductibilité et assurer son adaptabilité, il est indispensable de s'appuyer sur des standards et d'outiller le statisticien pour qu'il soit le plus autonome possible. Le langage VTL est une solution permettant de répondre à ces deux besoins majeurs. La transformation de la donnée administrative en donnée statistique brute est composée d'une succession de tâches

élémentaires (de façon analogue à une composition de fonctions en mathématiques) qui peuvent se décomposer selon les six catégories suivantes :

- renommer les variables sélectionnées ;
- restructurer les données par unité statistique ;
- recoder ;
- calculer les variables dérivées ;
- filtrer les enregistrements utiles ;
- pseudonymiser.

Cet ordre est indicatif et peut différer d'une source à l'autre. En particulier, les phases de pseudonymisation et de filtrage peuvent intervenir à différents moments de la séquence.

Cette catégorisation s'inspire du service d'accueil des sources mis en place par le programme Résil (Répertoire statistique des individus et des logements) *(Durr et alii, 2022)* qui va jouer ce rôle d'intégration de données administratives dans le système d'information démographique et social en utilisant l'outil Accueil Réception Contrôle ARC ***(encadré 1)***.

▶ Encadré 1 : Le module Accueil Réception Contrôle (ARC)

Le module Accueil Réception Contrôle (ARC) a été conçu comme un bloc d'infrastructures mutualisées pour l'ensemble des déclarations, mobilisé par le Système d'Information sur l'Emploi et les Revenus d'Activités. Il a été développé en *open source*, ce qui facilite sa mutualisation Il est aujourd'hui utilisé par d'autres systèmes à l'Insee et même en dehors, par l'Istat (Institut de statistique italien) notamment.

Ce module fonctionnel permet l'accueil, le contrôle et la transformation des données administratives en données statistiques élémentaires, avec une possibilité de filtrage en amont.

Le paramétrage du chargement (fichier plat ou fichier XML*), la reconnaissance des normes associées à des déclarations, les contrôles appliqués aux données et la transformation des données administratives en données statistiques (le *mapping*) peuvent être modifiés et améliorés au cours du temps de façon interactive par l'utilisateur. Étant donné les volumes de données très importants à traiter, la possibilité de modifier les traitements de façon interactive ainsi que d'éventuelles erreurs de spécification pourraient entraîner des risques sur la production.

**XML : Extensible Markup Language.*

Pour limiter ce risque, les utilisateurs disposent d'espaces de tests similaires mais distincts de l'espace de production, afin de tester et qualifier les nouvelles règles spécifiées et de mesurer leur impact sur les données chargées avant leur mise en production.

Pour que le module ARC puisse réceptionner et contrôler ces différents types de fichiers, il faut lui spécifier au préalable trois informations indispensables sur les caractéristiques des fichiers à charger et contrôler :

1. la famille spécifique de rattachement du fichier administratif : (exemple : DSN, Particuliers employeurs, etc.) ; ces familles distinguent les grands types de fichiers à contrôler.
2. la norme de rattachement du fichier administratif. La norme est un ensemble de caractéristiques décrivant le fichier (sous-ensemble de la famille), le plus souvent valable durant une période précise et qui permet d'étalonner les tests.
3. la périodicité du fichier reçu ; elle renseigne l'application sur la temporalité des fichiers devant faire l'objet du contrôle.

La suite de l'article sera illustrée à partir d'un exemple de fichier administratif fictif utilisé par l'Insee. Cet exemple est simpliste, mais permet de se rendre compte des différentes transformations possibles et d'illustrer la mise en place d'un « pipeline » pour l'intégration d'une source.

Imaginons que le contenu des données envoyées à l'Insee par fichier au format CSV (avec séparateur « ; ») soit le suivant :

▶ Exemple

- l'identifiant du foyer fiscal (**dirindik**) ;
- le taux d'imposition du foyer fiscal (**8OT**) ;
- l'adresse associée au foyer fiscal (**ZAFR1**) ;
- les nom (**LNCON**), prénom (**LNCOPF**), sexe (**LCCOT**), date de naissance (**DNCO**) et salaire (**1AJ**) du déclarant ;
- les nom (**LNCJN**), prénom (**LNEPP**), sexe (**LCPPT**), date de naissance (**DAEPAD**) et salaire (1BJ) du conjoint ;
- le nombre d'enfants du foyer fiscal (**7EA**).

Le début du fichier pourrait ressembler à ce qui suit :

```
dirindik;8OT;ZAFR1;LNCON;LNCOPF;LCCOT;DNCO;1AJ;LNCJN;LNEPP;LCPPT;DAE-
PAD;1BJ;7EA
570001;11;18 rue des bleuets 57 000
Metz;DURANT;Jonathan;1;01/12/55;10500;DURANT;Germaine;2;27/05/58;7500;3
570002;30;11 rue des Lilas 57000
Metz;DUPUIS;Sophie;2;30/02/74;30000;;;;;;2
570003;41;523, rue du Général De
Gaulle;WENDLING;Zoé;2;23/07/74;50000;HENRY;Jérôme;;;25000;
570004;42;47 rue des Plantes 57 000
Metz;THIERY;Robert;1;17/08/80;1000000000;;;;;;52
540005;0;23 rue Jordan 54 000 Nancy;DURANT;Jonathan;1;01/12/55;;;;;;;0
```

Dans notre exemple, le salaire d'un individu n'est affecté qu'à un seul foyer fiscal, mais un individu peut être affecté à plusieurs foyers fiscaux, s'il possède par exemple une résidence secondaire dans un département différent de celui dans lequel il a déclaré son salaire : c'est ici le cas de Durant Jonathan.

