

Résil : un changement de paradigme pour l'élaboration des statistiques sociales



- 1 RESIL ET LES DONNÉES ADMINISTRATIVES**
- 2 LES TRAVAUX MÉTHODOLOGIQUES**
- 3 LES PERSPECTIVES**

01 RESIL ET LES DONNÉES ADMINISTRATIVES

« Le programme RÉSIL vise à construire un système de répertoires d'individus, de ménages et de locaux d'habitation, durable et évolutif, mis à jour à partir de sources administratives diverses. »

(source : Compte rendu du CD Insee du 5 octobre 2020)

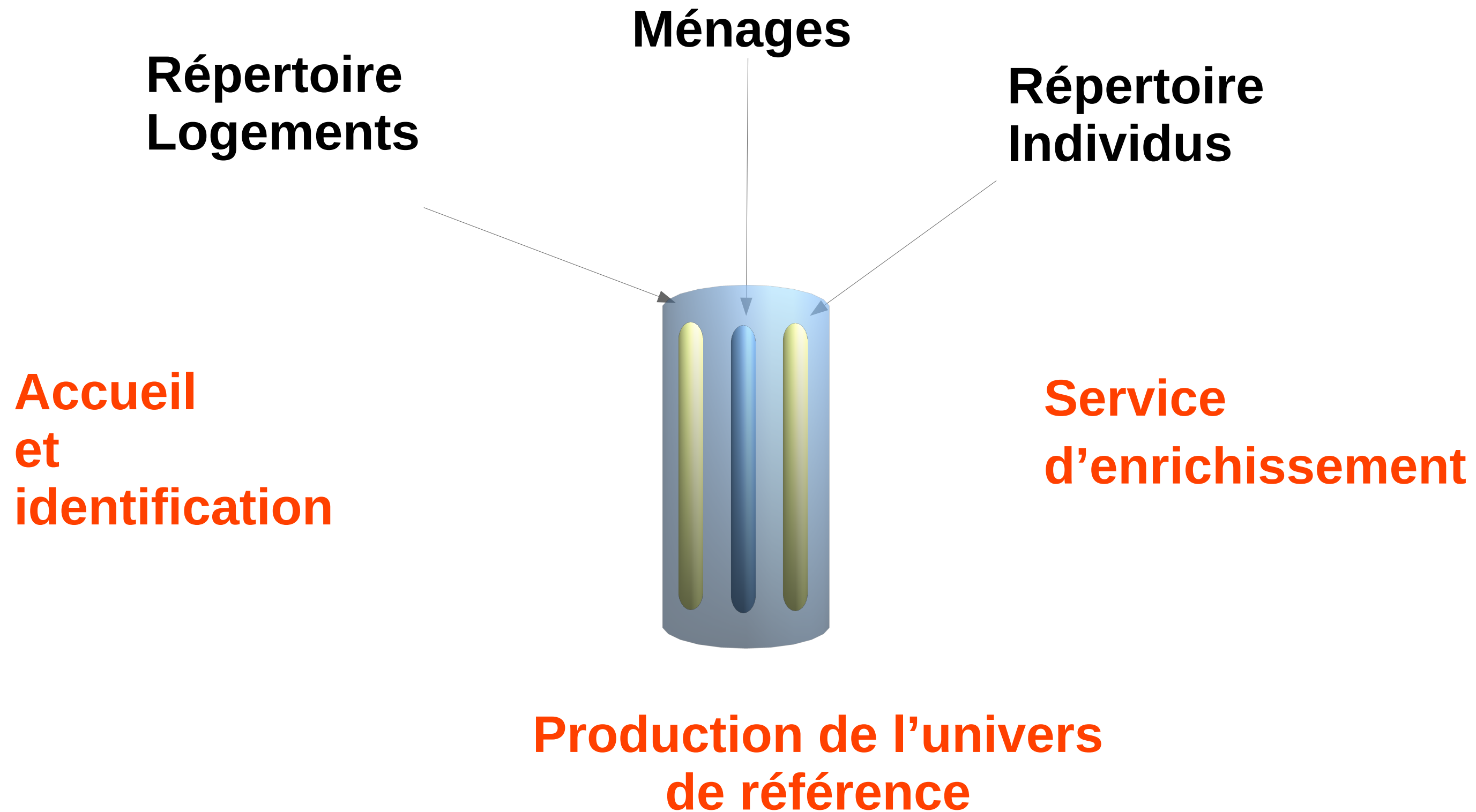
Résil produira :

- Une liste la plus exhaustive possible des personnes qui résident en France
- Une liste la plus exhaustive possible des logements implantés en France
- Une liste la plus exhaustive possible des ménages installés en France c'est-à-dire des personnes vivant dans un même logement

- 2 répertoires (individus et logements) mis à jour au fil de l'eau,
- un univers de référence consistant en trois bases annuelles
(individus, logements et ménages)

cet univers de référence peut notamment servir à produire des bases de sondage utiles pour l'échantillonnage et le calage, et à vérifier la qualité de couverture des sources administratives ou à faciliter les appariements de sources ;

- et trois services (accueil des sources, production d'univers de référence et production de fichiers enrichis par appariement) ;
- dont la gestion des informations localisantes est assurée par le référentiel Gaïa.



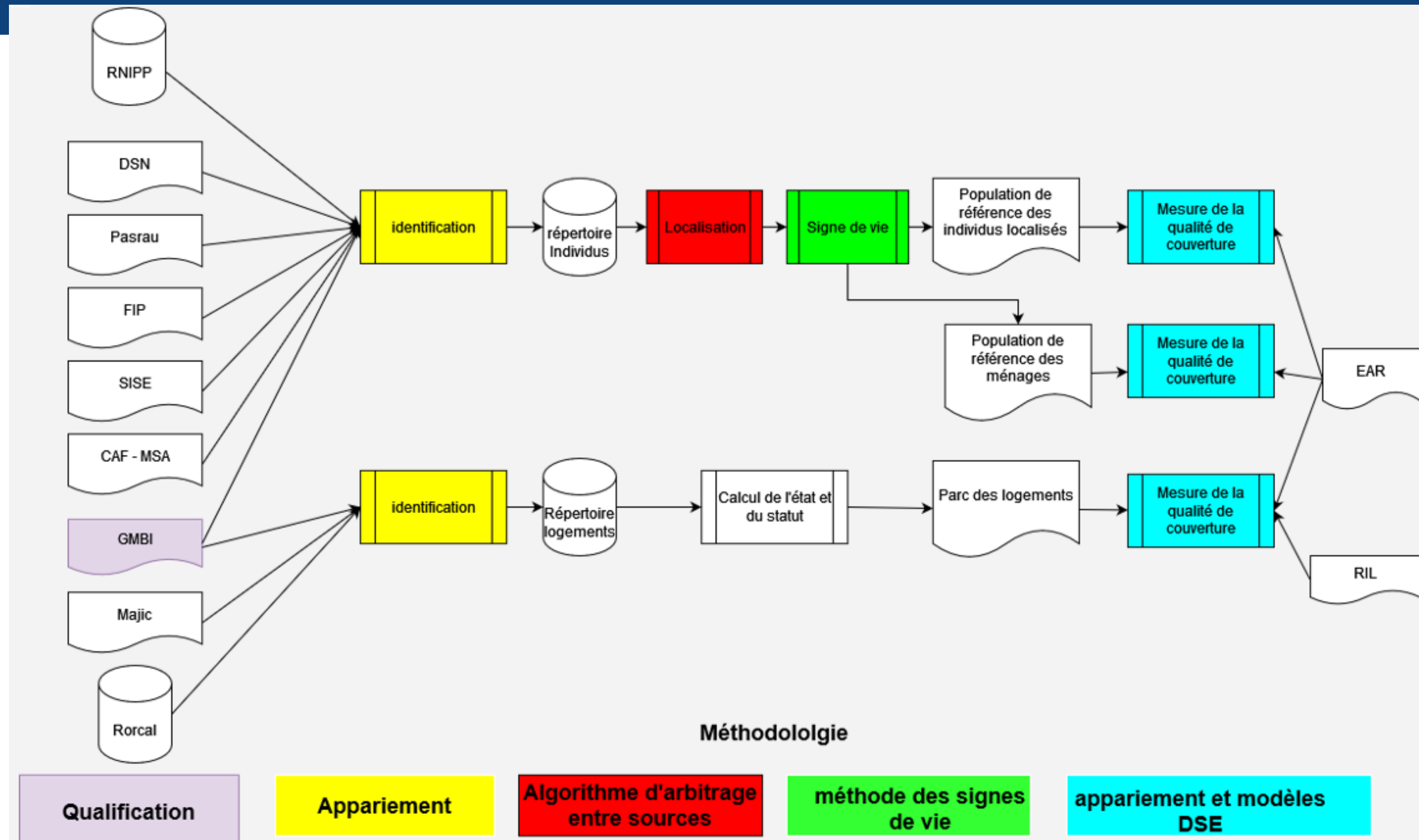
- Les données administratives alimentent Résil
- Résil va faciliter leur appariement
- Résil va permettre de mesurer leur qualité de couverture
- Résil va permettre de corriger les éventuels biais de couverture

