Statistique fondée sur données administratives

Éléments pour un cadre général

Insee

Measuring, understanding

1 CADRE MÉTHODOLOGIQUE : ENQUÊTES VS SOURCES ADMINISTRATIVES
2 DONNÉES ADMINISTRATIVES : DÉFINITIONS ET CONSÉQUENCES
3 DE LA DONNÉE ADMINISTRATIVE À LA DONNÉE STATISTIQUE

01 CADRE MÉTHODOLOGIQUE
ENQUÊTES VS SOURCES ADMINISTRATIVES

CADRE GÉNÉRAL DU PROCESSUS D'ENQUÊTE

Principes généraux

- On veut comprendre des phénomènes socio-économiques
  - Et pour cela se donner les moyens *d'observer* la réalité "macro"
  - A l'aide d'une représentation synthétique quantifiée : des statistiques
- Pour cela on conçoit un système d'observation
  - Définition de la population observée
    - **Unités statistiques de référence**
    - **Champ, répertoire et principes de sélection**
  - Définition des variables observées
    - **concepts, référence à des nomenclatures**
  - Définition de la temporalité observée
  - Elaboration de l'instrument d'observation : le questionnaire

" Quantifier, c'est convenir puis mesurer " (A. Desrosières)

CADRE GÉNÉRAL DU PROCESSUS D'ENQUÊTE

On dispose aussi d'un cadre mathématique

- Théorie des sondages
- Calcul de variance
- Traitement de la non-réponse
- Calage sur marges
- … tout cela étant fondé sur la théorie des probabilités

Ce cadre permet de maîtriser l'erreur, en la quantifiant

- Composantes biais et variance … surtout variance

CADRE GÉNÉRAL DU PROCESSUS D'ENQUÊTE : CADRE OPÉRATIONNEL

On met en oeuvre le système d'observation

- Construction de la base de sondage
- Tirage d'échantillon
- Mise au point du (des) support(s) de collecte
- Mise au point du protocole de collecte
- Réalisation de la collecte sur une certaine ***temporalité***
  - En appliquant le protocole de collecte préalablement défini
  - Avec une organisation et des ressources dédiées
- Traitements statistiques
- Interprétation des résultats

Avec l'enquête, on a donc un certain degré de maîtrise

- Choix des unités à observer
- Choix des variables à observer
- Choix du champ d'observation
- Choix du protocole d'observation
- Choix des principes de catégorisation : les nomenclatures
- Maîtrise de la double temporalité … *et données figées*
- Capacité à estimer les erreurs

" Statistics is the technology of extracting meaning from data and of handling uncertainty " (D. Hand)

Avec les données administratives, on perd le plus souvent tout ce cadre :

- Concepts : imposés par le processus administratif
- Unités de gestion : pas nécessairement celles souhaitées
- Champ : spécifique à l'usage administratif
- Nomenclatures de gestion imposées
- En général, données vivantes, temporalité imposée par la gestion
- En général, pas de protocole d'observation, pas de collecte
- … et donc pas de "mécanisme de non-réponse"
- Pour l'essentiel, incapacité à estimer les erreurs

PERTE DU CADRE THÉORIQUE D'UNE ENQUÊTE

$$\hat{X}_{C_k} = \sum_{\substack{i \in S \subset U \\ i \in C_k}} w_i X_i^t$$

PERTE DU CADRE THÉORIQUE D'UNE ENQUÊTE : CONCEPTS

$$\hat{\bar{X}}_{C_k} = \sum_{\substack{i \in S \subset U \\ i \in C_k}} w_i \bar{X}_i^t$$

PERTE DU CADRE THÉORIQUE D'UNE ENQUÊTE : UNITÉS

$$\hat{\bar{X}}_{C_k} = \sum_{\substack{i \in S \subset U \\ i \in C_k}} w_i \bar{X}_i^t$$

