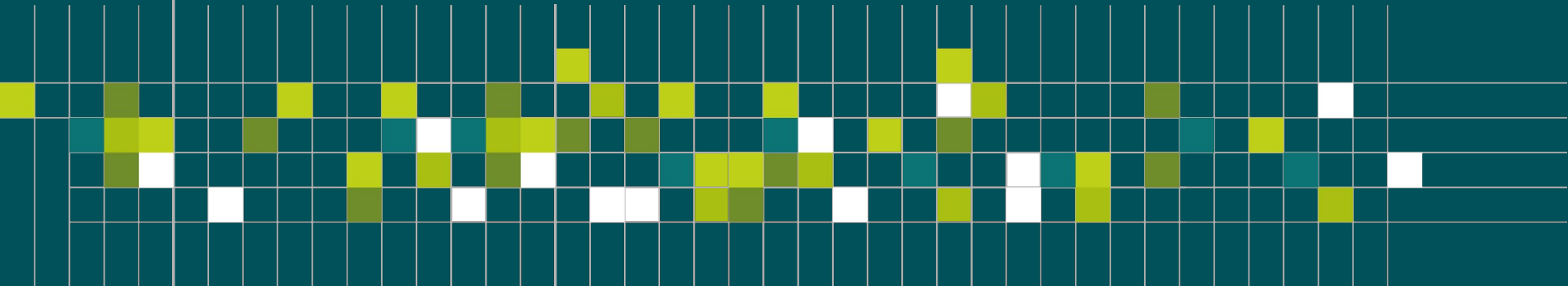
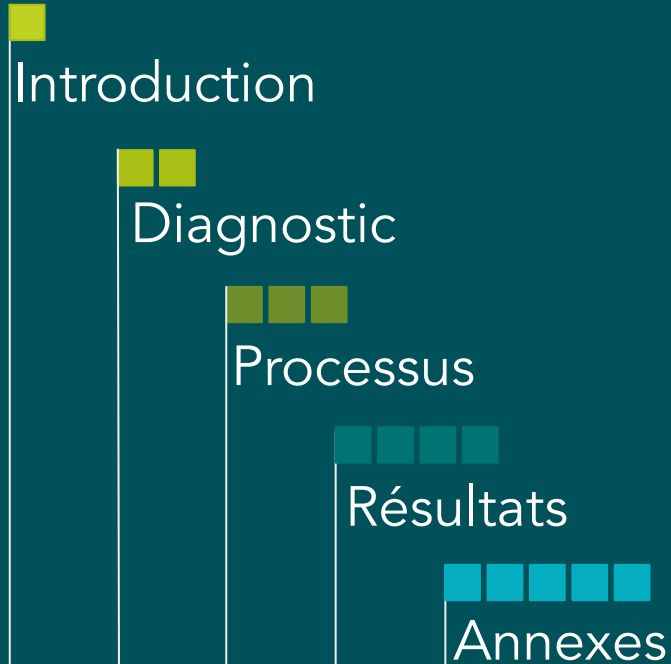