02 LES TRAVAUX MÉTHODOLOGIQUES

LE PROCESSUS ENVISAGÉ POUR LA DÉFINITION DES UNIVERS DE RÉFÉRENCE

INDIVIDUS

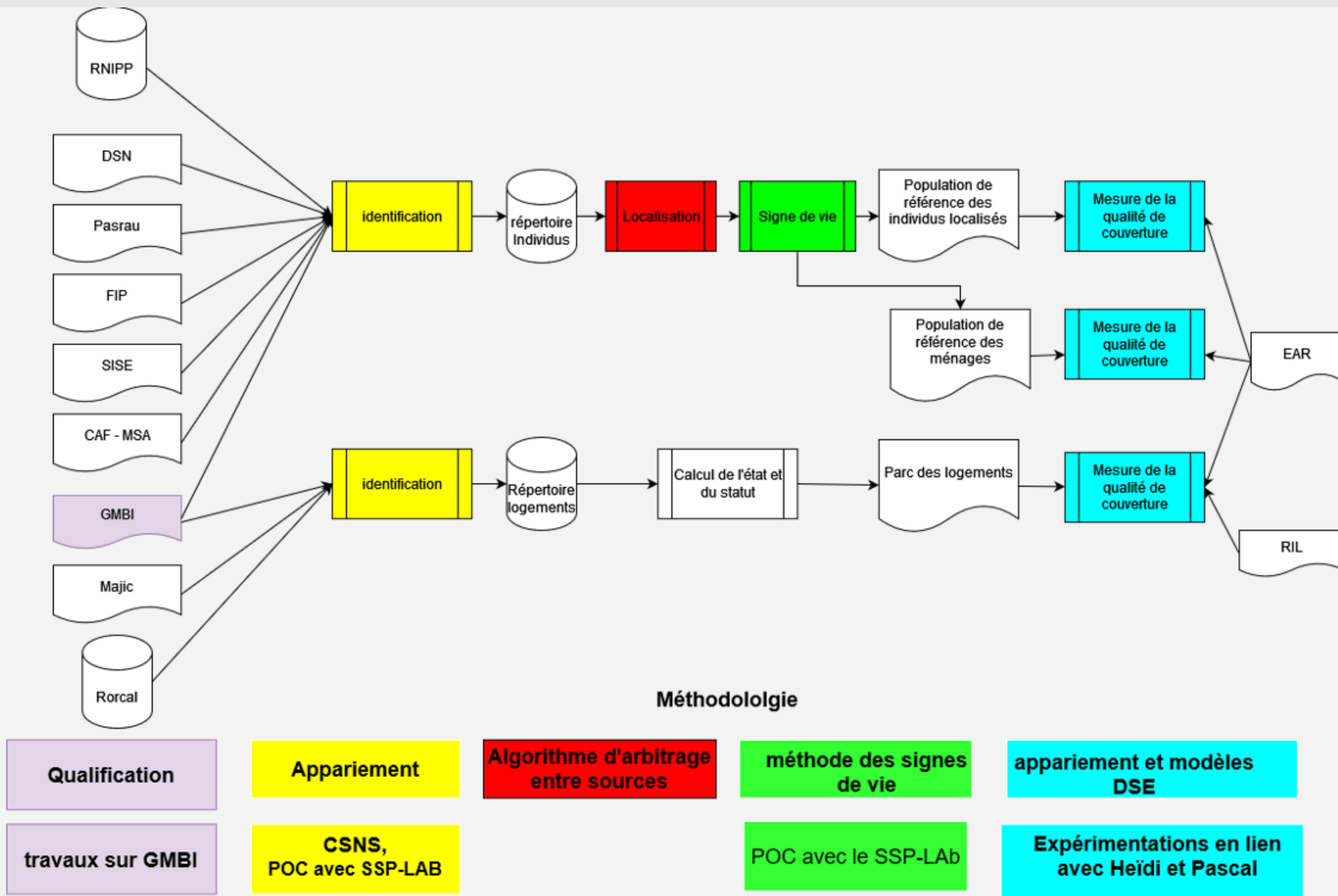
LOGEMENTS



LE PROCESSUS ENVISAGÉ POUR LA DÉFINITION DES UNIVERS DE RÉFÉRENCE

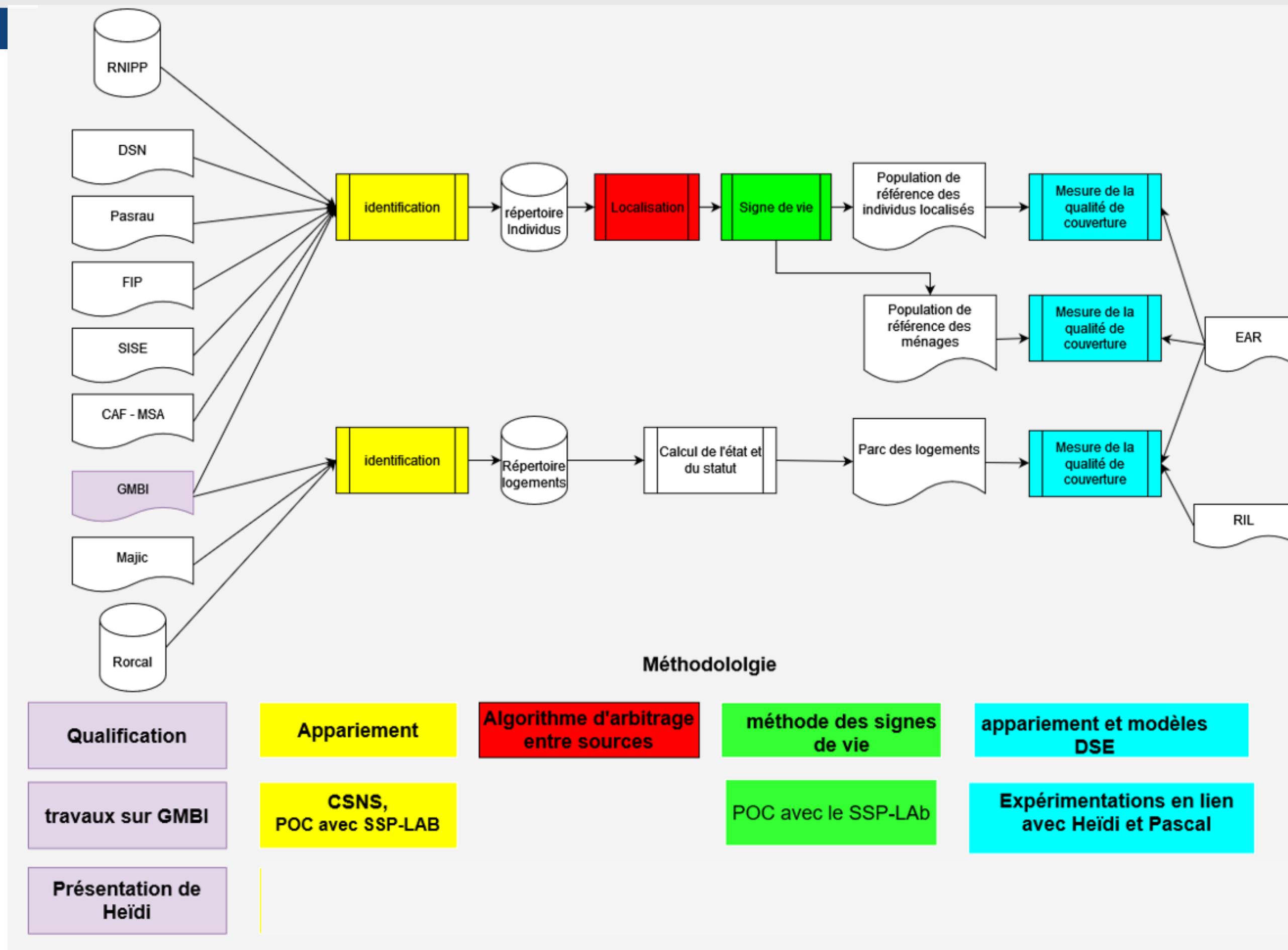
INDIVIDUS

LOGEMENTS



INDIVIDUS

LOGEMENTS



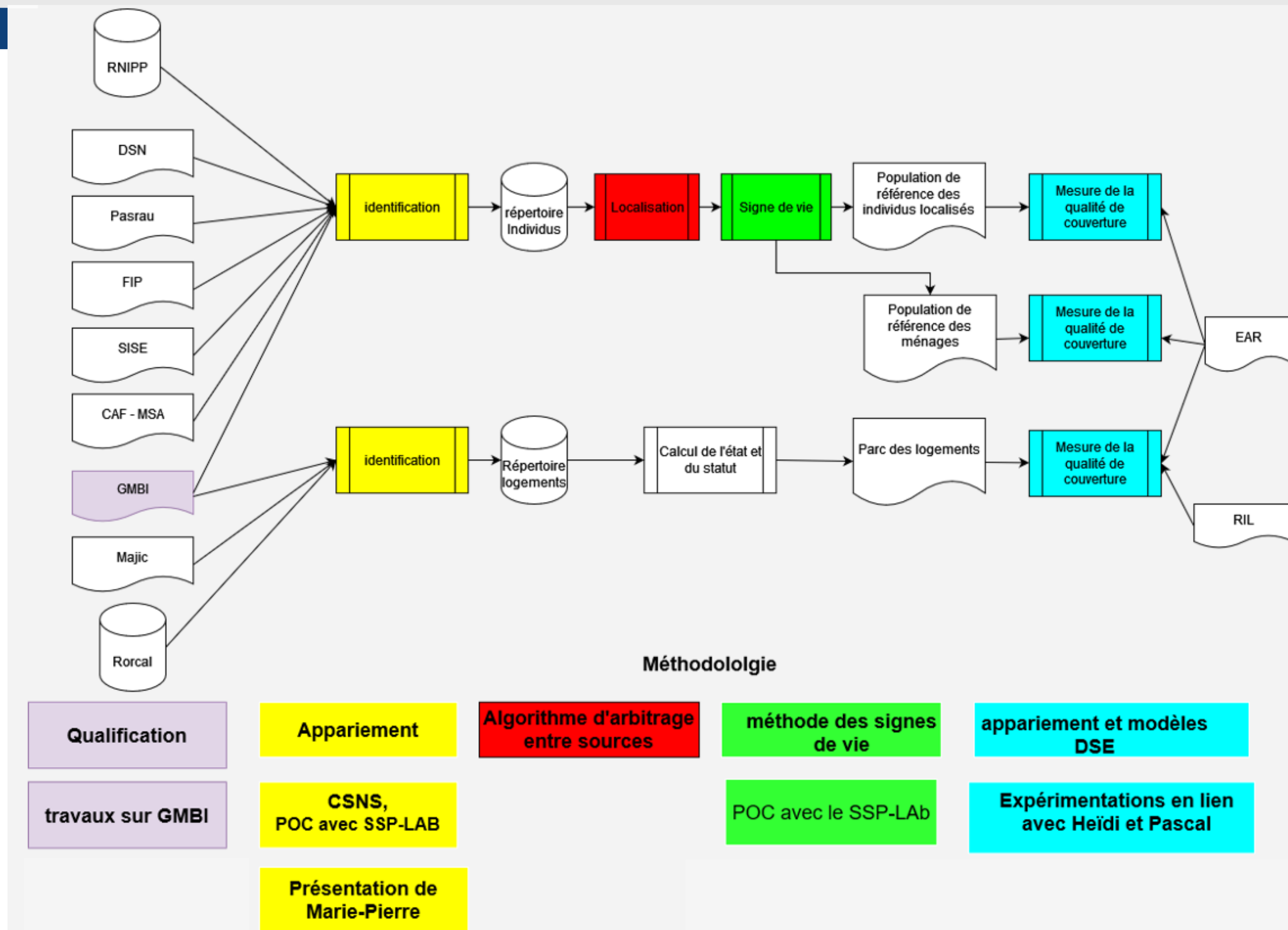
La qualification est une étape incontournable de l'utilisation statistique de données administratives :

- L'exploitation statistique de données administratives est un réemploi : il s'agit de s'assurer de la qualité des données pour ce nouvel usage
 - Conformité des données, couverture, pertinence...

- Profiter de la variété des points de vue et des connaissances :
 - La connaissance de la littérature sur le sujet apportée par le DMS
 - La connaissance du processus métier apportée par Résil
 - Pour le DMS, la confrontation de principes, propositions méthodologiques à un cas d'usage concret est indispensable
 - Pour Résil, un regard externe qui pousse à formaliser l'expression du besoin, l'analyse des objets
- Un sujet pour Résil au-delà de GMBI :
