

Guide de prise en main de la base du dossier complet et du comparateur de territoires

Insee

22 juillet 2026

Ce guide présente la [base du dossier complet](#) et la [base du comparateur de territoires](#), avec quelques exemples d'utilisation.

La base du dossier complet et du comparateur de territoires

Sur insee.fr, le [Dossier complet](#) présente les principaux indicateurs diffusés par l'Insee, classés par thème, pour le territoire de son choix (une commune, un département, un EPCI, etc.). Le dossier complet est mis à jour à chaque nouvelle diffusion de données, et permet donc de visualiser les données disponibles les plus récentes.

Chaque tableau ou graphique du dossier complet est alimenté à partir d'un ou plusieurs jeux de données accessible sur [le catalogue de données \(Melodi\)](#) : par commodité, la **base du dossier complet** propose d'agréger toutes ces données en un unique fichier.

Cette base permet notamment de récupérer :

- Toutes les données du dossier complet pour un ou plusieurs territoires de son choix
- Les données associées à un tableau ou graphique du dossier complet, à l'aide de son identifiant

La base est proposée aux formats CSV et Parquet : pour travailler de manière plus performante sur ce fichier volumineux, il est recommandé d'utiliser la version au format Parquet. Pour plus d'informations sur le format **Parquet**, se référer à [Courrier des statistiques - Quels formats pour quelles données ?](#), pour des exemples d'utilisation de données dans ce format, se référer au [Guide d'utilisation des données du recensement de la population au format Parquet](#)

Suivant le même principe, la [base du comparateur de territoires](#) est également proposée : le [comparateur de territoires](#) permet de comparer plusieurs territoires sur une sélection d'indicateurs. La base du comparateur de territoires est un extrait de la base du dossier complet.

Liste des variables

Variable	Libellé
GEO_REF	Millésime géographique de référence
GEO_OBJECT	Niveau géographique - Code
GEO_OBJECT_LABEL (*)	Niveau géographique - Libellé (*)
GEO	Géographie - Code
GEO_LABEL	Géographie - Libellé
TIME_PERIOD	Période temporelle
ID_TAB	Identifiant du tableau du dossier complet - Code
ID_TAB_LABEL (*)	Identifiant du tableau du dossier complet - Libellé (*)
DS	Jeu de données source du catalogue Melodi
TAB_MEASURE	Mesure du tableau - Code
TAB_MEASURE_LABEL (*)	Mesure du tableau - Libellé (*)
OBS_VALUE	Valeur de l'observation
OBS_STATUS	Statut de l'observation
UNIT_MULT	Quantité d'unités (en puissance de 10)
UNIT_MEASURE	Unité de mesure
CONF_STATUS	Statut de confidentialité

(*) Par commodité d'utilisation, les libellés associés à des codes ont été intégrés directement dans le fichier au format Parquet. Au format CSV, les libellés sont présents dans le fichier metadata.csv, et nécessite au besoin de faire une jointure pour les récupérer.

Structure de la base

La base du dossier complet contient l'ensemble des données permettant de reproduire les tableaux et graphiques du dossier complet. Les indicateurs visibles dans le dossier complet sont soit directement présents dans la base, soit peuvent nécessiter un calcul à partir de plusieurs données de la base.

Exemple : dans le tableau POP T1, pour obtenir la part de la population de moins de 15 ans, la base du dossier complet fournit la population de moins de 15 ans (numérateur) et la population totale (dénominateur), permettant ainsi de recalculer la part de la population de moins de 15 ans sur un ou plusieurs territoires de son choix.

Comme l'ensemble des jeux de données du catalogue de données Melodi, la base du dossier complet est présentée sous forme de cube (cf. [Courrier des statistiques N11 - Faciliter l'accès aux données de l'Insee - Cubes, catalogue et métadonnées](#)). Cette présentation, dite en **format long**, est optimale pour le stockage, l'analyse et le filtrage des données. Pour certains besoins il peut être utile de passer in fine les données en **format large** : une ligne par territoire, une colonne par indicateur. Un exemple pour passer d'un format à l'autre est fourni ci-après.

Quelques exemples d'utilisation

Les exemples de code ci-après sont formulés en SQL, applicables notamment dans le cadre d'une utilisation du fichier au format Parquet avec duckdb.

1) Récupérer toutes les données du dossier complet pour mon territoire

Un territoire (une commune, un département, une région, etc.) est identifié de manière unique par son niveau géographique (GEO_OBJECT), et son code géographique (GEO).