# Intégrité des données IRVE

Proposition d'amélioration

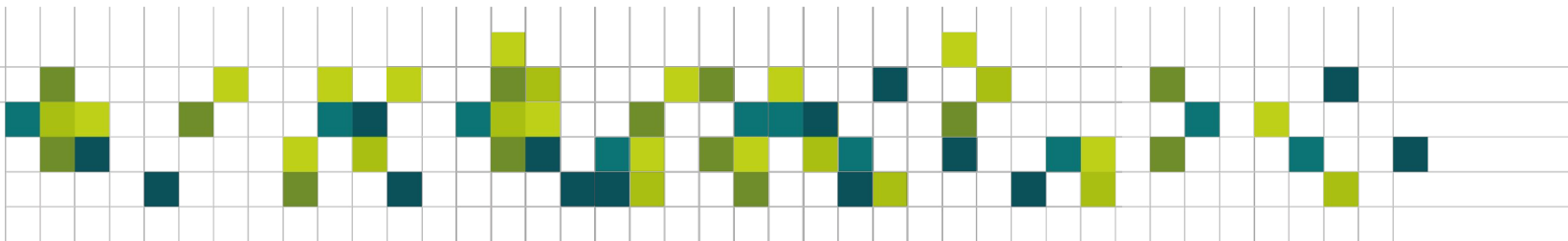




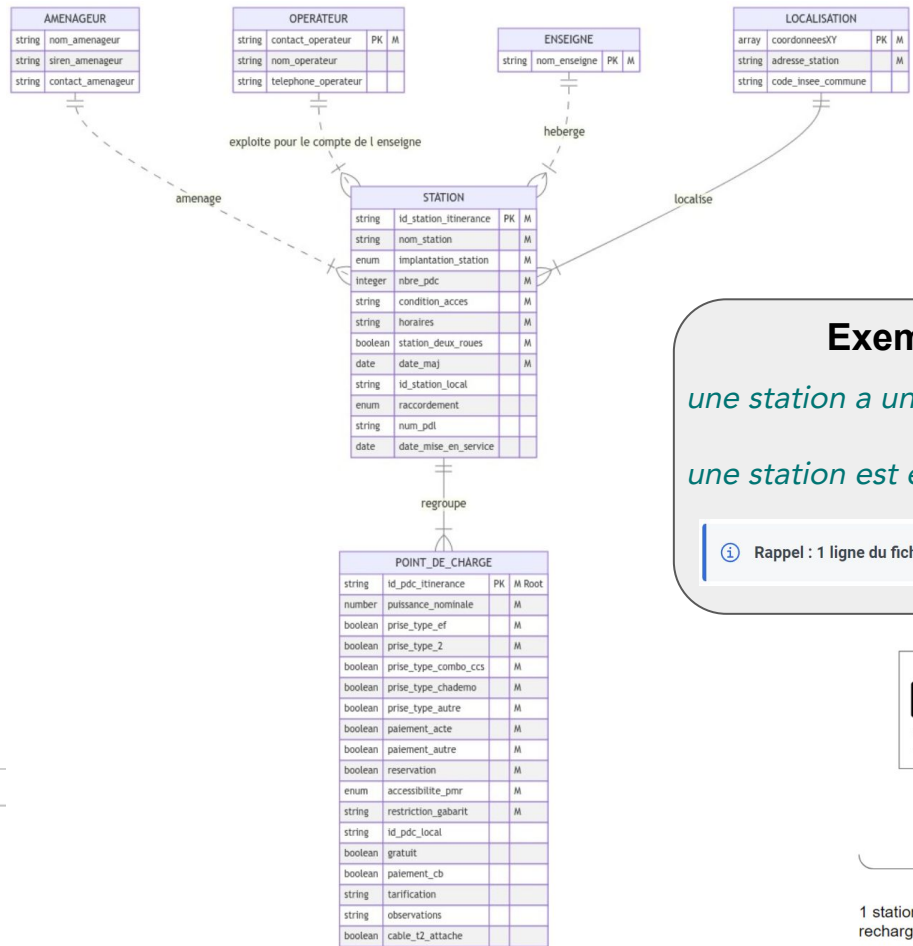
■	Introduction
■ ■	Diagnostic
■ ■ ■	Processus
■ ■ ■ ■	Résultats
■ ■ ■ ■ ■	Annexes

# Pourquoi aborder les données IRVE ?

- Contexte historique
  - Projet "environmental sensing" retenu dans le cadre du programme BlueHats
  - Travaux sur l'interopérabilité et le partage des données
- Echanges "data.gouv"
  - Clarification du rôle des modèles de données dans les jeux de données (cf mise à jour récente des guides)
  - Intégration d'une propriété "relationship" dans les schémas de données (issue TableSchema en cours)
  - Création d'outils de contrôle des relations entre champs des jeux de données tabulaires
- Objectifs
  - Valider l'utilisation d'un modèle de données en complément d'un schéma de données
  - Identifier les apports que pourraient avoir les contrôles d'intégrité (validation des relations entre champs)
- Spécificités IRVE
  - Données et processus de production complexes
  - Questions utilisateurs sur la qualité des données