 - Pour GMBI mais aussi les autres sources accueillies par Résil, il sera nécessaire de généraliser l'évaluation de la qualité (éléments de rapport type proposés)
- Au-delà de Résil, évaluer la qualité de GMBI est utile pour d'autres processus statistiques :
 - Fidéli et par ricochet la base de sondage des enquêtes, différents produits statistiques...

INDIVIDUS

LOGEMENTS



Des travaux exploratoires sur les appariements



Journées de méthodologie statistique de l'Insee

2022

*PROBABILISTES OU DÉTERMINISTES, DES MÉTHODES D'APPARIEMENTS AU
BANC D'ESSAI DU PROGRAMME RÉSIL*

Olivier Haag (), Heïdi Koumarianos(**), Lucas Malherbe(**)*

Le CSNS (2022)

**Le Code statistique non signifiant
(CSNS) : un service pour faciliter
les appariements de fichiers**

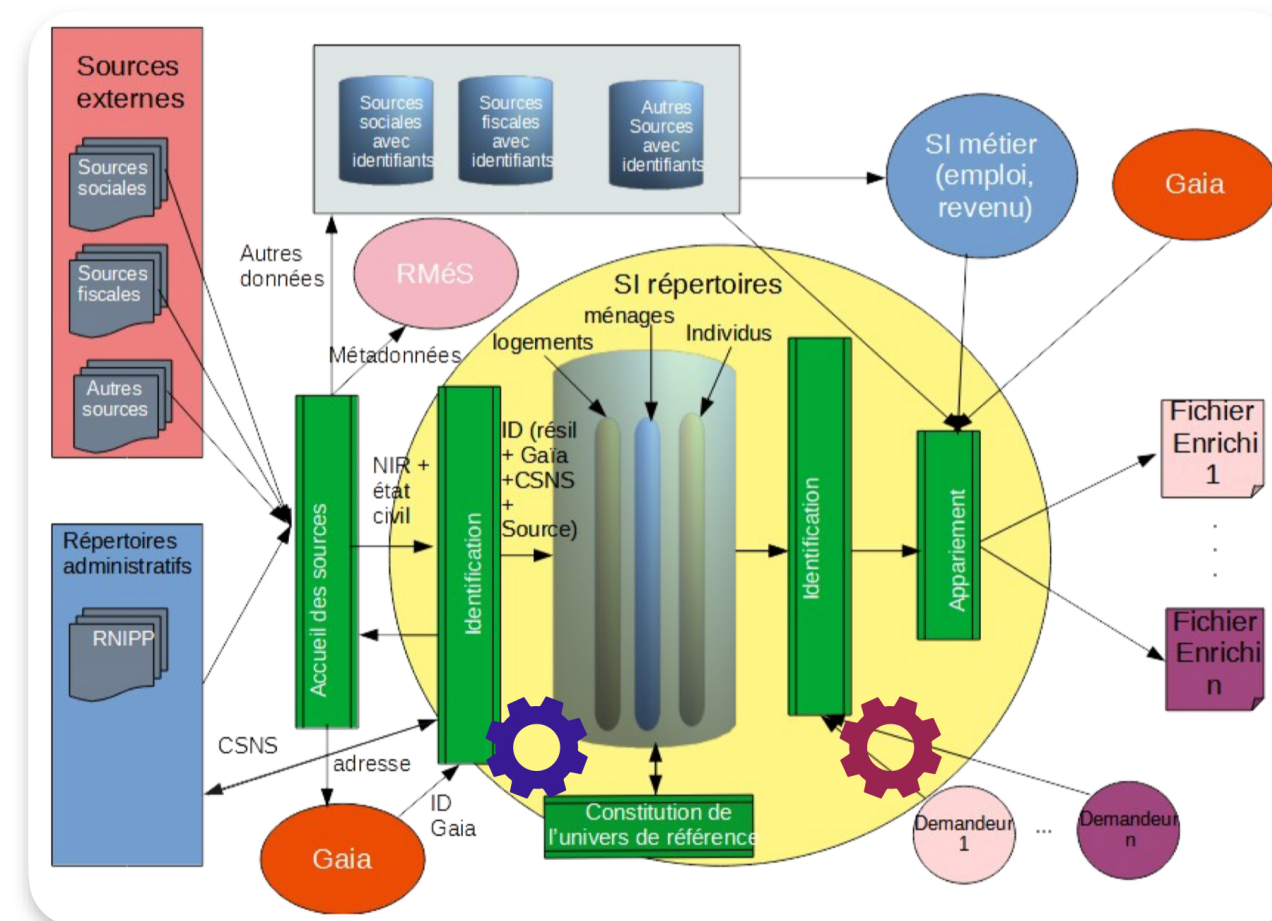


Yves-Laurent Bénichou, Lionel Espinasse** et Séverine Gilles****

Objectif : compléter le moteur d'identification du CSNS grâce à :

- variables supplémentaires disponibles dans Résil
- plus de flexibilité pour le réglage des paramètres suivant les sources à appairer

Deux cas d'usage :



- ⚙️ Appariement pour constituer l'univers de référence
- ⚙️ Appariement pour enrichir un fichier externe ou identifier un individu externe