PERTE DU CADRE THÉORIQUE D'UNE ENQUÊTE : CHAMP

$$\hat{X}_{C_k} = \sum_{\substack{i \in S \subset U \\ i \in C_k}} w_i X_i^t$$

PERTE DU CADRE THÉORIQUE D'UNE ENQUÊTE : NOMENCLATURES

$$\hat{X}_{C_k} = \sum_{\substack{i \in S \subset U \\ i \in C_k}} w_i X_i^t$$

PERTE DU CADRE THÉORIQUE D'UNE ENQUÊTE : PÉRIODES

$$\hat{X}_{C_k} = \sum_{\substack{i \in S \subset U \\ i \in C_k}} w_i X_i^t$$

PERTE DU CADRE THÉORIQUE D'UNE ENQUÊTE : PROTOCOLE

$$\hat{X}_{C_k} = \sum_{\substack{i \in S \subset U \\ i \in C_k}} w_i X_i^t$$

PROCESSUS FONDÉ SUR DONNÉES ADMINISTRATIVES : PREMIERS ÉLÉMENTS DE COMPARAISON

… mais l' "existence" de multiples sources de données administratives offre des opportunités tout à fait nouvelles

- On s'affranchit en quelque sorte de tout le processus de collecte
- On a une forme d'exhaustivité, certes imparfaite
- Le caractère obligatoire de certaines données peut fournir quelques garanties de qualité

Il est donc essentiel de mieux comprendre les données administratives pour fixer un cadre de travail efficace

02 DONNÉES ADMINISTRATIVES : DÉFINITIONS ET CONSÉQUENCES

Une définition à plusieurs dimensions

- **Une donnée, c'est un triplet (concept, domaine de validité, valeur)**
  - Exemple : (activité principale d'une entreprise, nomenclature NAF, 62.01Z)
- **Le concept se décrit comme l'attribut d'un objet**
  - Activité principale *d'une entreprise*, âge *d'une personne*, date de début *d'un contrat*, lieu *d'un accident*, ...
  - Objets : en statistique, souvent individus, ménages, entreprises
- **Le domaine de validité, c'est l'ensemble des valeurs admissibles**
  - Intervalle de valeurs, liste de codes, … ou règles de validité pour une date
- **Mais une donnée se caractérise aussi par son mode d'obtention**
  - Comment la donnée naît-elle ?

DONNÉES ADMINISTRATIVES : MODE D'OBTENTION

Caractéristiques :

- Elle est issue d'une institution dont le rôle n'est pas de produire cette information
- Elle est obtenue lors d'une opération de gestion et conservée dans une base de données
- Sa naissance est liée à des "événements" administratifs (ou traitements automatiques liés)
  - Internes à l'administration : décision d'attribution d'une prestation
  - Externes à l'administration : changement d'adresse, naissance, décès, demande de prestation, ...
- Elle a des impacts dans la vie réelle
  - Ex : clôture d'une offre d'emploi, immatriculation entreprise
- Elle sert fondamentalement de support à des processus opérationnels

La donnée peut donc être vue comme un "effet de bord" d'un processus de gestion

- *Data exhaust*, ce qui reste d'un processus opérationnel : déchet, coproduit (Hand 2018)

" Administrative data are data exhaust: that which is left over after the organizational machinery has used the data to drive itself forward " (D. Hand)

DONNÉES ADMINISTRATIVES : LES “OBJETS” ADMINISTRATIFS

Objets :

- **Nombreux, et liés entre eux**
  - Ex. dans la DSN : contrat, salarié, rémunération, établissement
  - Ex. cadastre : local, personne, transaction
- **Subordonnés à un usage**
  - Ex. sécurité sociale : attribuer une prestation
- **… ce qui implique des “rôles” associés à une entité, liés aux objectif de gestion**
  - Ex. usager, patient, étudiant, bénéficiaire, ayant-droit, ...

Ces objets s’inscrivent dans un “SI administratif”, complexe

C'est l"habitat" naturel des données

- Il a des finalités opérationnelles
  - Ex. assurer un parcours de soin, verser une retraite à bon droit
- Il est constitué d'un ensemble de processus
  - Dont processus RH, comptable, métier
- Il est mis à jour via des événements internes ou externes
- Tout cela est supporté par des bases de données
- C'est un système qui "vit" en permanence
  - La plupart du temps, il n'y a pas de "collecte"

    … sauf cas particulier des déclarations administratives
  - Donc inutilisable tel quel pour la statistique

"Les sources administratives portent plus sur les moyens et sur les actions publiques que sur l'état de la société." (A. Desrosières)

DONNÉES ADMINISTRATIVES : QUELLES TYPOLOGIES ?