## Données IRVE



Contexte réglementaire

Structure multi-entités

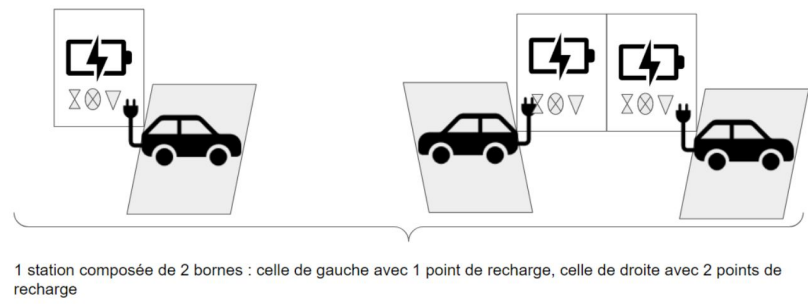
Processus de consolidation et de mise à jour

### Exemple de règles

*une station a une seule localisation*

*une station est exploitée par un seul opérateur*

**Rappel : 1 ligne du fichier de données = 1 point de recharge = 1 id\_pdc\_itinerance**



# Etat des lieux

- Données consolidées
  - 51 000 pdc documentés (50 000 en itinérance)
  - 16 000 stations
- Niveau d'intégrité faible
  - 18 000 lignes présentent une incohérence
  - 32 000 lignes sont cohérentes

```

index - id_pdc_itinerance      16123
contact_operateur - id_station_itinerance  10719
nom_enseigne - id_station_itinerance      7514
coordonneesXY - id_station_itinerance     11825
id_station_itinerance - id_pdc_itinerance  578
nom_station - id_station_itinerance       1865
implantation_station - id_station_itinerance 1245
nbre_pdc - id_station_itinerance         1458
condition_acces - id_station_itinerance    35
horaires - id_station_itinerance          9869
station_deux_roues - id_station_itinerance 10968
adresse_station - coordonneesXY          1360

nombre d'enregistrements sans erreurs : 32553
nombre d'enregistrements avec au moins une erreur : 18116
taux d'erreur : 36 %

```

Les règles d'intégrité ne sont ni exprimées ni contrôlées.

Le processus de mise à jour autorise la conservation de l'historique.

## Doublons?

[Copier le lien vers la discussion](#)


Genevieve Hines

12 avril 2023

Bonjour,

J'aimerais vérifier auprès de vous que je comprends bien les données représentées dans la base.

Le nom des stations n'est pas unique mais les lignes correspondant à un même nom ont souvent des informations identiques mis-à-part par exemple, parfois, le numéro d'identifiant, parfois la date de mise-en-service, à quelques jours près. De telles répétitions doivent elles être interprétées comme de multiples entrées de des mêmes points de charges ou comme des entrées distinctes réelles?  
Toutes mes excuses si j'ai mal lu une explication dans la documentation...

Tous mes remerciements par avance.

## Plusieurs incohérences dans la base

[Copier le lien vers la discussion](#)


Thomas Capdevielle

23 mars 2023

Bonjour,

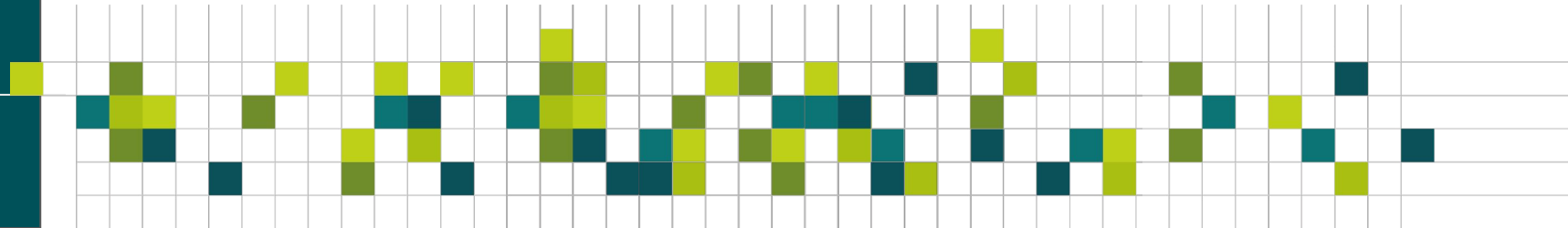
Je me permets de remonter plusieurs points d'interrogation concernant cette base de données.  
En effet après plusieurs analyses je me demande :