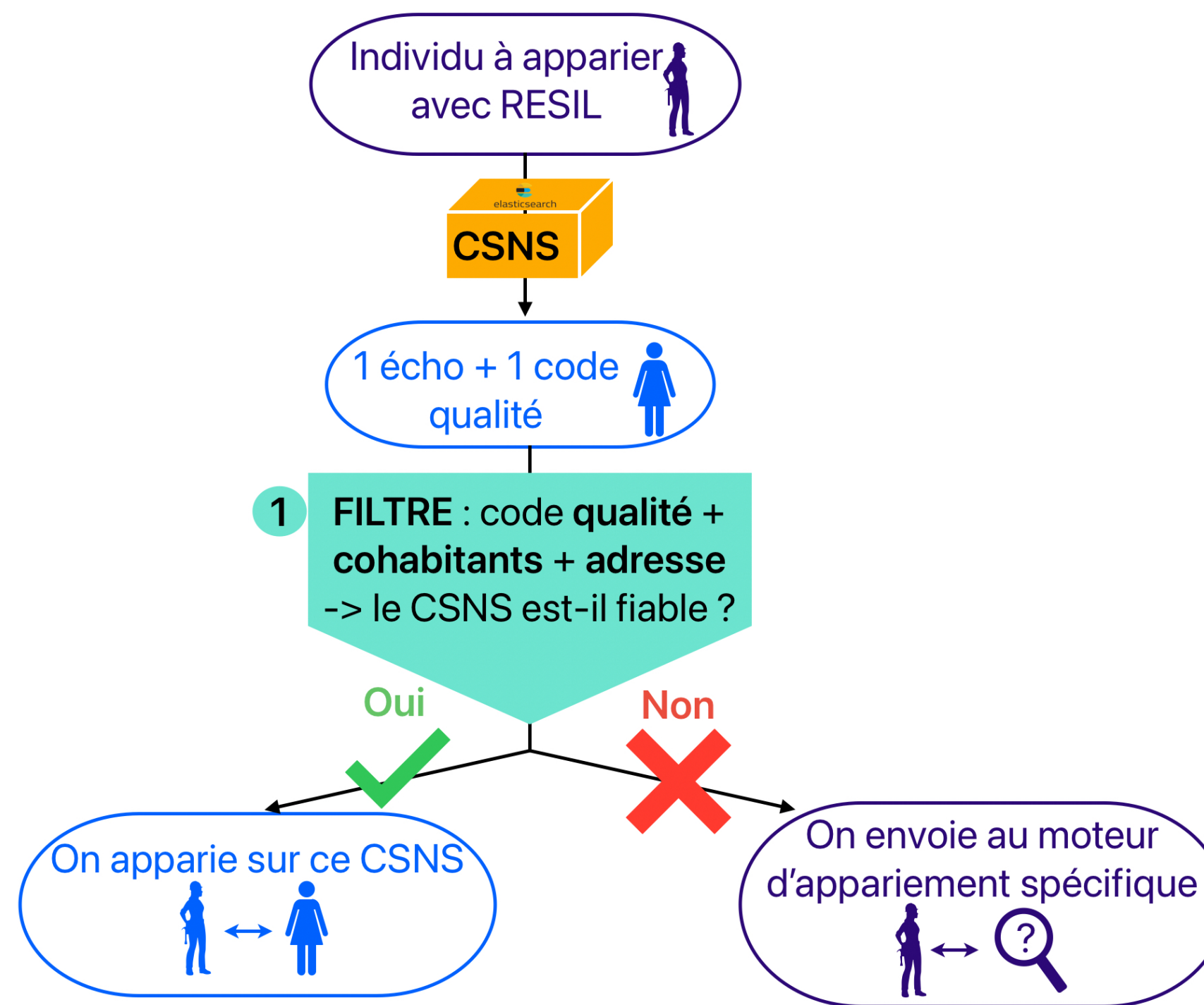
Deux sprints mobilisant la DMCSI et la DSDS (RESIL et le DRCVM)

- 1er sprint de 10 jours en mai 2023
- 2ème sprint de 5 jours en octobre 2023

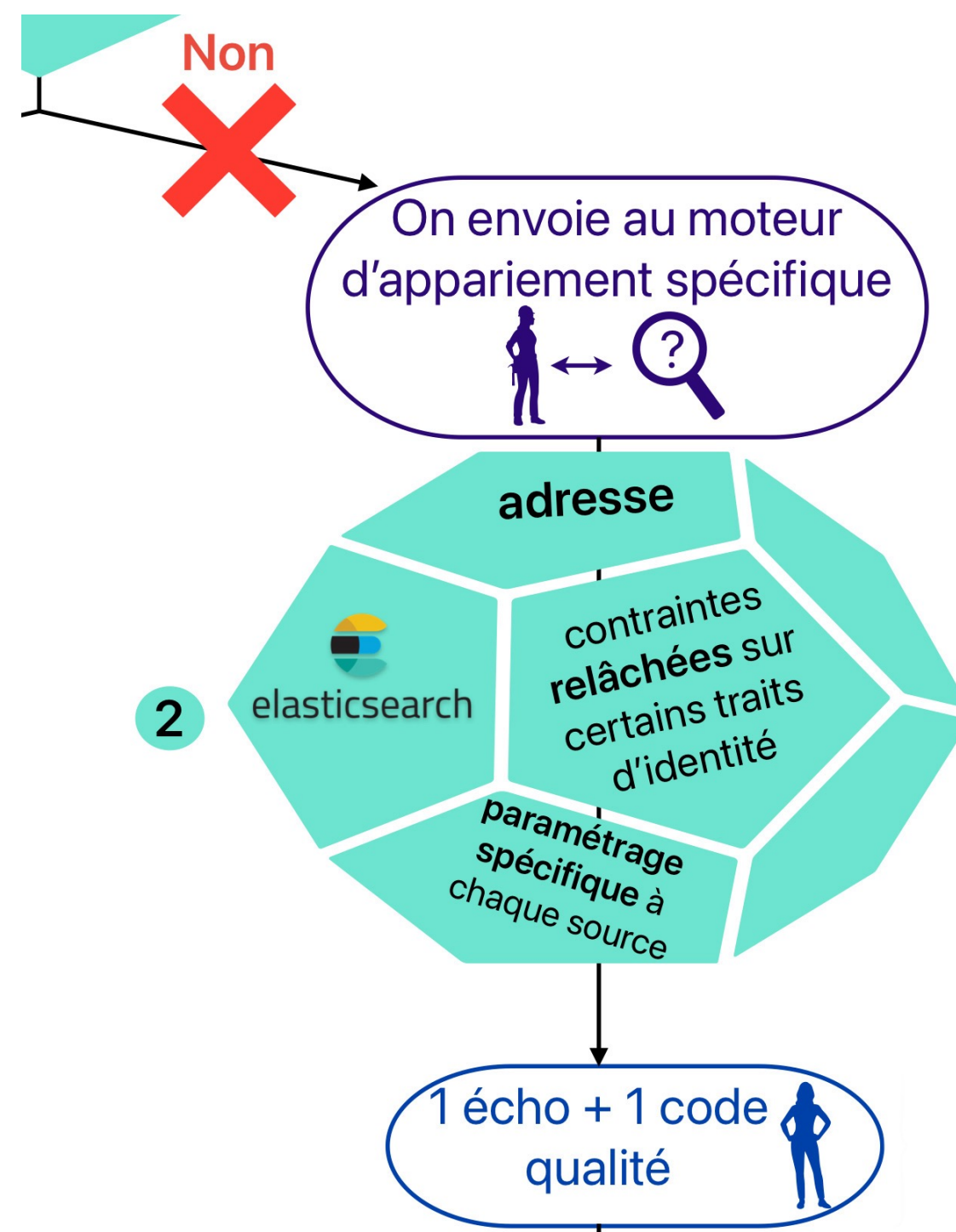
Utilisation du "pseudo-RESIL" pour les premiers travaux

- Répertoire RESIL pas encore disponible au moment du sprint.
 - Pseudo-RESIL construit en **empilant des sources** proches de celles du futur RESIL.
- => Pas de prédiction qualitative d'efficacité du moteur, mais suffisamment proche de la base finale pour avancer dans les travaux exploratoires

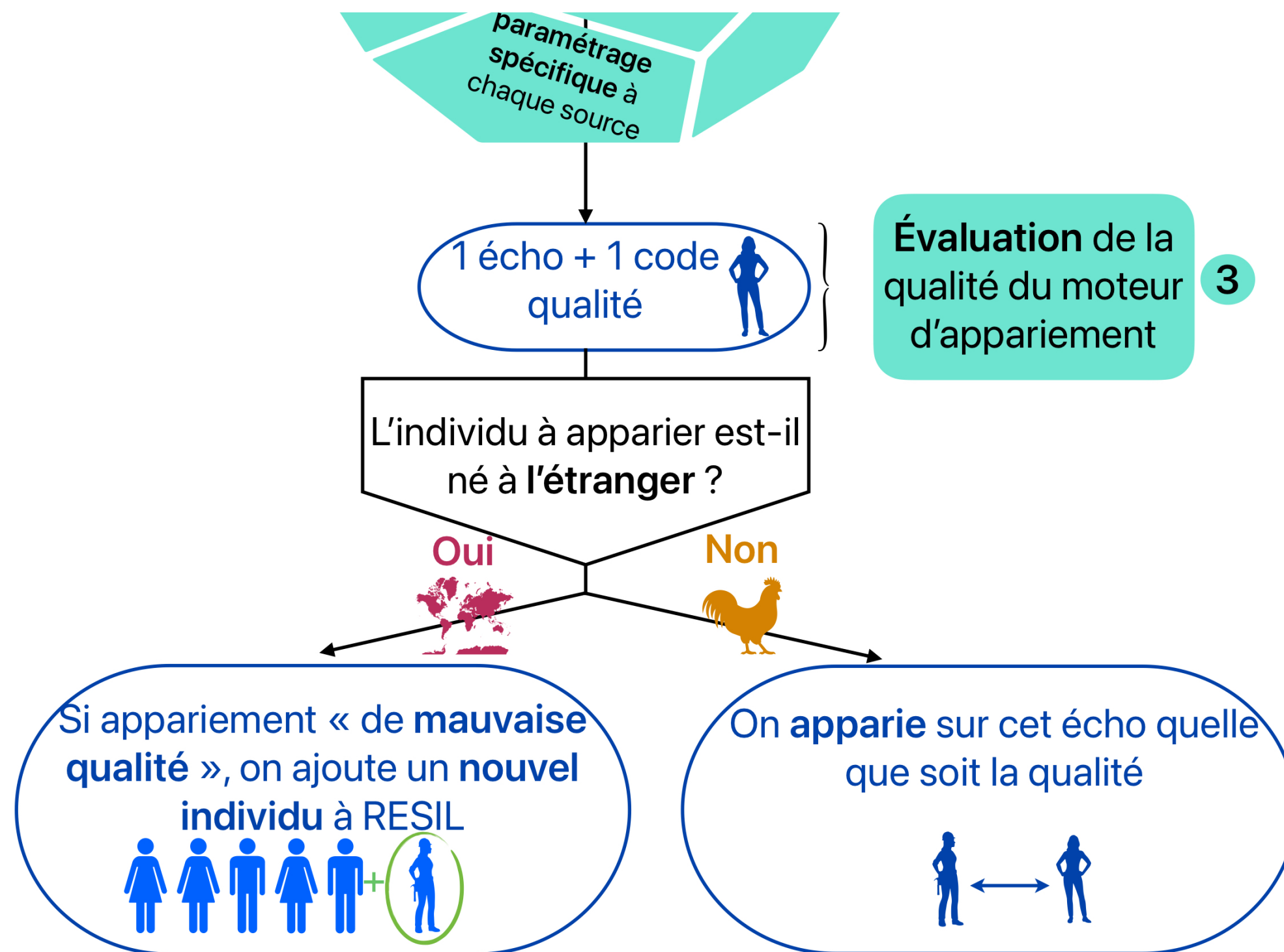
Étape 1 : Filtre des observations à envoyer au moteur