Renommer les variables : une nécessité pour mieux les comprendre

> ***Pour exploiter les variables provenant d'une source administrative, il est plus facile qu'elles aient un nom compréhensible.***

Pour exploiter les variables provenant d'une source administrative, il est plus facile qu'elles aient un nom compréhensible (par exemple : `salaire_decl`) plutôt que de conserver le nom du fichier d'origine qui correspond par exemple au nom de la case de la déclaration fiscale (`1AJ`) dans l'exemple.

Les noms des variables des fichiers externes sont souvent liés au contexte dans lequel elles ont été collectées et il est souhaitable de les renommer dès cette phase afin de faciliter leur utilisation par des statisticiens (qui n'ont pas forcément participé directement à la phase d'intégration des données). Renommer les données pour en faciliter l'appropriation par le statisticien n'est toutefois pas suffisant et s'accompagne de la saisie de métadonnées descriptives pour décrire plus précisément le concept, le format des variables, etc. Les nouveaux noms des variables se rapprochent alors des concepts statistiques qu'elles recouvrent. Ainsi dans notre exemple :

Exemple

• `dirindik` devient `id_foyer_fisc`	• `8OT` devient `tx_imposition`
• `ZAFR1` devient `adresse`	• `LNCON` devient `nom_decl`
• `LNCOPF` devient `prenom_decl`	• `LCCOT` devient `sexe_decl`
• `DNCO` devient `d_nais_decl`	• `1AJ` devient `salaire_decl`
• `LNCJN` devient `nom_conj`	• `LNEPP` devient `prenom_conj`
• `LCPPT` devient `sexe_conj`	• `DAEPAD` devient `d_nais_conj`
• `1BJ` devient `salaire_conj`	• `7EA` devient `nb_enfants`

Restructurer les données par unité statistique et les relier

Les unités de gestion administratives (compteur électrique, foyer fiscal) sont en général différentes des unités statistiques diffusées (logement, ménage, individu, entreprise, etc.). Il faut donc dériver ces unités statistiques si elles ne sont pas directement présentes dans la source administrative. Cette étape se traduit par l'agrégation de plusieurs enregistrements ou la suppression d'enregistrements pour créer une unité statistique. Par exemple, on peut regrouper des données à un niveau « ménage » à partir de données d'individu ou encore regrouper des données au niveau « individus » à partir de données de contrat de travail, etc. Ceci est possible si les données contenues dans la source externe le permettent grâce à la présence de l'identifiant du ménage dans les enregistrements d'individus, par exemple. Si un tel traitement nécessite des données externes à la source, cette opération se fera dans une phase ultérieure du processus.

Inversement, on peut vouloir éclater un enregistrement du fichier en entrée en plusieurs enregistrements dans le modèle de données statistiques (cas des individus dans les fichiers fiscaux structurés en foyers fiscaux). Il est alors nécessaire d'expliciter les liens entre différentes unités statistiques. Dans le cadre des données fiscales, faire le lien entre individus et foyers fiscaux d'une part, et foyers fiscaux et adresses d'autre part, permet *in fine* de constituer des ménages.

Dans l'exemple, le fichier envoyé « mélange » deux types d'unités statistiques au sein d'un même enregistrement.

Dans une première étape, on créerait deux tables différentes, individus et foyers fiscaux, dans le modèle de données « statistique ».

Les données sont intégrées dans les deux tables suivantes :

Table 1 : La table des foyers fiscaux

ID	TAUX IMPOSITION	ADRESSE	NB ENFANTS
570001	11	18 rue des bleuets 57000 Metz	3
570002	30	11 rue des Lilas 57000 Metz	2
570003	41	523, rue du Général De Gaulle	
570004	42	47 rue des Plantes 57000 Metz	52
540005	0	23 rue Jordan 54000 Nancy	0

Table 2 : La table des individus

NOM	PRÉNOM	SEXE	DATE DE NAISSANCE	SALAIRE	STATUT	ID FOYER
DURANT	Jonathan	1	01/12/55	10500	déclarant	570001
DURANT	Germaine	2	27/05/58	7500	conjoint	570001
DUPUIS	Sophie	2	30/02/74	30000	déclarant	570002
WENDLING	Zoé	2	23/07/74	50000	déclarant	570003
HENRY	Jérôme			25000	conjoint	570003
THIERY	Robert	1	17/08/80	1000000000	déclarant	570004
DURANT	Jonathan	1	01/12/55		déclarant	540005

Une jointure de ces deux tables sur la variable id_foyer présente dans les deux tables permet de reconstituer les données initiales. On ne change pas l'information, on la restructure.

Pseudonymiser les données

Les sources administratives peuvent contenir des informations nominatives (NIR[5], nom, prénoms etc.) inutiles à la production statistique et qu'il faut supprimer dès cette phase afin de respecter la confidentialité des données au plus tôt dans le processus de production. On pseudonymise. Pour séparer au plus tôt les données d'état civil (nom, prénom, date et lieu de naissance et adresse) des données « métier » utiles pour produire des statistiques (information sur l'emploi, le revenu, etc.), on apparie les données d'état civil de la source avec un référentiel d'identité pour remplacer, dans les données intégrées, les données d'état civil par l'identifiant leur correspondant dans ce référentiel[6].

La mise en œuvre de cette pseudonymisation dans notre exemple conduit à scinder la table des individus en deux : une table nommée « individu_etat_civil » contenant les données d'état civil et une deuxième (« individu_salaire ») contenant les autres variables. Le lien entre les deux tables se fait par le biais de la variable id_ind qui identifie un individu de façon univoque selon son nom, son prénom et sa date de naissance. On constate que la table individu_etat_civil ne contient que 6 lignes car DURANT Jonathan est en double dans le fichier.