- 1 ligne = 1 point de charge ? Il semblerait que dans certains cas cela fonctionne mais pas toujours
- dans certains cas on a une ligne unique pour un id\_station\_itinerance avec nbre\_pdc > 1 ce qui est contradictoire avec l'hypothèse ci-dessus
- si l'on somme simplement le nbre\_pdc on obtient 600k, est-ce normal ?
- différence de 10000 lignes entre deux extractions à 1 mois d'intervalle, quelle est la justification ?

En vous remerciant par avance,

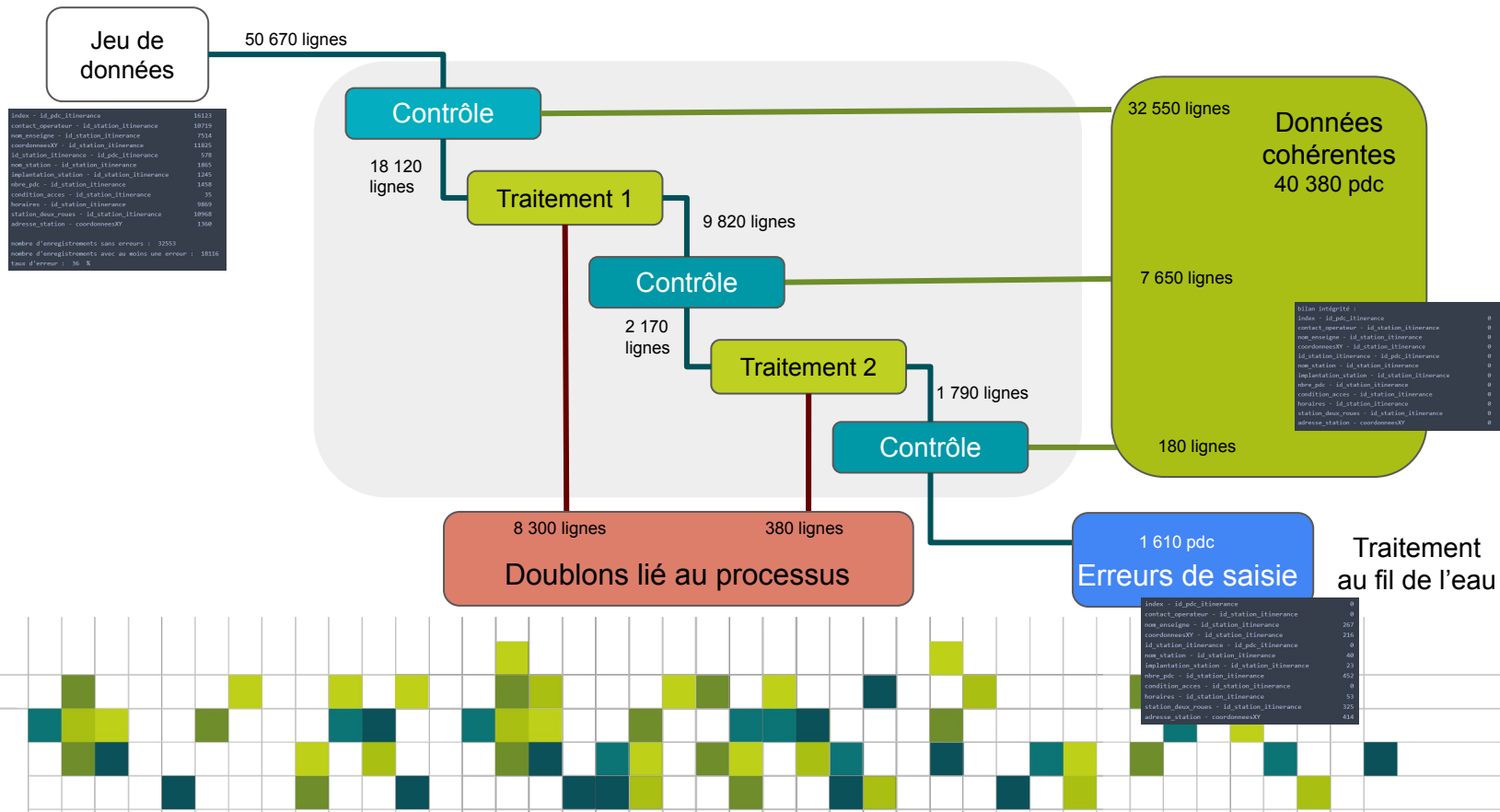
# Méthodologie

- Construire le modèle de données
  - Proposition faite, cohérente avec les données existantes. A valider
- Exprimer les règles d'intégrité
  - Découle du modèle de données. Intégration dans le format du schéma de données
- Mesurer l'état d'intégrité des données
  - Outils existants, mesure effectuée
- Identifier les incohérences
  - Analyse des données effectuée pour tous les écarts
- Traiter les incohérences
  - Effectué pour la majorité des écarts (doublons)
- Intégrer la démarche dans la chaîne de traitement
  - A effectuer



# Mise en cohérence des données

## Processus



# Bilan

## ■ Données

- Données cohérentes (40 000 pdc)
  - Aucune incohérence de structure (reste les éventuelles incohérences de valeurs)
  - Apporte une réponse aux questions posées
- Doublons (8 700 pdc)
  - Anciennes données, à archiver ou supprimer ?
- Données résiduelles ( 1 600 pdc)
  - Représente 500 stations pour 30 opérateurs
  - Erreurs identifiées par catégorie

```

index - id_pdc_itinerance          0
contact_operateur - id_station_itinerance 0
nom_enseigne - id_station_itinerance 267
coordonneesXY - id_station_itinerance 216
id_station_itinerance - id_pdc_itinerance 0
nom_station - id_station_itinerance 40
implantation_station - id_station_itinerance 23