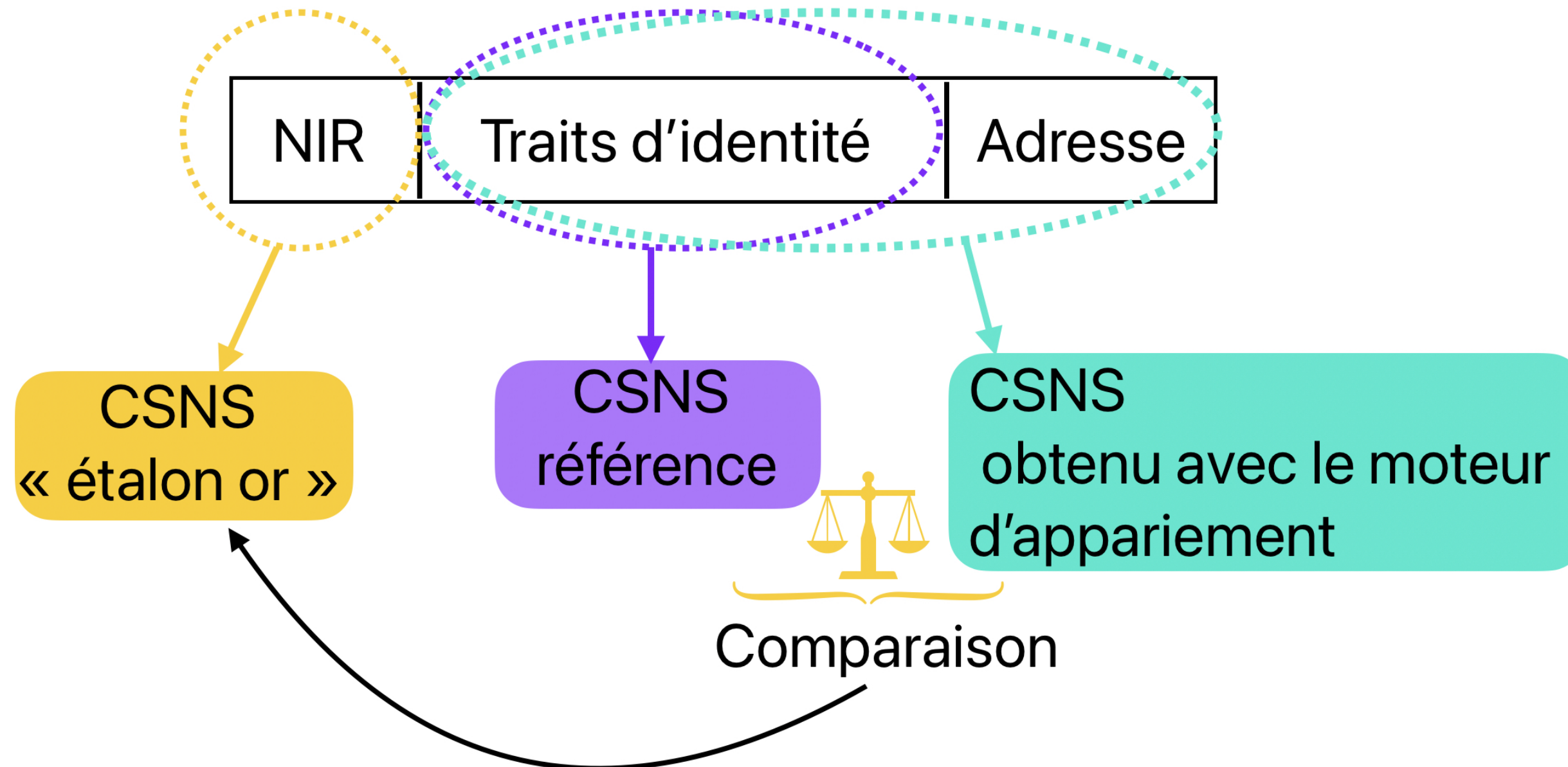
Étape 2 : Outil d'appariement



Étape 3 : Évaluation qualité



Étape 3 : Évaluation qualité : sur source administrative



Étape 3 : Évaluation qualité : avec l'outil d'annotation

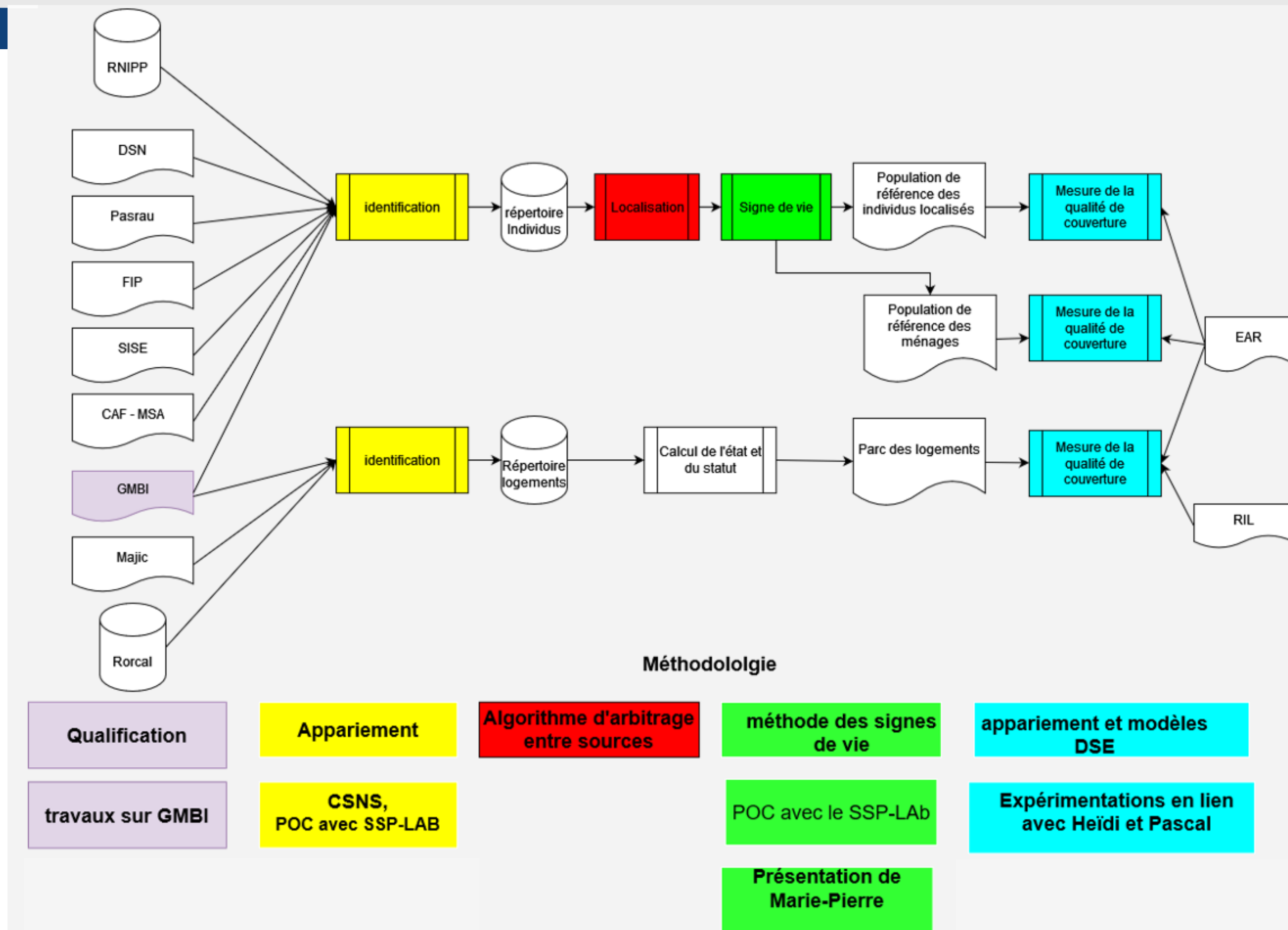
Variable d'intérêt	Valeur observée	Paire potentielle
Nom	TERREL	LE GUILLOUX
Prenoms	FELIX EUGENE MARIE	JEROME EUGENE PIERRE MARIE
Nom naiss		
Dep naiss	99	56121
Lieu naiss		LORIENT
Sexe	m	m
Date naiss	19850926	19850926

Prochain sprint DMCSI - RESIL fin 2024

- Prise en compte des spécificités de la base RESIL définitive
- Amélioration du moteur d'appariement (paramétrage des requêtes Elastic, ...)