Peu de typologies présentées dans la littérature :

- Approche thématique/ par oganisation (fiscale, sociale, santé...), Unece 2011
- Approche par grande finalité du processus (flux de biens et personnes, prestations, enregistrements d'évènements, réglementation..), Statistique Canada 1987

Constat : comprendre la manière dont naissent les données est décisif en vue d'usages ultérieurs

- Entité à l'origine de la naissance de la donnée
- Lien avec la vie réelle (naissance, création d'une entreprise, modification d'une variable de gestion…)
- Temporalités

Comment faire le passage SI → source ?

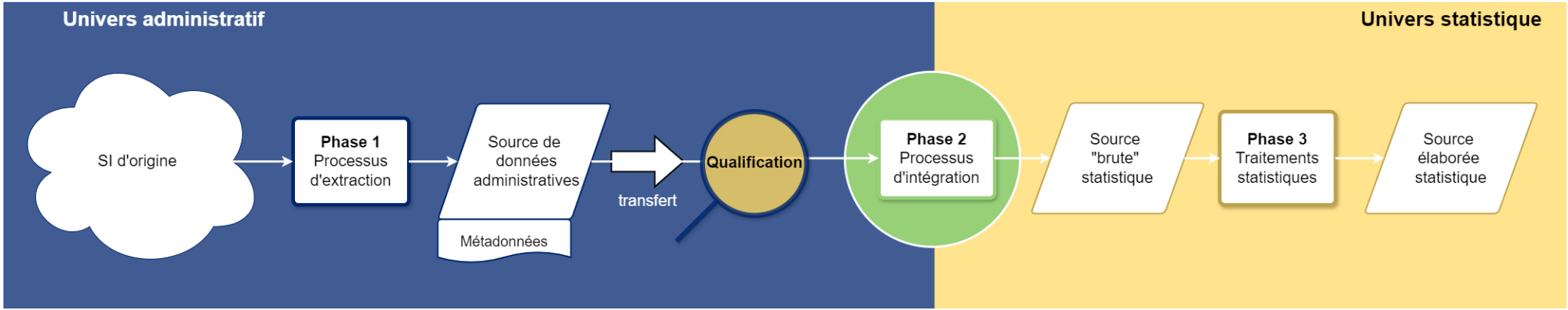
- **Cas favorable : les déclarations administratives**
  - Passage à source figée “relativement” simple
- **Cas semi-favorable : bilans annuels intégrés au process**
  - Données figées … mais processus de figeage méconnu
  - L’administration impose son mode de calcul
- **Cas général**
  - Soit on demande à l’administration de produire le fichier
    - **Problème de maîtrise**
  - Soit les statisticiens le font eux-mêmes
    - **Problème de coût**
- **Et surtout : il y a plusieurs possibilités de passage => pas de définition univoque de la notion de “source administrative”**

DONNÉES ADMINISTRATIVES : COMMENTAIRES

- **Données dans une configuration inhabituelle, et souvent méconnue**
- **Déclaration = protocole de collecte formalisé**
  - Plus proche d'une logique "statistique"
  - Raison pour laquelle ce sont les sources les plus utilisées
- **La "source administrative" est un objet étrange, hybride, qui contient souvent des données de différents types :**
  - Données issues de déclaration, données de gestion, données de répertoires, et parfois données agrégées
- **Pour la statistique publique il est essentiel de se doter d'un cadre partagé pour une meilleure maîtrise**

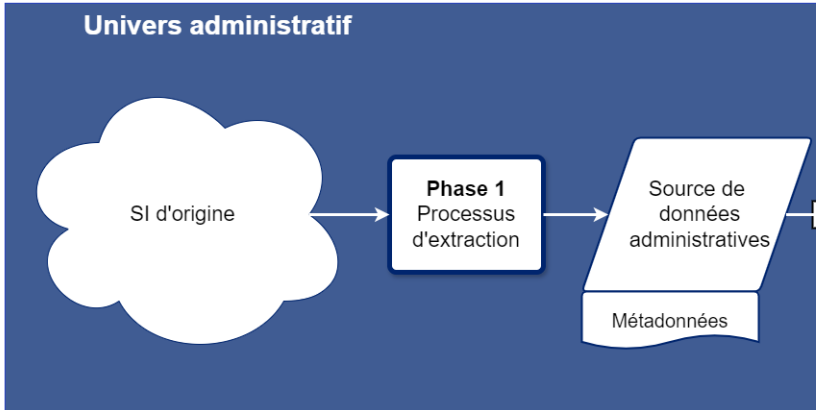
03 DE LA DONNÉE ADMINISTRATIVE À LA DONNÉE STATISTIQUE

Trois grandes phases distinctes ...