Table 1 : La table individus_etat_civil

ID_IND	NOM	PRÉNOM	SEXE	DATE DE NAISSANCE
1	DURANT	Jonathan	1	01/12/55
2	DURANT	Germaine	2	27/05/58
3	DUPUIS	Sophie	2	30/02/74
4	WENDLING	Zoé	2	23/07/74
5	HENRY	Jérôme		
6	THIERY	Robert	1	17/08/80

Table 2 : La table individu_salaire (pseudonymisée)

ID_IND	SALAIRE	STATUT	ID FOYER
1	10500	déclarant	570001
2	7500	conjoint	570001
3	30000	déclarant	570002
4	50000	déclarant	570003
5	2500	conjoint	570003
6	1000000000	déclarant	570004
1		déclarant	540005

5 Numéro d'identification au répertoire national des personnes physiques (RNIPP), plus connu comme le « numéro de sécurité sociale ».
6 Voir l'article de Yves-Laurent Bénichou, Lionel Espinasse et Séverine Gilles sur le Code statistique non signifiant (CSNS) dans ce même numéro.

Recoder les valeurs des variables

> ***Les sources administratives peuvent utiliser des nomenclatures différentes de celles utilisées pour la statistique.***

Les sources administratives peuvent utiliser des nomenclatures différentes de celles utilisées pour la statistique. Cette étape permet de se conformer à ces dernières à condition que cette transformation soit totalement automatisée (passage de « H » à 1 pour le sexe par exemple). Pour les enregistrements où une intervention humaine complémentaire est nécessaire (codification de la PCS[7] par exemple), un indicateur de reprise peut être posé durant l'intégration et le recodage « manuel » sera reporté à une phase ultérieure du processus de production.

Ce traitement concerne directement les valeurs d'une ou plusieurs variables du fichier en entrée. Il peut s'agir :

- d'une harmonisation des valeurs manquantes ou « refuge »[8] dans le cas où les valeurs manquantes sont traitées de façon différente selon les variables. Ceci facilitera leur repérage dans la phase de redressement[9] ;
- d'une correction des valeurs aberrantes. Ce processus peut être reporté dans les phases aval du processus statistique mais s'il est mis en œuvre dès l'intégration, il doit pouvoir être fait sans intervention humaine. Dans l'exemple, il s'agirait de traitement simple comme, par exemple, la mise à valeur manquante d'une date erronée (30 février) ;
- de la normalisation d'une variable. Il peut s'agir de supprimer des caractères spéciaux, du passage en lettres majuscules, etc. Ce traitement peut concerner des variables utiles pour l'identification afin de s'assurer que les règles de normalisation sont les mêmes dans le fichier à identifier et le référentiel auquel il doit être confronté. De tels traitements sont mis en œuvre pour l'attribution du Code Statistique Non Signifiant par exemple.

Calculer des variables dérivées

Pour ses études, le statisticien utilise souvent des nomenclatures (tranche d'âge, catégorie d'entreprises etc.). Le calcul de variables dérivées permet, par exemple, de passer d'une année de naissance contenue dans la source administrative à une tranche d'âge.

> ***Créer des nouvelles variables qui vont être utiles dans les phases ultérieures.***

Il s'agit de créer des nouvelles variables qui vont être utiles dans les phases ultérieures. Par exemple, créer une variable statistique par agrégation ou éclatement de variables en entrée. Le revenu peut être défini comme la somme de différents postes de la liasse fiscale, l'adresse peut être décomposée en plusieurs champs, etc.

7 PCS : nomenclature des professions et catégories socioprofessionnelles.
8 Le jour de naissance est par exemple mis à 00 plutôt que laissé à blanc.
9 Il faut bien distinguer cette étape, qui reste au niveau de la représentation des variables, des traitements d'imputation qui viendront plus tard dans le processus et qui eux s'appuient sur un modèle statistique. Imaginons que pour une variable donnée, il y ait plusieurs façons de représenter la valeur manquante (un blanc, un 00 etc.). Lors de l'intégration, ces valeurs seront harmonisées (toutes mises à blanc par exemple). Leur imputation n'interviendra qu'ultérieurement.

Dans l'exemple, des recodages et des créations de variables dérivées automatiques ont été faits. Ils ne demandent aucune intervention humaine :

- codage du sexe en : 1 = « M » et 2 = « F », autre = « blanc » ;
- calcul d'une nouvelle variable (salaire_foyer) dans la table foyer_fiscaux et qui correspond à la somme des salaires des déclarants et conjoints ;
- calcul d'une variable permettant de juger de la vraisemblance du salaire du foyer. Cette variable vaut 1 si le salaire du foyer est inférieur à 10 millions d'€ et vaut 0 sinon. À noter qu'aucune correction n'est effectuée lors de cette phase du processus. Les redressements manuels ou automatiques seront réalisés dans les phases ultérieures du traitement. Ce calcul n'est pas forcément appliqué à toutes les variables. Dans l'exemple, on laisse la valeur aberrante des 52 enfants. Cette valeur sera corrigée dans les traitements postérieurs à la phase d'intégration ;
- calcul du nombre de membres du foyer, qui est égal au nombre d'enfants + 2 si un conjoint est présent et au nombre d'enfants + 1 sinon ;
- calcul d'une variable « unique » pour les individus. Cette variable vaut 1 si l'identifiant de l'individu n'est présent qu'une fois dans la table individus_salaire et 0 sinon.