INDIVIDUS

LOGEMENTS



Enjeu : comment utiliser la **présence détectée** dans différentes **sources administratives** pour estimer la **population résidente** à un instant t ?

Principe : combiner les informations de présence issues de différentes sources pour estimer la **probabilité que tel individu réside effectivement sur le territoire** → quelle importance accorder à chaque signe de vie ?

ex : la présence dans la source de la caisse des retraites est très peu probable pour un individu de 20 ans, mais très indicative pour un individu de plus de 60 ans.

Difficulté : évaluer le **poids** à accorder à chaque signe de vie

Objectifs de l'expérimentation SSP Lab - Résil : évaluer le **potentiel** d'une poursuite des travaux sur cette thématique et identifier les **pistes à approfondir**

Méthode *Hypothèse : un individu recensé dans une commune une année donnée réside véritablement dans cette commune*

- Modèle qui estime la **probabilité de résider dans une commune** sachant qu'on a observé un ensemble de signes de vie dans différentes sources administratives
- estimation par **apprentissage supervisé** (approche deep learning) pour plus facilement capturer la **complexité des données** et **automatiser** la résolution du modèle

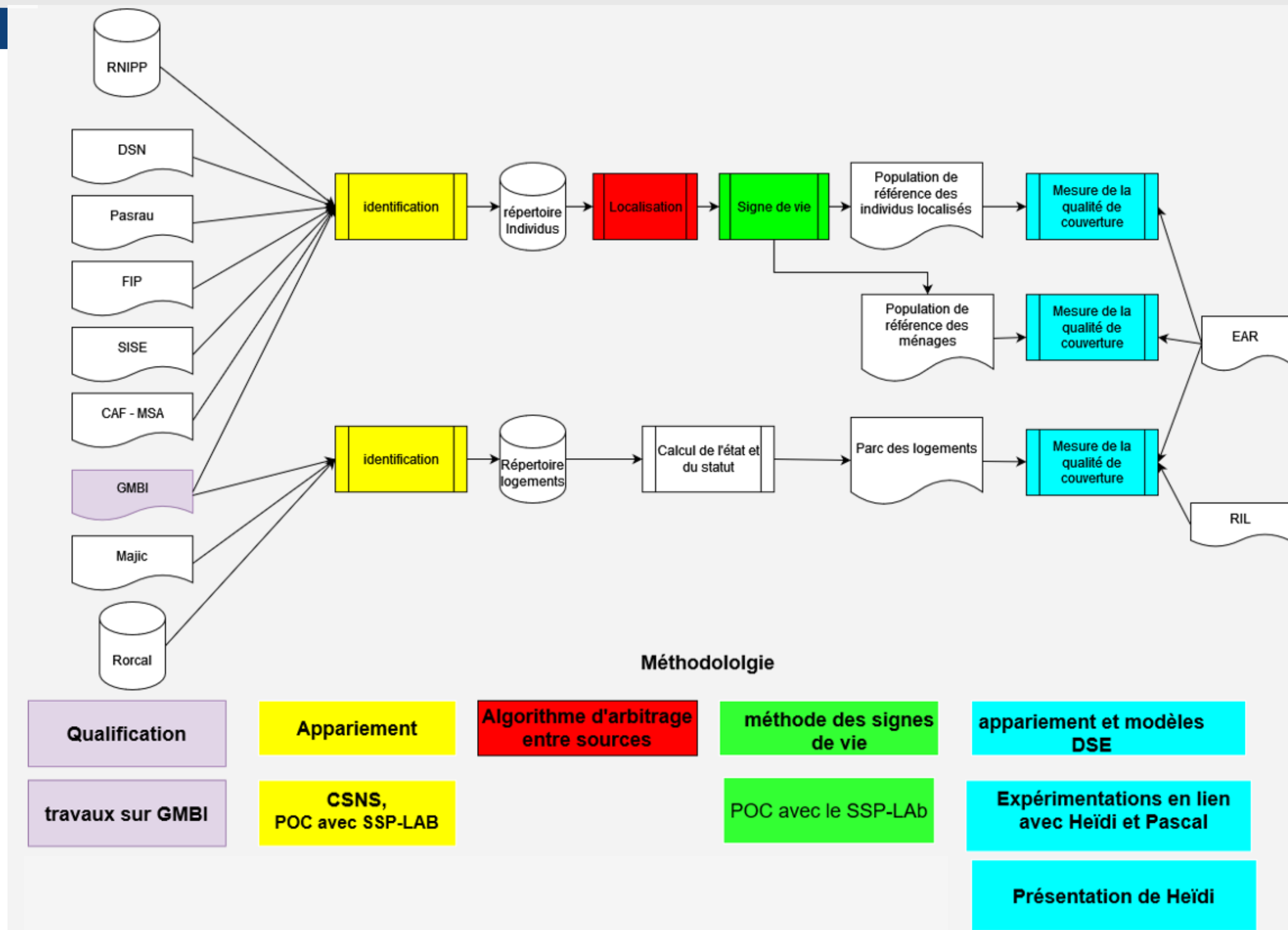
Conclusion :

- le modèle possède un pouvoir prédictif intéressant → une **approche prometteuse**, à poursuivre.
- les résultats pourront être améliorés avec un **meilleur appariement** entre recensement et sources administratives et avec **plus de sources**
- nombreux **raffinements possibles** en complexifiant les signes de vie (variations suivant les caractéristiques individuelles, prorata du nombre de mois de présence, ...)
- travaux repris en septembre 2024 par la DMTR (nouveau poste de data scientist) en collaboration avec Résil.