… pour traiter les différentes difficultés (champ, unité, variables, temporalité, estimation des erreurs)

PHASE 1 : DU SI ADMINISTRATIF À LA SOURCE ADMINISTRATIVE



D'UN SI VIVANT À DES DONNÉES FIGÉES

La première phase se passe dans l'univers administratif

Elle doit régler plusieurs obstacles à l'exploitation statistique des données :

- Leur caractère vivant du SI d'origine
- L'immensité du SI d'origine
- L'absence éventuelle de définition de période de collecte et de référence

En conséquence, cette phase peut comporter

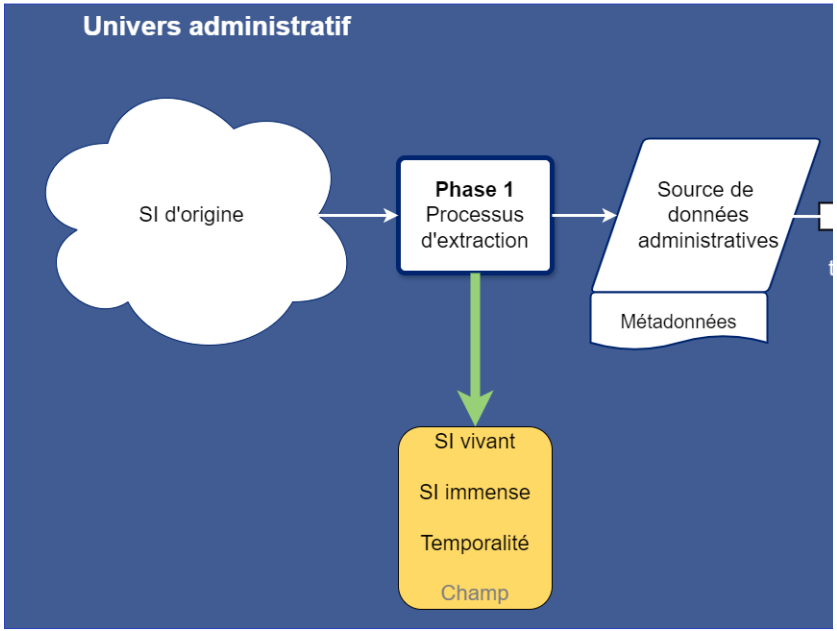
- Des filtres sur les objets
- Des sélections sur les attributs
- Des agrégations
- Des jointures entre plusieurs tables du SI d'origine

Il y a également une transformation de format physique pour rendre les données transportables