Les tables obtenues à l'issue de cette phase d'intégration sont les suivantes :

Table 1 : La table foyers_fiscaux

ID	TAUX IMPOSITION	ADRESSE	NB ENFANTS	NB FOYER	SALAIRE	SALAIRE COHÉRENT
570001	11	18 rue des bleuets 57000 Metz	3	5	18000	1
570002	30	11 rue des Lilas 57000 Metz	2	3	30000	1
570003	41	523, rue du Général De Gaulle	0	2	75000	1
570004	42	47 rue des Plantes 57000 Metz	52	53	1000000000	0
570005	0	23 rue Jordan 54000 Nancy	0	1		1

Table 2 : La table individus_etat_civil

ID IND	NOM	PRÉNOM	SEXE	DATE DE NAISSANCE
1	DURANT	Jonathan	M	01/12/55
2	DURANT	Germaine	F	27/05/58
3	DUPUIS	Sophie	F	30/02/74
4	WENDLING	Zoé	F	23/07/74
5	HENRY	Jérôme		
6	THIERY	Robert	M	17/08/80

Table 3 : La table individu_salaire (pseudonymisée) :

ID IND	SALAIRE	STATUT	ID FOYER	UNIQUE
1	10500	déclarant	570001	non
2	7500	conjoint	570001	oui
3	30000	déclarant	570002	oui
4	50000	déclarant	570003	oui
5	2500	conjoint	570003	oui
6	1000000000	déclarant	570004	oui
1		déclarant	540005	non

Mettre en place des filtres

> ***Le champ de la source en entrée peut être beaucoup plus vaste que le champ d'intérêt statistique.***

Le champ de la source en entrée peut être beaucoup plus vaste que le champ d'intérêt statistique. Dans ce cas, si la source contient des variables permettant de filtrer la population d'intérêt, il est possible de le faire dès l'étape d'intégration afin d'alléger les bases statistiques. Ceci permet également de respecter les principes de minimisation et de nécessité qui obligent tout statisticien à se limiter autant que possible aux données dont il a un réel besoin. Par exemple, supprimer les locaux commerciaux d'un fichier administratif lorsqu'on ne s'intéresse qu'aux locaux d'habitation.

Dans l'exemple, les lignes des individus en double ayant un montant de salaire égal à valeur manquante ne sont pas intéressantes. La variable unique est recalculée sur cette nouvelle table.

Au final la table individu_salaire est donc la suivante :

Table : La table individu_salaire

ID IND	SALAIRE	STATUT	ID FOYER	UNIQUE
1	10500	déclarant	570001	non
2	7500	conjoint	570001	oui
3	30000	déclarant	570002	oui
4	50000	déclarant	570003	oui
5	2500	conjoint	570003	oui
6	1000000000	déclarant	570004	oui

▶ Construire le *pipeline* de données

La phase d'intégration des sources doit être totalement automatisée et reproductible. Elle consiste à mettre en place un cadre général pour une démarche systématique sur des données structurées par un producteur externe. Elle doit se traduire par la mise en place d'un *pipeline* jalonné de points de contrôle pour s'assurer que la succession des tâches se déroule correctement et d'arrêter le processus dès qu'un éventuel problème est rencontré.

L'objectif de cette industrialisation est de mettre en place l'enchaînement de traitements génériques paramétrables les plus indépendants possibles de la source en entrée. L'intégration d'une nouvelle source (ou la mise à jour d'une source existante) revient alors à décrire la source en entrée, le modèle de données souhaité en sortie et les transformations permettant de passer de l'un à l'autre. L'enchaînement des étapes d'intégration des données se fait ensuite de façon automatique à partir de ces métadonnées qui deviennent actives. Ainsi, la spécification de la représentation attendue pour une variable (M, F pour la variable sexe par exemple) permet de générer automatiquement des traitements de contrôle des valeurs.

Dans ces conditions, certaines métadonnées plus précises ont un intérêt particulier dans le cadre des transformations de données. En amont, la documentation des traitements peut aller jusqu'à leur spécification dans un langage plus ou moins formel, par exemple BPMN[10] ou graphe acyclique direct (DAG)[11] pour le processus d'ensemble, SQL ou VTL ***(encadré 2)*** pour les transformations elles-mêmes. Cela présente l'avantage de permettre l'automatisation complète ou partielle du processus dans une démarche de métadonnées actives. La spécification en VTL est une métadonnée (elle décrit la transformation dans un langage compréhensible par statisticien) qui est ensuite implémentée automatiquement à l'aide d'outils informatiques ce qui la rend donc directement active. En aval, le traçage précis des opérations (quelles variables ont servi pour le calcul de telle autre, par exemple), permet une meilleure maîtrise du processus et favorise la reproductibilité et la transparence. On parle de métadonnées de provenance ou de lignage *(data lineage)*.

> ***Ces métadonnées actives permettent au concepteur d'être le plus autonome possible et d'adapter plus facilement son pipeline.***

Ces métadonnées actives permettent au concepteur d'être le plus autonome possible et d'adapter plus facilement son *pipeline* aux évolutions des sources externes. Par exemple, la DGFiP a récemment modifié sa façon d'enregistrer les dépendances des maisons dans son système d'information ; par le passé, elles étaient liées à un local principal, maintenant elles sont considérées comme des logements à part entière. Une telle mise à jour conduit donc à revoir le filtre de définition du champ, voire à ajouter de nouvelles variables. Par le biais des métadonnées actives, le statisticien peut créer une nouvelle variable dans le modèle de données d'accueil de RMéS[12] (*Bonnans, 2019*) qui permet de repérer les dépendances, d'établir le mode de calcul de cette nouvelle variable dans ARC à partir des variables du fichier de la DGFiP et de modifier son script

10 Le Business Process Model and Notation (BPMN) est la norme standard pour la modélisation de processus métier.
11 Un graphe acyclique direct ou orienté (directed acyclic graph ou DAG) est une façon standard de décrire des processus.
12 Référentiel de métadonnées statistiques.