Ainsi, si je souhaite récupérer les données de la commune de Nantes, je filtre sur le niveau géographique Commune GEO_OBJECT='COM' et sur le code géographique 44109 GEO='44109' :

```
SELECT
    *
FROM
    base_dossier_complet
WHERE
    GEO_OBJECT = 'COM' -- Niveau géographique : Commune
    AND GEO = '44109' -- Code commune de Nantes
```

Pour sélectionner toutes les données de la commune de Nantes, de l'EPCI de Nantes métropole et du département de Loire-Atlantique :

```

SELECT
    *
FROM
    base_dossier_complet
WHERE
    (
        GEO_OBJECT = 'COM' -- Niveau géographique : Commune
        AND GEO = '44109' -- Code commune de Nantes
    )
    OR (
        GEO_OBJECT = 'EPCI' -- Niveau géographique : EPCI
        AND GEO = '244400404' -- Code géographique Nantes métropole
    )
    OR (
        GEO_OBJECT = 'DEP' -- Niveau géographique : Département
        AND GEO = '44' -- Code géographique de Loire-Atlantique
    )

```

À noter que la variable GEO_REF contient le millésime géographique de diffusion de la donnée, information à étudier avec vigilance notamment si les territoires étudiés ont connu des changements de géographie (typiquement fusion ou scission de commune).

2) Récupérer les données d'un tableau du dossier complet

Le [dossier complet](#) est composé de tableaux et graphiques identifiés de manière unique par un code visible en titre :



FIGURE 1 : Identifiant d'un tableau du dossier complet

Pour obtenir uniquement les données associées à un tableau, je filtre la base à l'aide de la variable ID_TAB :

```
SELECT
    *
FROM
    base_dossier_complet
WHERE
    GEO_OBJECT = 'COM' -- Niveau géographique : Commune
    AND GEO = '44109' -- Code commune de Nantes
    AND ID_TAB = 'ACT_T4' -- Identifiant du tableau du dossier complet ACT T4
```

La variable DS (pour dataset) permet d'identifier le jeu de données d'origine de la donnée. À l'aide de cet identifiant il est ainsi possible de consulter la documentation et les métadonnées complètes du jeu de données via le [catalogue de données Melodi](#).

The screenshot shows the Insee website's data catalog. The URL bar at the top contains 'catalogue-donnees.insee.fr/lr/catalogue/recherche/DS_RP_EDUCATION_PRINC', with 'DS_RP_EDUCATION_PRINC' highlighted by a red box. The main header features the Insee logo and navigation links. Below the header, a breadcrumb trail reads 'Accueil insee.fr / Catalogue / Éducation - Scolarisation - Principaux indicateurs (dossier complet)'. A large blue banner displays the title 'Éducation - Scolarisation - Principaux indicateurs (dossier complet)' with the Melodi logo. Below this, a table with three tabs: 'Description', 'Métadonnées', and 'Données'. The 'Description' tab is active, showing the dataset identifier 'Identifiant du jeu de données : DS_RP_EDUCATION_PRINC', which is also highlighted by a red box. Below the identifier, the title, subtitle, and summary of the dataset are provided.

FIGURE 2 : Catalogue Melodi : identifiant du jeu de données

3) Passer les données en format large (en colonnes)

Supposons que j'ai filtré la base sur un territoire (Nantes), sur un tableau du dossier complet (RFD G1) et deux années (2023, 2024). Le résultat obtenu est au format long :

```
SELECT TAB_MEASURE_LABEL, TIME_PERIOD, OBS_VALUE
FROM base_dossier_complet
WHERE GEO = '44109' AND GEO_OBJECT = 'COM'
AND ID_TAB = 'RFD_G1'
AND TIME_PERIOD IN ('2023', '2024')
ORDER BY TIME_PERIOD
```

TAB_MEASURE_LABEL	TIME_PERIOD	OBS_VALUE
Nombre de naissances vivantes	2023	4,044
Nombre de décès	2023	2,179
Nombre de naissances vivantes	2024	4,051
Nombre de décès	2024	2,213

Je souhaite passer le résultat au format large, c'est à dire une colonne par année.

Pour cela j'applique un pivot, qui permet de créer autant de colonnes que de millésimes de données (TIME_PERIOD) disponibles :

```
WITH data AS (
  SELECT TAB_MEASURE_LABEL, TIME_PERIOD, OBS_VALUE
  FROM base_dossier_complet
  WHERE GEO = '44109' AND GEO_OBJECT = 'COM'
  AND ID_TAB = 'RFD_G1'
  AND TIME_PERIOD IN ('2023', '2024')
)
PIVOT data
ON TIME_PERIOD -- passage des années en colonne
USING sum(OBS_VALUE) -- contenu en colonne
```

TAB_MEASURE_LABEL	2023	2024
Nombre de naissances vivantes	4,044	4,051
Nombre de décès	2,179	2,213

3.2) Passer les données en format large en R

En R, suivant le même principe, le package `tidyr` peut être utilisé pour pivoter les données au format large :

TAB_MEASURE_LABEL	TIME_PERIOD	OBS_VALUE
Nombre de naissances vivantes	2023	4,044
Nombre de décès	2023	2,179
Nombre de naissances vivantes	2024	4,051
Nombre de décès	2024	2,213

```
data_pivot <- data |>
  tidyr::pivot_wider(
    id_cols = c("TAB_MEASURE_LABEL"),
    names_from = c('TIME_PERIOD'),
    values_from = OBS_VALUE
  )
```

TAB_MEASURE_LABEL	2023	2024
Nombre de naissances vivantes	4,044	4,051
Nombre de décès	2,179	2,213