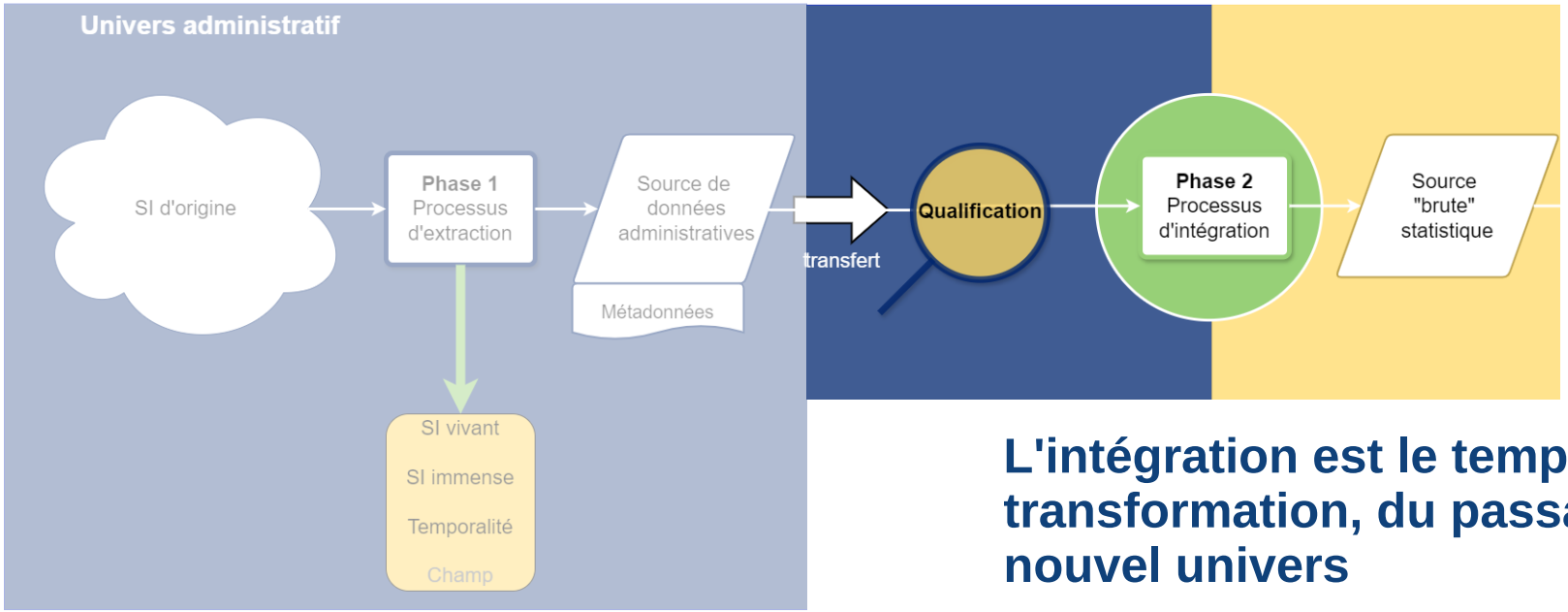
La première phase nécessite :

- Un effort important pour comprendre l'univers de gestion des données
- Des échanges indispensables avec le fournisseur
- Une coopération pouvant revêtir différentes formes (ONS)

PHASE 1 : LA PROJECTION D'UNE RÉALITÉ COMPLEXE



PHASE 2 : DE LA SOURCE ADMINISTRATIVE AUX DONNÉES STATISTIQUES BRUTES



L'intégration est le temps de la transformation, du passage dans un nouvel univers

Il y a un enjeu statistique et technique

QUALIFIER LES DONNÉES : UN PRÉALABLE INDISPENSABLE À L'EXPLOITATION

L'exploitation statistique de données administratives est une réutilisation

Qualifier une source, ou évaluer la qualité de données administratives :

- C'est s'assurer que la qualité des données est adaptée à l'usage statistique qui en est envisagé
- C'est s'assurer que le changement d'univers des données est possible

Qualifier la source administrative avant transformation facilite les échanges avec le fournisseur

La qualification des données est indispensable

- Pour valider l'opportunité d'une première utilisation
- Lors d'une utilisation répétée

“ Quality means fitness for use ” (J. Juran)

QUALIFICATION – PRIMO ACQUISITION

Avant même l'usage effectif d'une source, il y a une étape de qualification, liée à une primo acquisition

- Evaluer la pertinence de la source, en lien avec le besoin exprimé
- Analyser les objets et concepts administratifs, évaluer la qualité de la transformation en unités statistiques et des variables d'intérêt
- Evaluer la couverture, la distribution des variables d'intérêt

QUALIFICATION – ACQUISITION RÉPÉTÉE

Dans un contexte d'acquisition répétée, plusieurs étapes d'évaluation sont nécessaires :

- À la réception des données, évaluer leur complétude (nombre de fichiers par exemple)
- Lors de l'intégration, s'assurer de la conformité du format
- Après l'intégration et avant les traitements statistiques, évaluer les principales variables (comptages des objets, comptages des unités créées, variables de filtres)

L'évaluation peut s'appuyer sur des programmes automatisés, des comptages simples

Cette évaluation est utile :