VTL de définition du champ pour prendre en compte cette nouvelle variable. Le statisticien est alors totalement autonome pour modifier son *pipeline*, prendre en compte la modification de structure du fichier en entrée et les tracer.

Dans le *pipeline* relatif aux différentes étapes de l'intégration des données ***(figure)***, le statisticien saisit en amont les métadonnées (définition des variables, attachement à un concept statistique, règle de transformation, etc.). Les étapes s'enchaînent ensuite automatiquement à partir de ces métadonnées, conduisant à une standardisation et une séparation claire entre spécification logique et réalisation technique et permettent une plus grande autonomie du statisticien.

En sortie de ce processus, les données administratives sont donc intégrées dans le système d'information sous forme de données brutes. En outre, une évaluation de la qualité est également produite en parallèle afin de disposer de premiers indicateurs qualité (nombre d'enregistrements, taux de non-réponse partielle, etc.) qui permettent de donner des premiers éléments sur la qualité des données brutes.

▶ Encadré 2 : VTL, langage de validation et de transformation de données

VTL (*Validation and transformation language*) est un langage permettant de spécifier des traitements de validation ou de transformation de données développé dans le cadre du standard SDMX*, norme d'échange de données et métadonnées statistiques. Destiné aux statisticiens, il fournit une vue neutre (indépendante de l'implémentation technique) du processus de données au niveau métier. En tant que langage de spécification, VTL est suffisamment riche et expressif pour définir des traitements relativement compliqués.

VTL possède des caractéristiques qui le rendent particulièrement intéressant dans un contexte d'industrialisation et d'automatisation des processus statistiques. VTL est donc adéquat pour l'intégration de sources externes.

Tout d'abord, comme VTL se positionne au niveau logique, intermédiaire entre le concept et l'implémentation, il n'est pas directement exécutable comme peuvent l'être Java, R ou Python. Les expressions VTL doivent être transmises à un moteur qui l'exécutera sur une plate-forme de plus bas niveau, par exemple Java, Python ou C#. Ceci permet une claire séparation des préoccupations (ou **séparation des responsabilités****) entre le statisticien qui se concentre sur la spécification des traitements et l'informaticien qui se charge de l'implémentation. Dans les langages directement exécutables, la formulation logique du traitement est souvent noyée dans les détails d'implémentation et difficile à reconstituer. Avec VTL, la spécification est traitée comme un objet en soi, qui peut donc être géré, versionné, tracé, partagé, documenté, etc.

Une autre propriété intéressante de VTL est d'être basée sur un modèle de données qui dérive des standards internationaux (GSIM, SDMX, DDI***) et qui est adapté à la statistique et à différents types de données (détaillées, agrégées, qualitatives, quantitatives, etc.). Au cœur du modèle se trouve le *Data Set*, composé de *components* (les colonnes dans un fichier tabulaire) jouant différents rôles (identifiants, mesures et attributs) et de lignes (*Data Points*). Ce modèle permet de simplifier les expressions : ainsi, si l'on effectue une somme sur un jeu de données, il n'est pas utile de préciser que l'opération ne s'applique qu'aux mesures et non aux identifiants ou aux attributs.

Enfin, VTL est décrit par une grammaire formelle, qui assure l'assise logique du langage et permet de l'exploiter de façon automatisée, notamment par la construction d'outils comme des éditeurs ou des moteurs d'exécution pour différentes plates-formes techniques. Ceci assure qu'une même expression sera exécutée de façon cohérente dans différents langages de plus bas niveau.