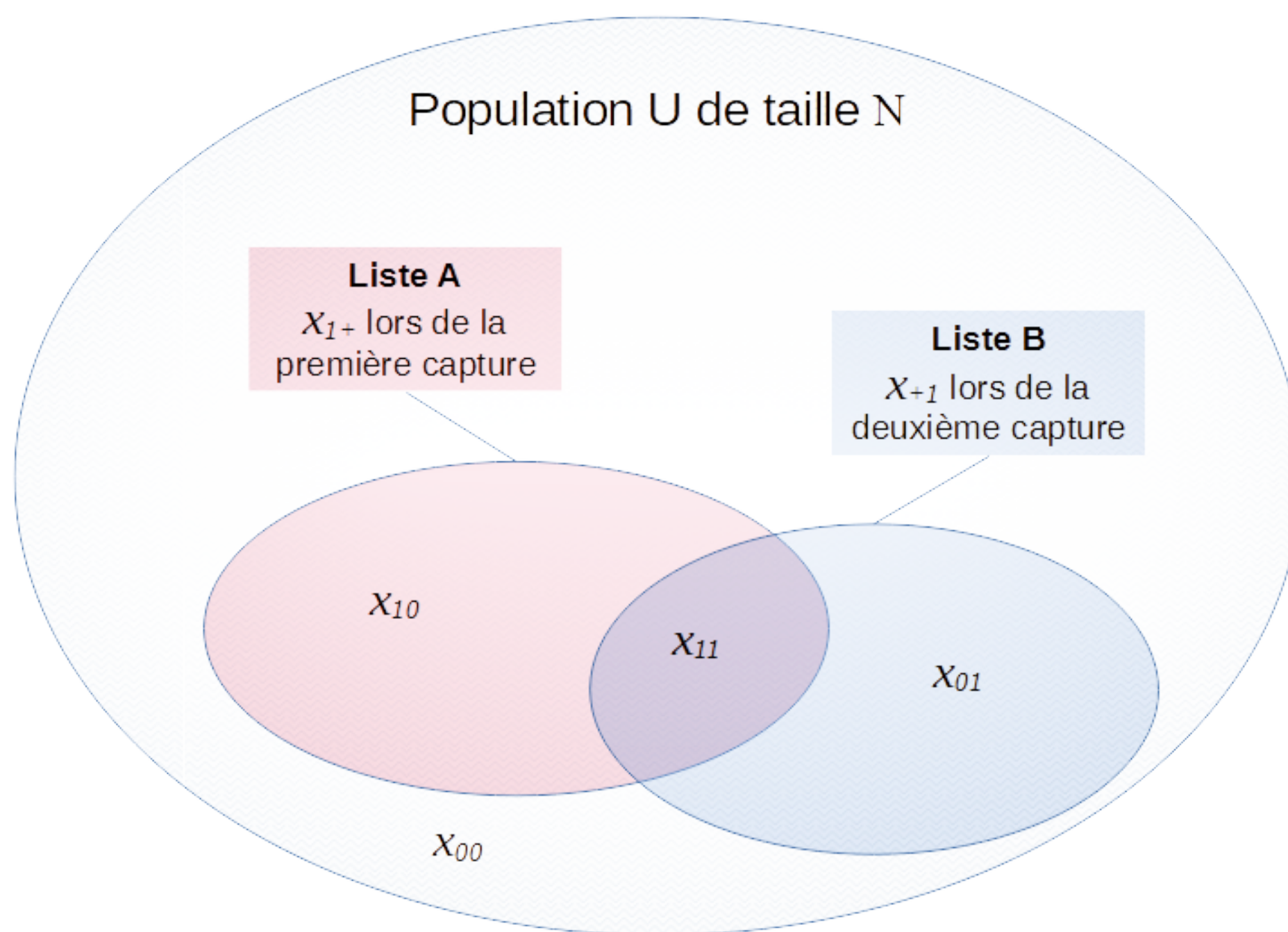
LE PROCESSUS ENVISAGÉ POUR LA DÉFINITION DES UNIVERS DE RÉFÉRENCE

INDIVIDUS

LOGEMENTS



L'objectif de ces méthodes est : **estimer la taille d'une population** sans recensement exhaustif, mais à partir d'énumérations partielles, en tirant partie de la connaissance de l'**intersection de ces énumérations**



$$\hat{N} = \left[\frac{X_{1+} \cdot X_{+1}}{X_{11}} \right]$$

- Mesurer les populations animales, dénombrer des espèces menacées
- Estimer des populations humaines, difficilement mesurables par le biais d'enquêtes : populations rares, contexte illégal ou violent
 - Personnes sans domicile (étude du CepiDC, à partir de ses données et du collectif 'Les morts de la rue' en 2014)
 - Personnes souffrant de maladies rares (étude InVs, incidence de la dengue en France métropolitaine en 2012)
 - Victimes de trafic humain, de crimes contre l'humanité ; Personnes décédées lors de l'attentat du 11 septembre ; Usagers de drogue à risque
- Evaluer la couverture d'une source de données
 - Evaluer la couverture du recensement avec une enquête post censitaire
 - Evaluer la couverture du recensement par des sources administratives (papier de Statistics New Zealand)
- Remplacer un recensement par une estimation de population fondée sur des sources administratives

- Les méthodes d'estimation par système dual reposent sur plusieurs hypothèses très fortes

(celles-ci explicitent les liens entre les listes et la population, et le comportement des individus) :

- homogénéité des probabilités de capture
- indépendance des captures (inter individus, et inter captures)
- population fermée
- absence de surcouverture
- appariement sans erreur entre les énumérations

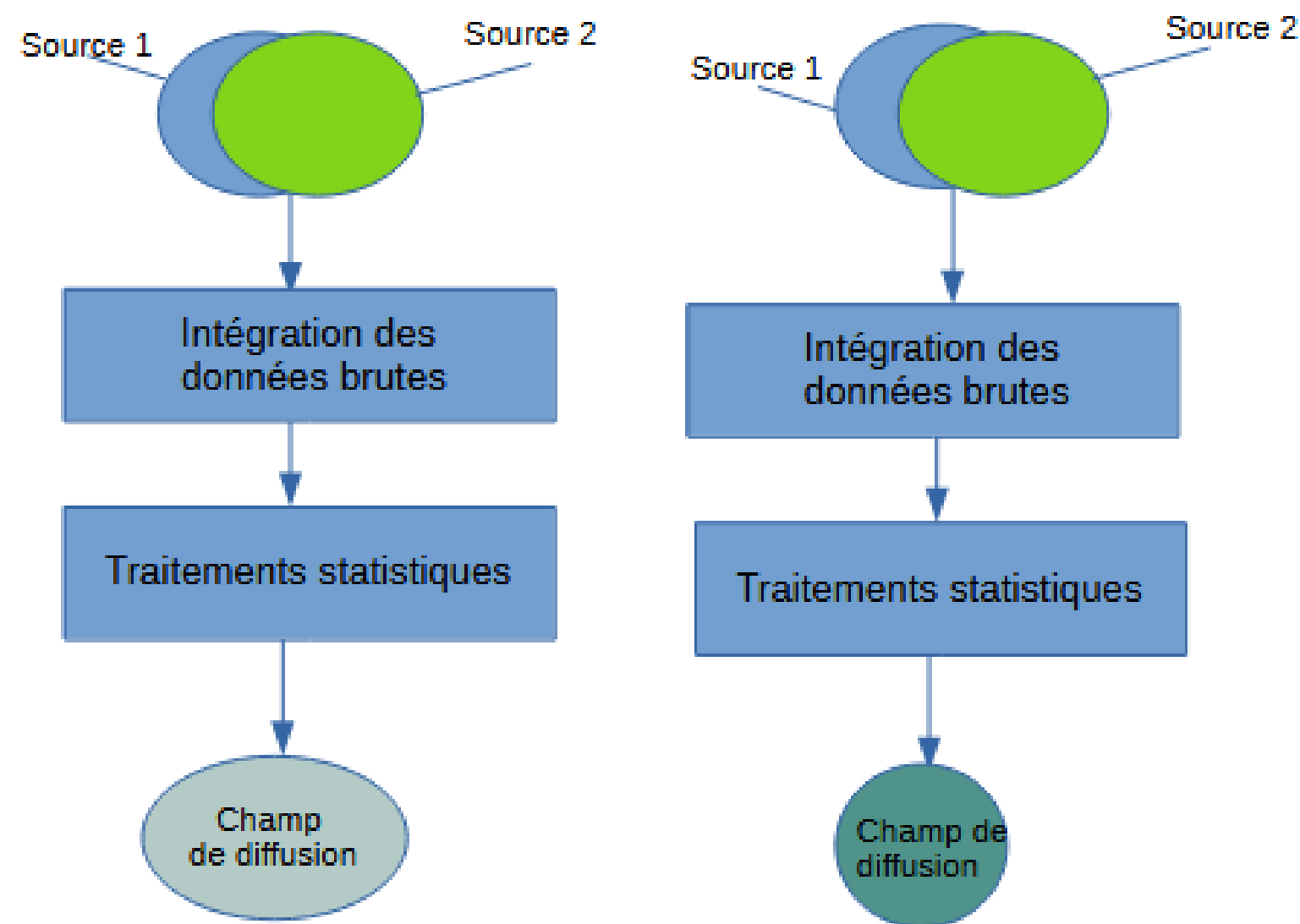
- L'objectif pour Résil est la mesure de la qualité de couverture du répertoire, en estimant un total de population à partir du répertoire et de l'enquête annuelle de recensement (EAR)
- L'application de ces méthodes suppose que l'on s'assure au préalable de la validité des hypothèses, et que l'on mette en œuvre un certain nombre de travaux :
 - Correction de la sur couverture du répertoire
 - Évaluation quantitative des appariements de sources qui constituent le répertoire : faux positifs et faux négatifs
- La collaboration a pris la forme de :
 - GT méthodologique : présentation des méthodes, et échange sur les premières mises en œuvre par l'équipe Résil
 - Rédaction d'un document de travail avec Pascal Ardilly

03

LES PERSPECTIVES

- Au delà de sa construction à partir de données administratives
- Résil offre de nouvelles perspectives dans la production de statistiques
- Mais aussi de nouveaux sujets méthodologiques à étudier