- Pour valider le bon déroulement du processus :
  - De la phase 1
  - Du transfert des données
  - De leur intégration dans le SI
- Pour détecter de nouveaux comportements de gestion, les conséquences des changements de législation etc...
  - L'impact du prélèvement à la source sur le nombre d'individus

PHASE DE TRANSFORMATION : SIX CATÉGORIES DE TÂCHES

Trois tâches qui peuvent s'appliquer aussi aux données d'enquête :

- Renommer les variables sélectionnées
- Pseudonymiser
- Calculer les variables dérivées

Trois tâches plus spécifiques à l'intégration de données administratives qui peuvent être plus délicates :

- Restructurer les données par unité statistique
- Recoder
- Filtrer les enregistrements utiles

Voir Cotton et Haag 2023

Restructurer les données par unité statistique

- Il n'y a pas nécessairement de bijection objet <-> unité
  - il peut y avoir découplage ou à l'inverse fusion d'objets
- Les liens inter-objets peuvent participer à la définition des unités statistiques :
  - lien logement occupant pour définir un ménage
  - lien pérenne entre un payeur et un bénéficiaire pour suivre des revenus
- La constitution des unités statistiques a un impact sur des variables liées aux objets
  - agrégation de périodes et gestion de date de début, date de fin
  - affectation des revenus d'un individu à un seul ménage (même si cet individu appartient à plusieurs ménages)

RECODER : COMPRENDRE LE LANGAGE DE L'UNIVERS ADMINISTRATIF

Recoder

- Normaliser (une adresse), coder dans une nomenclature
- Traiter les valeurs refuge (hors ou dans le domaine de définition), de gestion, aberrantes

Dans l'univers des données administratives, on rencontre les faits suivants :

- Votre parking en sous-sol se situe à l'étage `81`
- Une date de naissance manquante peut valoir

  `32/13` `00/00` `99/99` `WW/WW` `./.` `01/01`

- Les naissances le `15` du mois sont trois fois plus fréquentes que la moyenne des autres jours
- Certains individus se prénomment `SNP` (sans nom prénom)

FILTRER : CHAMP ET COUVERTURE

Passer du champ administratif au champ statistique :

- Traiter la surcouverture statistique (doublons, hors champ stat)

Filtrer les enregistrements utiles à partir d'une information administrative

- Implique une très bonne qualité des variables de filtre
  - Variables distinguant les locaux d'habitation des autres locaux
- Et une évaluation rigoureuse des évolutions de ces variables afin de détecter l'impact éventuel sur les résultats
  - Bascule dans ou hors du champ statistique non détectable sur le volume initial de données