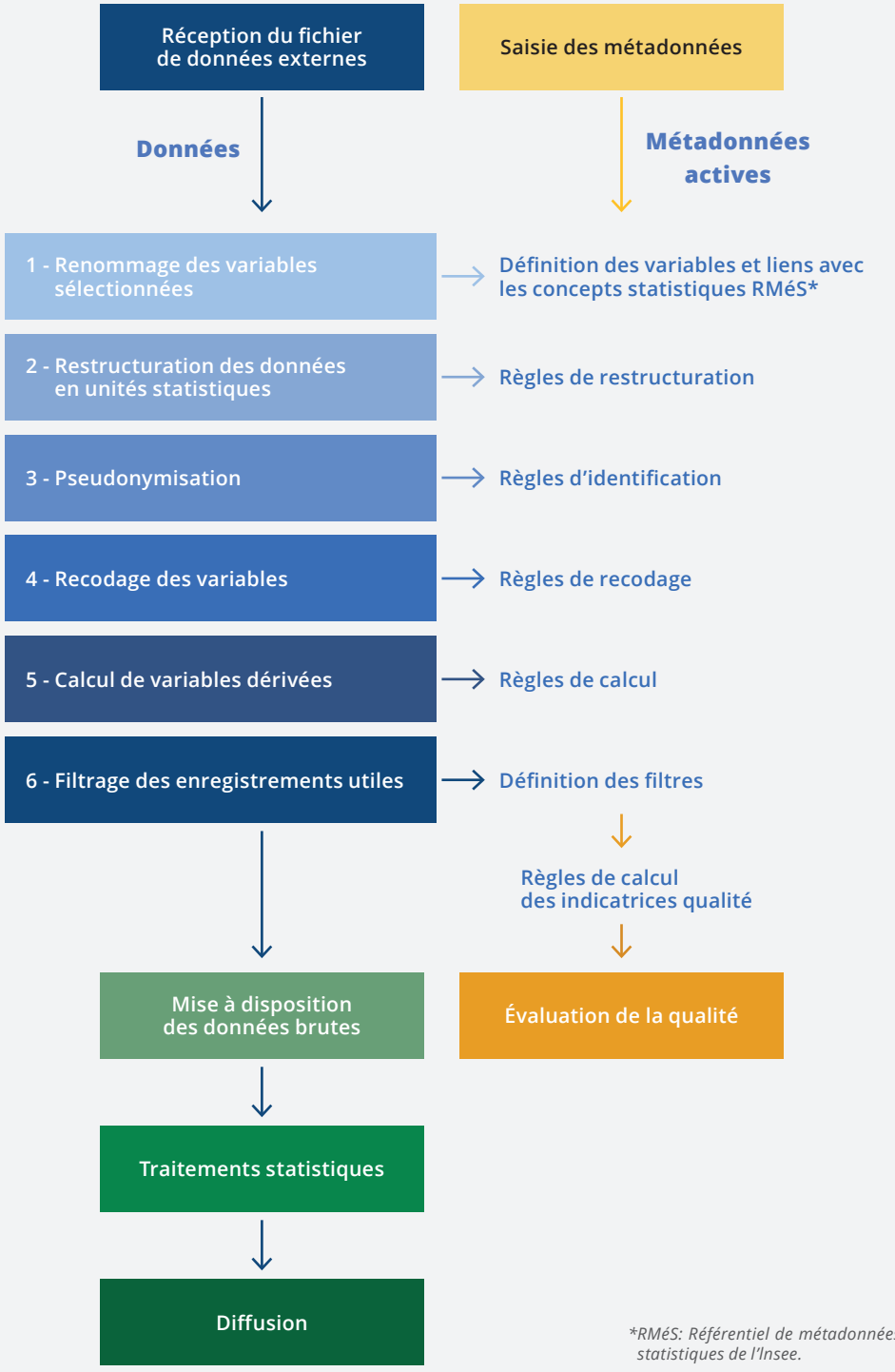
De nombreux outils ont été développés autour de VTL par les communautés statistiques et bancaires. Parmi les intervenants les plus actifs, on peut citer la Banque d'Italie, l'Insee ainsi que des sociétés privées.

* *SDMX pour Statistical Data and Metadata eXchange.*

** *https://fr.wikipedia.org/wiki/S%C3%A9paration_des_pr%C3%A9occupations.*

*** *GSIM pour Generic Statistical Information Model et DDI pour Data Documentation Initiative.*

Figure - Le pipeline de données et de métadonnées



RMéS: Référentiel de métadonnées statistiques de l'Insee.

▶ Intégrer les données c'est bien mais avec les métadonnées c'est mieux…

> ***L'existence de métadonnées complètes et exactes est fondamentale pour le bon déroulement des opérations.***

Pour les transformations de données comme pour toute étape du processus statistique, l'existence de métadonnées complètes et exactes est fondamentale pour le bon déroulement des opérations. Plus encore, comme les traitements de transformation se situent souvent au tout début du processus, il faut capturer ou créer les métadonnées qui seront réutilisées dans les étapes ultérieures.

En règle générale, les données administratives sont décrites avec leurs propres métadonnées. Certaines pourront être reprises mais pas forcément toutes. Et les métadonnées statistiques issues du processus d'intégration sont différentes de celles d'origine (les nomenclatures peuvent être différentes, les concepts statistiques associés sont spécifiques aux données statistiques, etc.).

Différents types de métadonnées revêtent une importance particulière dans le cadre de la transformation de données.

Les métadonnées peuvent, elles-mêmes, être l'objet d'une transformation pour accompagner la transformation des données. Par exemple, le renommage des variables peut être vu comme une transformation des métadonnées.

Les métadonnées de structure, à savoir la définition des variables utilisées, les concepts statistiques auxquels elles se raccrochent, le type et le domaine de valeurs, parfois contraints par des listes de codes, doivent être définies au plus tôt. Il est important qu'elles soient disponibles dans des formats qui permettent aisément d'en automatiser l'utilisation pour valider les traitements effectués. Par ailleurs, les métadonnées descriptives constituent une autre catégorie importante de métadonnées. On range ici tout ce qui permet de faciliter la découverte et la compréhension des données, d'évaluer leur qualité ou leur adéquation à une utilisation spécifique, etc. La documentation des traitements et des méthodes utilisées et les conditions de fourniture des données peuvent également être considérées comme des métadonnées descriptives.

▶ … avec une mesure de la qualité au plus tôt

> ***Il est important d'avoir une première mesure de la qualité des données dès leur intégration.***

Même si la source a été qualifiée en amont et que sa qualité a été jugée suffisante pour produire des statistiques, il est important d'avoir une première mesure de la qualité des données dès leur intégration (*Six et Kowarick, 2022, UNECE Statswiki Quality*). Ceci permet d'avoir une première idée de la qualité intrinsèque de la source et de pouvoir enclencher au plus tôt un processus d'amélioration continue de la qualité. Cette phase permet également de se familiariser avec les nouvelles sources en comprenant mieux leur structure et leur contenu.