nbre_pdc - id_station_itinerance 452
condition_acces - id_station_itinerance 0
horaires - id_station_itinerance 53
station_deux_roues - id_station_itinerance 325
adresse_station - coordonneesXY 414

```

## ■ Outils

- Outil de contrôle (génère les données résiduelles)
  - À intégrer au traitement quotidien (outil générique disponible via pip)
- Outil de traitement (génère les doublons)
  - A intégrer au traitement quotidien (outil spécifique simple)

### Plusieurs incohérences dans la base

[Copier le lien vers la discussion](#)


Thomas Capdevielle

23 mars 2023

Bonjour,

Je me permets de remonter plusieurs points d'interrogation concernant cette base de données.

En effet après plusieurs analyses je me demande :

- 1 ligne = 1 point de charge ? Il semblerait que dans certains cas cela fonctionne mais pas toujours
- dans certains cas on a une ligne unique pour un id\_station\_itinerance avec nbre\_pdc > 1 ce qui est contradictoire avec l'hypothèse ci-dessus
- si l'on somme simplement le nbre\_pdc on obtient 600k, est-ce normal ?
- différence de 10000 lignes entre deux extractions à 1 mois d'intervalle, quelle est la justification ?

En vous remerciant par avance,

### Doublons?

[Copier le lien vers la discussion](#)


Genevieve Hines

12 avril 2023

Bonjour,

J'aimerais vérifier auprès de vous que je comprends bien les données représentées dans la base.

Le nom des stations n'est pas unique mais les lignes correspondant à un même nom ont souvent des informations identiques mis-à-part par exemple, parfois, le numéro d'identifiant, parfois la date de mise-en-service, à quelques jours près. De telles répétitions doivent elles être interprétées comme de multiples entrées de des mêmes points de charges ou comme des entrées distinctes réelles?