L'INTÉGRATION : ÉGALEMENT UNE PHASE TECHNIQUE

Avec les objets administratifs, vient une structure de données parfois complexe

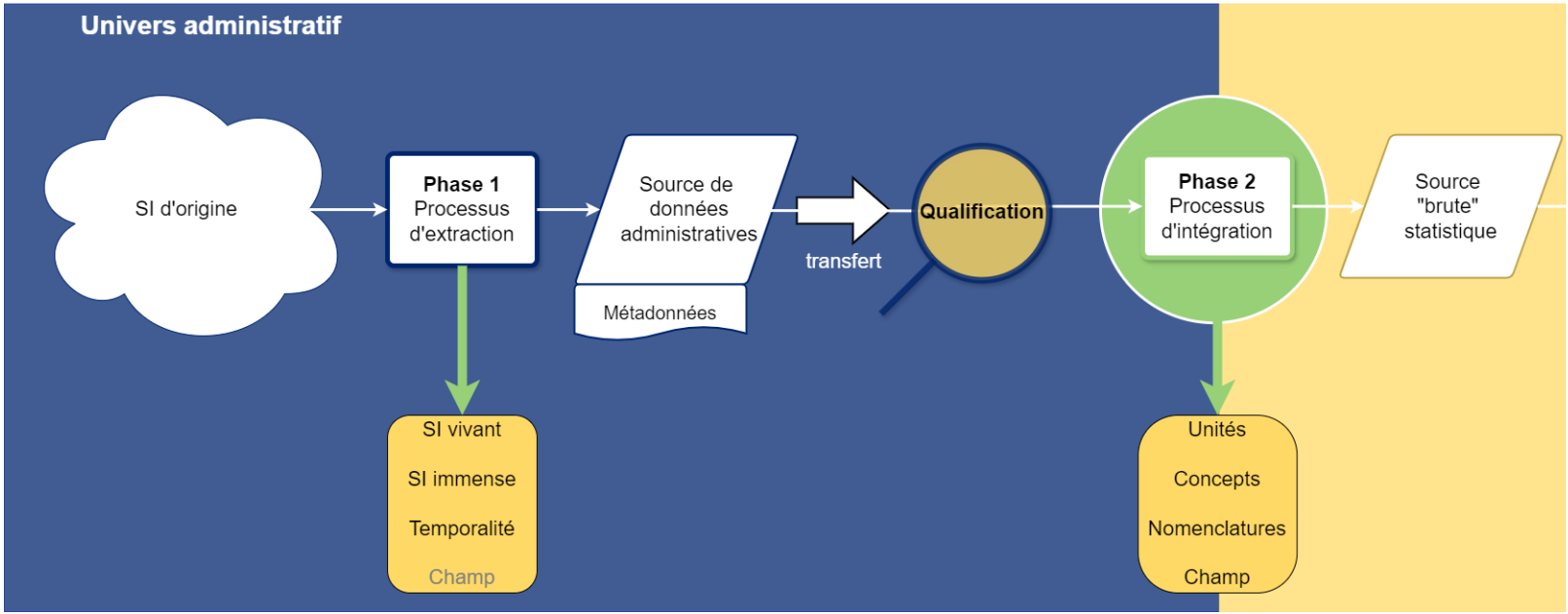
- Lecture de fichiers positionnels hiérarchiques
- Choix du format des bases de données statistiques (empiler des objets différents ou multiplier les bases et les relations)

C'est une étape intermédiaire qui doit prendre en compte la constitution de données suffisamment stables pour l'analyse statistique :

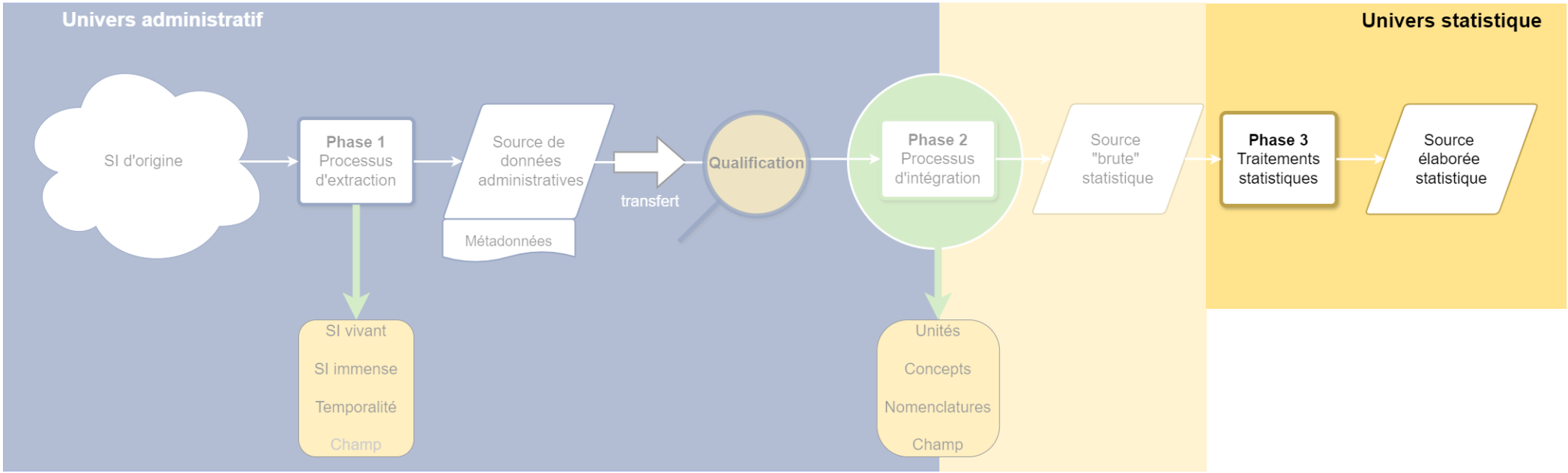
- Enjeu autour de la réception des données PASRAU (livraison quotidienne)