Au niveau de l'intégration, cette mesure de la qualité (communément appelée *data profiling* dans la littérature internationale) consiste à définir des indicateurs qui permettront d'avoir une première évaluation de la qualité de la source et de suivre son évolution au cours du temps, en comparant les livraisons successives de la source (*monitoring*). Ces indicateurs sont par exemple :

- le nombre d'enregistrements reçus par type d'unité statistique ;
- les totaux de variables d'intérêt ;
- le taux de non-réponse partielle par variable ;
- la fréquence des modalités dans le cas de variables qualitatives, permettant d'identifier des modalités peu ou pas utilisées ;
- la distribution et mise en évidence de valeurs aberrantes ;
- l'identification des codes non utilisés pour les variables associées à une nomenclature ;
- l'identification de doublons, etc.

L'analyse de ces indicateurs nécessitera souvent, au moins lors des premières réceptions de fichiers, de demander au producteur des précisions. Ceci peut également conduire à ce que le producteur ajoute des contrôles dans son propre processus afin d'améliorer la qualité de la collecte. Par exemple les travaux menés à l'Insee lors de l'accueil de la déclaration sociale nominative (DSN) ont conduit le GIP-MDS[13] à ajouter un contrôle du NIR dans son interface de saisie des données de la DSN, ce qui a amélioré notablement la qualité du NIR dans la DSN. De tels indicateurs permettent de repérer au plus tôt si les données fournies sont incomplètes, lorsque le nombre d'enregistrements ou les totaux de certaines variables sont plus faibles que lors des livraisons précédentes. Ces indicateurs servent également à identifier des problèmes qui pourront être corrigés par les processus ultérieurs (exemple de l'indicateur qualité du salaire du foyer qui permet d'identifier le salaire d'un milliard d'euros qui est suspect). Ces indicateurs « qualité » seront donc utiles à l'utilisateur des données brutes pour piloter ses propres traitements.

Cette phase étant primordiale, des outils spécifiques ont été développés.

▶ Perspectives

> ***… changement de nature et de «propriétaire» de la donnée, qui devient donc statistique, avec le cadre légal, méthodologique et organisationnel correspondant.***

L'intégration des données administratives dans un processus statistique est une phase essentielle dont l'importance est reconnue au niveau international ***(encadré 3)***. Elle doit être soigneusement identifiée, spécifiée et outillée. Elle correspond à un changement de nature et de « propriétaire » de la donnée, qui devient donc statistique, avec le cadre légal, méthodologique et organisationnel correspondant.

13 Groupement d'intérêt public Modernisation des déclarations sociales.

Elle doit se traduire par l'implémentation d'un *pipeline* documenté et reproductible de transformations élémentaires ; il est souhaitable de le définir dans un formalisme indépendant d'une technologie particulière. Lors de la spécification de ce *pipeline*, il est important de prendre en compte les métadonnées, que ce soient celles de la source, qui peuvent elles-mêmes faire l'objet de transformations, ou celles qui découlent du processus d'intégration comme les métadonnées de provenance. Il faut également, comme pour tout processus, produire des indicateurs de qualité qui permettront de contrôler et d'améliorer les traitements.

Au-delà des données administratives, l'utilisation de données externes pour la statistique est appelée à se développer, et la mise en place d'un cadre méthodologique et d'outils communs pour l'acquisition de données permettra d'industrialiser cette fonction, à l'instar de ce qui a pu être fait pour la filière d'enquêtes à l'Insee (*Cotton et Dubois, 2019 ; Koumarianos et Sigaud, 2019*). Comme pour cette dernière, le pilotage des processus par l'activation des métadonnées, conduisant à une standardisation et une séparation claire entre spécification logique et réalisation technique, permettra une plus grande autonomie du statisticien et une meilleure réactivité des systèmes, importante pour s'adapter aux changements externes.

▶ Encadré 3 : Transformation de données, le cadre international

La collaboration internationale pour la **modernisation de la statistique officielle*** pilotée par l'Unece s'est initialement attachée à une vision orientée processus avec en particulier le modèle GSBPM, puis le moins connu GSDEM (2015). Les aspects relatifs aux données sont arrivés sur le devant de la scène dans le cadre de différentes initiatives, d'abord sous l'égide de l'Unece :

- projet «*Data Integration*» (2016) : il met l'accent sur l'intégration de données (après des réflexions sur le *Big data*), définie comme l'activité consistant à combiner au moins deux sources de données différentes dans un ensemble de données. La transformation des données n'y est pas définie comme une étape en soi, l'accent étant plutôt mis sur les méthodes d'appariement ;
- conception d'une architecture de données statistique commune (**CSDA**** 2017 – 2018) : son ambition est de « soutenir les organismes statistiques dans la conception, la collecte, l'intégration, la production et la diffusion de statistiques officielles basées à la fois sur les types de sources de données traditionnelles et nouvelles. » Elle s'appuie sur des **principes** courants dans les démarches modernes orientées données (donnée en tant qu'actif, accessibilité, ré utilisabilité, emploi de modèles standard), et définit les «capacités» (*capabilities*) nécessaires pour utiliser et gérer données et métadonnées statistiques. La transformation des données est une des capacités de haut niveau ;
- projet «*Data Governance Framework*» (2022, piloté par le Mexique) : il couvre certains domaines abordés ici, tels que gouvernance des données, qualité, métadonnées, etc.