Toutes mes excuses si j'ai mal lu une explication dans la documentation...

Tous mes remerciements par avance.



# Contrôles d'intégrité

## Clefs primaires des entités

- Un pdc est identifié par id\_pdc\_itinerance
- Une station est identifiée par id\_station\_itinerance
- Un opérateur est identifié par contact\_operateur
- Une localisation est identifiée par coordonneesXY
- Une enseigne est identifiée par nom\_enseigne

index - id_pdc_itinerance	0
contact_operateur - id_station_itinerance	0
nom_enseigne - id_station_itinerance	267
coordonneesXY - id_station_itinerance	216
id_station_itinerance - id_pdc_itinerance	0
nom_station - id_station_itinerance	40
implantation_station - id_station_itinerance	23
nbre_pdc - id_station_itinerance	452
condition_acces - id_station_itinerance	0
horaires - id_station_itinerance	53
station_deux_roues - id_station_itinerance	325
adresse_station - coordonneesXY	414

## Contrôles

1. Un pdc est unique et associé à une ligne du tableau
2. Un pdc est intégré dans une et une seule station
3. Une station est opérée par un et un seul opérateur
4. Une station est hébergée par une et une seule enseigne
5. Une station a une et une seule localisation
6. Une station a un et un seul "nom\_station"
7. Une station a une et une seule "implantation\_station"
8. Une station a un et un seul "nbre\_pdc"
9. Une station a un et un seul "condition\_accès"
10. Une station a un et un seul "horaires"
11. Une station a un et un seul "station\_deux\_roues"
12. Une localisation correspond à une et une seule "adresse\_station"

## Points non contrôlés

- Les champs associés au pdc sont implicitement validés par l'unicité du pdc
- Les champs facultatifs ne font l'objet d'aucun contrôle
- Les données hors itinérance ne sont pas prises en compte (modèle de données non applicable -> à faire dans un second temps)
- La cohérence du champ "nbre\_pdc" avec les pdc n'est pas traitée (voir Annexe spécifique)
- L'unicité de la date de mise à jour par station n'est pas prise en compte



# Traitement des doublons

date 1

Station 1

pdcc1  
pdcc2  
pdcc3

date 2

Station 1

pdcc1  
pdcc3  
pdcc4

Suppression des doublons pdc  
*(on garde les pdc avec la date la plus grande)*

Suppression des doublons stations  
*(on garde les pdc de la station avec la date de la station la plus grande)*

station	pdc	date
station1	pdcc1	date 1
station1	pdcc2	date 1
station1	pdcc3	date 1
station1	pdcc1	date 2
station1	pdcc3	date 2
station1	pdcc4	date 2

station	pdc	date
station1	pdcc2	date 1
station1	pdcc1	date 2
station1	pdcc3	date 2
station1	pdcc4	date 2

Etat incohérent

station	pdc	date
station1	pdcc1	date 2
station1	pdcc3	date 2
station1	pdcc4	date 2

La méthode est valide si (1) tous les pdc ont une date de mise à jour valide (2) les mises à jour se font sur une station complète (la mise à jour est associée à la station et tous les pdc de la station portent la même date de mise à jour)

Nota : la gestion par date ne garantit pas la cohérence des mises à jour sur les entités ou champs associés aux stations. Par exemple, si deux stations partagent la même localisation (ex. Parking) et que l'on modifie l'adresse d'une station, l'adresse de la seconde ne sera plus cohérente. La modification d'adresse doit donc se faire sur tous les pdc de toutes les stations du parking.