Un processus qui doit être réplicable

PHASE 2 : LE PASSAGE À L'UNIVERS STATISTIQUE



PHASE 3 : DES DONNÉES BRUTES AUX DONNÉES ÉLABORÉES



PHASE 3 : LES TRAITEMENTS STATISTIQUES

Les méthodes habituelles de traitement s'appliquent (data editing), et peuvent nécessiter des adaptations

- Data editing

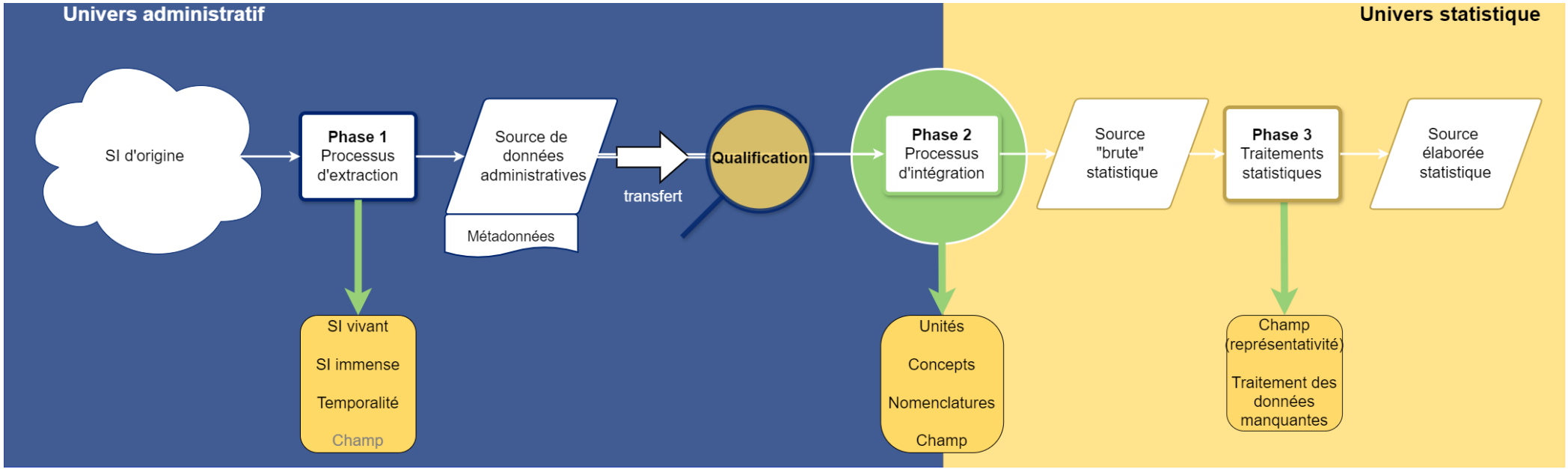
Certaines méthodes sont plus souvent utilisées dans un contexte de données administratives :

- Appariements, capture-recapture

Il est nécessaire de poursuivre la prise en compte de certaines spécificités des données :

- Mécanisme de données manquantes (versus traitement de la non réponse)

PHASE 3 : CONCLUSION



LE CADRE PROPOSÉ

Est un éclairage sur le “statut” des différentes sources de données en jeu

- Qu'est-ce qui est en entrée de mon processus ? La source “primaire” ? Quels sont les traitements opérés ?

Met l'accent sur le lien fort entre enjeux statistiques et processus d'exploitation des données administratives

- Quels “problèmes” statistiques doivent être résolus par chaque phase ?

Insiste sur la nécessité d'évaluer les données et leurs traitements

Le cadre proposé est volontairement simplificateur et sera détaillé dans un document de travail

- **L'exploitation des données administratives se fait très souvent dans un contexte multisources**

Les trois phases identifiées peuvent chacune être le point de départ de mutualisation, en fonction des besoins

La mesure des erreurs reste un sujet d'investissement important à mener pour la statistique publique

Résil, par son statut de référentiel, pose le cadre d'arrivée des statistiques sociales

Merci de votre attention

insee.fr

Insee

Measuring, understanding