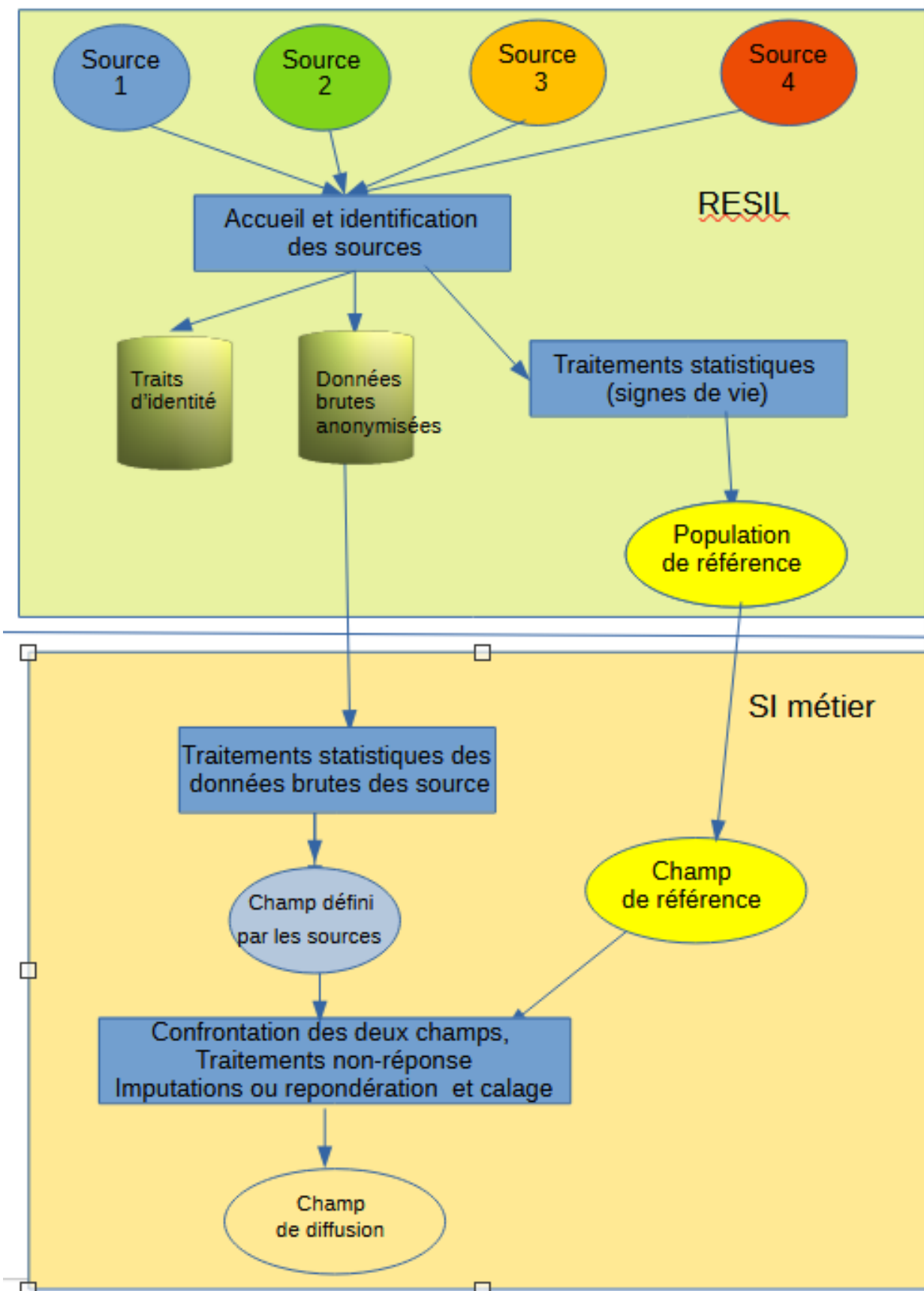
Aujourd'hui champ défini a posteriori par les unités présentes dans les sources externes



Modèle Fidéli

Modèle Filosofi

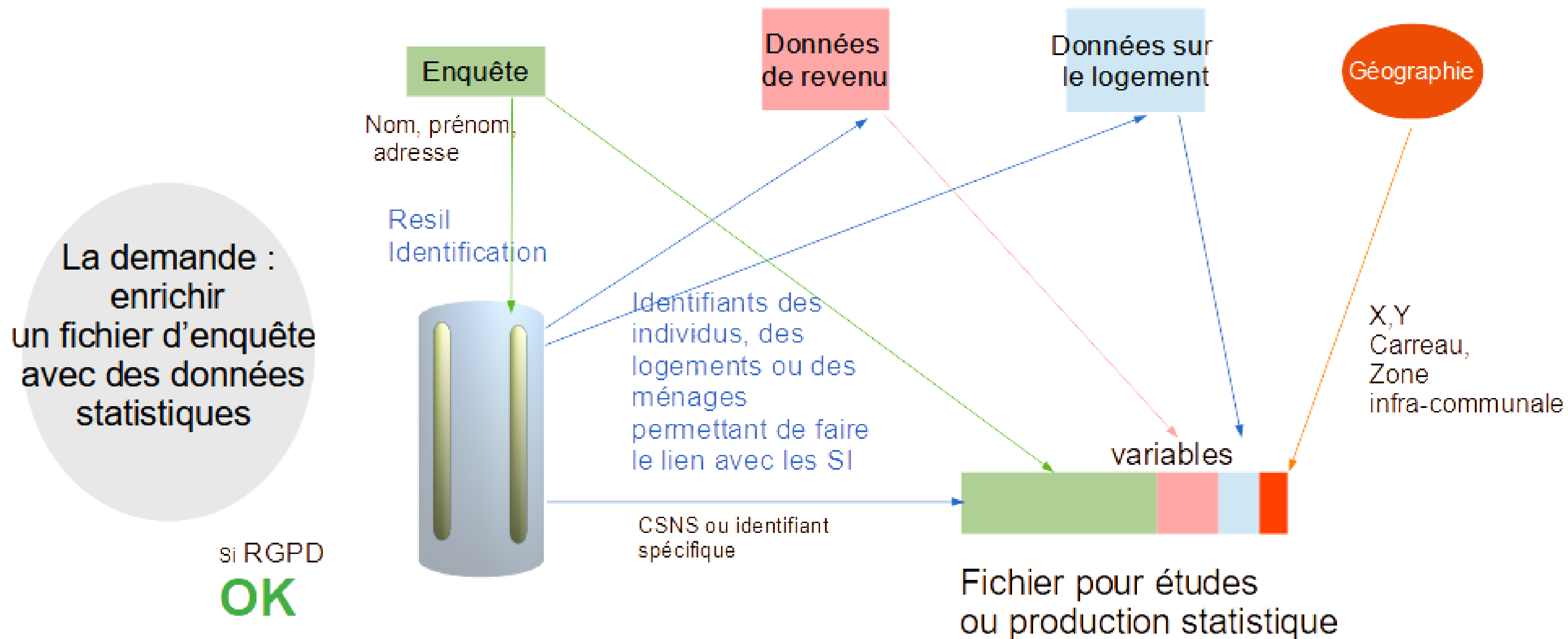
Avec Résil : volonté de définition d'un champ a priori en croisant un maximum de sources externes (signe de vie)



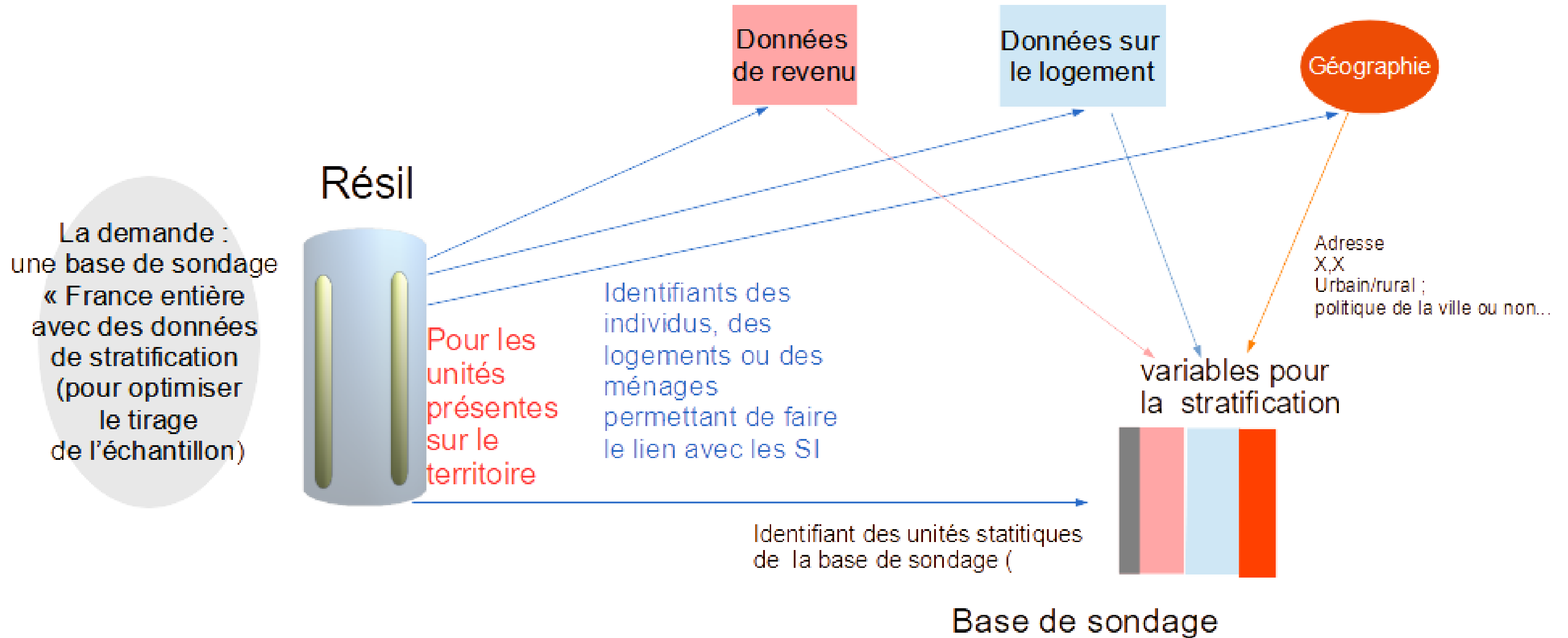
- Résil va produire un univers de référence ...
- ... et des indicateurs qualité issue du modèle DSE

Comment utiliser au mieux ces informations pour le calcul des marges de calage ?

Quels traitements des valeurs manquantes mettre en œuvre ?



LA CONSTRUCTION D'UNE BASE DE SONDAGE



- Résil fournira un fichier avec les variables souhaitées...
- ... et des indicateurs qualité de l'appariement

Comment traiter les valeurs manquantes (un salaire DSN mais pas de revenu du ménage) y compris dans la base de sondage ?

Quels contrôles mettre en œuvre pour repérer les erreurs d'identifications ? Et comment les corriger ensuite ?

Comment prendre en compte les erreurs d'appariement dans les estimations ?

Merci de votre attention