# Champ 'nbre\_pdc'

## Analyse :

- Avant purge des doublons, on a 24 000 valeurs du champ erronées sur 51 000 pdc
- Après purge des doublons, on a 9 900 valeurs du champ erronées sur 11 000 pdc qui présentent au moins un écart
- Sur ces 9 900 valeurs, 5 400 correspondent à des stations avec un seul pdc mais où 'nbre\_pdc' > 1
  - Cas d'erreur avec id\_station identique à id\_pdc : Si plusieurs stations ont la même localisation
  - Cas d'erreur avec un seul pdc fictif documenté : Si aucune station n'a la même localisation

## Synthèse :

- Ce champ est mal documenté et pourra difficilement être corrigé
- Il serait plus pertinent de le remplacer par un champ calculé (une seule ligne de code)
  - Exemple : `data['nb_pdc_calcul'] = data.groupby('id_station_itinerance')['index'].transform('count')`

# Données produites

## Fichiers de données

- Fichier csv initial dupliqué avec deux champs supplémentaires
  - lignes\_a\_corriger (booléen)
  - Doublons\_a\_supprimer (booléen)
- Fichier csv des lignes à corriger avec un champ (booléen) par contrôle (12 champs)
- Fichier csv des doublons à supprimer

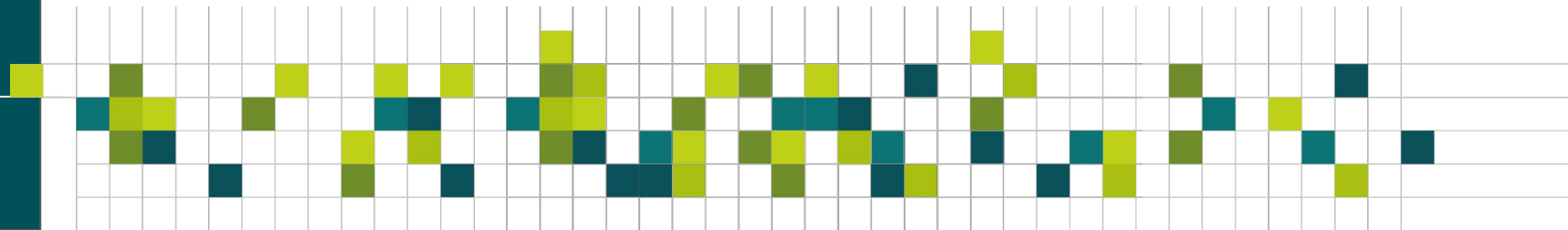
## Documentation

- Modèle de données
- Liste des contrôles au format du schéma de données (voir Notebook)
- Présentation de l'analyse

## Outils

- Outil d'analyse d'intégrité (bibliothèque open-source, accessible via pip)
- Outil Pandas d'élimination des doublons (fonction python simple)
- Programme de production des données de sortie (Jupyter Notebook)

Ces éléments sont accessibles sur le repository GitHub: <https://github.com/loco-philippe/Environmental-Sensing/tree/main/python/Validation/irve/Analyse>



# Points à valider

Modèle de données (uniquement pour les champs obligatoires)

- Opérateur, enseigne : Est-ce que le modèle proposé est correct pour 'nom\_enseigne' et 'contact\_operateur' (pas de relations entre opérateur et enseigne) ?
- Localisation : Est-ce que pour une coordonnée géographique ('coordonneesXY' , il y a bien une seule adresse ('adresse\_station') ?
- Station :
  - Est-ce qu'une station a bien une seule localisation ('coordonneesXY') ?
  - Est-ce qu'une station est bien associée à une seule enseigne ('nom\_enseigne') ?
  - Est-ce qu'une station a bien une seule localisation ('coordonneesXY') ?

Processus de mise à jour

- Est-ce que le champ 'date\_maj' est un champ associé à la station ou bien au pdc ?
  - Cf traitement des doublons (si c'est un champ associé au pdc, il faut préciser comment on supprime un pdc d'une station)
  - S'il est associé à la station, on doit contrôler qu'une station ne peut avoir plusieurs pdc avec des dates différentes
- Est-ce que les mises à jour s'effectuent par station complète (les nouveaux pdc remplacent les précédents) ou bien par pdc (dans ce cas il faut préciser comment on maintient la cohérence de la station : voir question précédente) ?
- Est-ce qu'il y a un mécanisme particulier pour les mises à jour des stations associées à la même localisation (ex. stations d'un même parking sur plusieurs étages) ?

Processus de mise à disposition

- Est-ce qu'il y a une contrainte qui impose de conserver les anciennes versions dans les données mises à disposition ?