L'évolution est similaire dans le Système statistique européen (SSE) :

- projet ESSnet ISAD (*Integration of Survey and Administrative Data*) terminé fin 2008 : il mentionne la transformation de données dans la phase de préparation des données avant l'intégration (par exemple, par appariement de fichiers) ;
- puis ESSnet «*Data Integration*» ou projet stratégique «ADMIN» (*ESS2020 ADMIN*) : les traitements de transformation des sources sont inclus dans une phase de préparation des données, peu détaillée, bien que connue pour mobiliser une part importante du travail statistique ;
- ESSnets *Big Data* I et II (2015 à 2021) : l'architecture métier élaborée identifie une fonction de *data wrangling*, c'est-à-dire « la possibilité de transformer les données du format source d'origine en un format cible souhaité, mieux adapté à une analyse et à un traitement ultérieur », et une fonction de représentation des données, c'est-à-dire l'ajout d'éléments de contexte et de structure (données dérivées, codes, catégories) aux données brutes. Ces éléments sont positionnés dans une couche de convergence des données, entre la couche des données brutes et celle des données statistiques ;
- démarche stratégique «*Trusted Smart Statistics*» : les récents développements lancés dans ce cadre réutilisent l'architecture définie par les projets *Big Data*.

À travers ces différents exemples, la problématique de transformation des données est clairement identifiée en tant que telle dans les instances internationales.

* *https://unece.org/statistics/modernization-official-statistics*.

***https://statswiki.unece.org/display/DA/CSDA+2.0*.

La spécification au niveau logique des traitements d'intégration, qui restent encore largement organisés en silos, permettra aussi de les rendre plus modulaires et mieux partageables. Dans cette optique, on peut envisager le déport en amont de certaines étapes chez le producteur de données (par exemple des traitements de pseudonymisation ou de recodification sur des smartphones avant le transfert de l'information). Cette problématique est notamment rencontrée dans le cadre des *smart statistics*[14]. Enfin, cette phase d'intégration automatisée des données prend encore plus son sens aujourd'hui avec la disponibilité massive de données protéiformes provenant de capteurs, d'outils numériques voire de réseaux électroniques/sociaux qui caractérisent la réalité sociale, environnementale et économique. Reste à s'assurer qu'on peut les utiliser pour produire des statistiques de qualité.

14 *https://www.cambridge.org/core/journals/data-and-policy/article/trusted-smart-statistics-how-new-data-will-change-official-statistics/380C6B6408D84C16164F33A1F4BF2F07.*

▶ Bibliographie

- BONNANS, Dominique, 2019. RMéS, le référentiel de métadonnées statistiques de l'Insee. In : *Courrier des statistiques*. [en ligne]. 27 juin 2019. Insee. N° N2, pp. 46-57. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4168396?sommaire=416841*.

- COTTON, Franck et DUBOIS, Thomas, 2019. Pogues, un outil de conception de questionnaires. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. N° N3, pp. 17-28. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4254216?sommaire=4254170*.

- COURMONT, Antoine, 2021. *Quand la donnée arrive en ville. Open data et gouvernance* urbaine. Presses universitaires de Grenoble. ISBN 2706147350.

- CROS, 2014. *Handbook on Methodology of Modern Business Statistics. In : site de Collaboration in Research and Methodology for Official Statistics*. [en ligne]. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://cros-legacy.ec.europa.eu/content/handbook-methodology-modern-business-statistics_en*.

- DURR, Jean-Michel, DUPONT Françoise, HAAG, Olivier et LEFEBVRE, Olivier, 2022. *Setting up statistical registers of individuals and dwellings in France: Approach and first steps.* In : Statistical Journal of the IAOS (SJIAOS). Volume 38, n°1, pp. 215-223. [en ligne]. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji210916*.

- ESS Vision 2020 ADMIN (Administrative data sources). In : *site de Collaboration in Research and Methodology for Official Statistics*. [en ligne]. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://cros-legacy.ec.europa.eu/content/ess-vision-2020-admin-administrative-data-sources*.

- FERMOR-DUNMAN, Verena, and PARSONS Laura, 2022. *Data Acquisition processes improving quality of microdata at the Office for National Statistics. Q2022 Vilnius*.

- HAND, David, 2018. *Statistical challenges of administrative and transaction data. In : Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 9 février 2018. Volume 181, N°3, pp. 555-605. [en ligne]. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://www.jstor.org/stable/48547504*.

- KOUMARIANOS, Heïdi et SIGAUD, Éric, 2019. Eno, un générateur d'instruments de collecte. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. N° N3, pp. 29-44. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4254218?sommaire=4254170*.

- LAMARCHE, Pierre et LOLLIVIER, Stéfan, 2021. Fidéli, l'intégration des sources fiscales dans les données sociales. In : *Courrier des statistiques*. [en ligne]. 8 juillet 2021. Insee. N° N6, pp. 28-46. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/5398683?sommaire=5398695*.

- SIX, Magdalena et KOWARIK, Alexander, 2022. *Quality Guidelines for the Acquisition and Usage of Big Data*. [en ligne]. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://lsvdv-my.sharepoint.com/:w:/g/personal/ingabal_stat_gov_lt/Evmr-Cle195BrkRRDtPr2OMBkJsZqEI7Arzy8H2auuaTPw?rtime=TaOXDVpp2kg*.

- UNECE Integration, statswiki, *A Guide to Data Integration for Official Statistics. In : site statswiki de l'UNECE*. [en ligne]. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://statswiki.unece.org/display/DI/Guide+to+Data+Integration+for+Official+Statistics*.

- UNECE quality, statswiki, Quality. In : *site statswiki de l'UNECE*. [en ligne]. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://statswiki.unece.org/display/DI/Quality*.

- UNECE GSBPM. In : *site statswiki de l'UNECE*. [en ligne]. [Consulté le 24 mai 2023]. Disponible à l'adresse : *https://statswiki.unece.org/display/GSBPM*.