

Faciliter l'accès aux données de l'Insee

Cubes, catalogue et métadonnées



Jocelyne Mauguin* et Nicolas Sagnes**

L'Insee publie sur son site une très grande quantité de données couvrant de nombreux thèmes économiques et sociaux, comme la démographie, l'emploi, les comptes nationaux ou encore les indices de prix. Face à la richesse de son offre, l'Insee doit accompagner son public dans le choix et la compréhension de la donnée. Présenter l'ensemble des statistiques de manière simple et cohérente sur *insee.fr* est donc un défi important. Un premier niveau de diffusion, la datavisualisation, permet d'appréhender un sujet donné à travers des indicateurs synthétiques sous forme de tableaux et de visuels simples, clairs et faciles à comprendre. Mais pour aller plus loin dans l'analyse, des données plus détaillées sont mises à disposition. Elles se présentent généralement sous une forme agrégée : des cubes multidimensionnels croisent différentes variables d'intérêt comme le genre, l'âge ou la catégorie socioprofessionnelle dans les enquêtes ménages. L'enjeu est alors de proposer ces informations dans des formats libres et bien normés, mais également de bien les documenter, en s'appuyant sur des standards internationaux. Ces données doivent aussi être bien cataloguées pour en faciliter la découverte. Pour y accéder, les services de consultation de la donnée de l'Insee se modernisent avec la possibilité de naviguer dans ces cubes. Enfin, les données doivent être accessibles tant par les internautes que par les machines qui les moissonnent : l'usage de ces dernières ouvre des perspectives de nouveaux modes de consommation de la donnée grâce à l'intelligence artificielle.

 INSEE (National Institute of Statistics and Economic Studies) publishes on its website a vast amount of data covering numerous economic and social themes such as demography, employment, national accounts, and price index. Given the richness of its website, INSEE must guide its users towards their desired data. Presenting all statistical data in a simple and coherent manner on the *insee.fr* website is therefore a significant challenge. A first level of dissemination, which is data visualisation, allows to comprehend a given subject through synthetic indicators presented as simple, clear, and easy-to-understand visuals. However, to go further in the analysis, more detailed data are made available. These are typically presented in an aggregated form: multidimensional cubes that cross-reference various variables of interest such as gender, age, or socio-professional category in household surveys. The challenge then becomes offering these informations in well-standardized and open-source formats, while also thoroughly documenting them, relying on international standards. These data must also be well-catalogued to facilitate discovery. To access them, INSEE's data consultation services are being updated to make it possible to navigate through these cubes. Finally, the data must be accessible both to internet users and to machines that harvest them: the use of the latter opens up new prospects for data consumption modes through artificial intelligence.

* Cheffe de projet statistique, Insee, DDAR.
jocelyne.mauguin@insee.fr

** Directeur de projet, Insee, DDAR.
nicolas.sagnes@insee.fr

L'Insee publie sur son site un très grand nombre de données de référence d'information économique et statistique. Avec le développement de la donnée, un enjeu très fort est de rendre cette offre la plus à jour, lisible et accessible. Pour cela, de nombreux défis sont à relever : ils portent sur l'importance des formats de données, la documentation et ses standards mais aussi sur les services comme la datavisualisation, le catalogage ou encore la navigation dans les données, sans oublier les APIs¹ indispensables à l'utilisation par des machines.

► Présenter simplement de très nombreuses données —

L'Insee diffuse des statistiques sur de nombreuses thématiques comme la démographie, l'emploi, les comptes nationaux ou encore les indices de prix. Celles-ci constituent une large part des statistiques publiques, les autres étant produites essentiellement par les services statistiques ministériels. Ces statistiques sont essentielles pour l'élaboration d'études économiques, afin d'éclairer des questions structurantes², décomposition de l'inflation, facteurs de la pauvreté, etc.

Cette diversité se traduit par un volume très important de données, dans le sillage de l'explosion de la data au cours des dernières décennies³. Quelques chiffres : l'Insee diffuse chaque année sur son site environ 5 000 fichiers XLSX⁴ ou encore 70 000 séries historiques (par exemple la série du produit intérieur brut depuis 1949 ou les séries mensuelles des indices de prix à la consommation).

Face à une telle richesse de l'offre, l'Insee doit accompagner ses utilisatrices et utilisateurs, dénommés par la suite « insee-nautes », dans le parcours vers la donnée souhaitée. La donnée doit être facilement trouvée et comprise. Pour cela, l'Insee s'efforce de suivre les grands principes du Code de bonnes pratiques de la statistique européenne⁵,

pierre angulaire du cadre qualité commun aux instituts statistiques européens. La cohérence/comparabilité et l'accessibilité/clarté sont les principes essentiels de la diffusion de statistiques :

- **Cohérence et comparabilité** : la comparaison des données sur une période raisonnable est possible ; les statistiques sont élaborées sur la base de normes communes pour les définitions, les unités et les nomenclatures dans les différentes enquêtes et sources de données.

- **Données accessibles et claires** : les statistiques sont présentées avec une documentation pour les interpréter correctement et les comparer utilement ; des technologies, des méthodes et des plateformes d'information



**L'Insee
accompagne
ses utilisatrices
et utilisateurs
dans le choix
et la compréhension
de la donnée.**



¹ *Application Programming Interface* ou Interface de Programmation d'Application en français. On parle aussi de service web ou web service. Le site [insee.fr](https://www.insee.fr) propose actuellement un service web dont le résultat respecte le standard international SDMX.

² Voir (European Commission, 2015).

³ Voir par exemple le site <https://project.opendatamonitor.eu/>.

⁴ XLSX est une extension de nom de fichier pour tableur au format Office Open XML utilisé par Microsoft Office à partir de la version 2007.

⁵ <https://www.insee.fr/fr/information/4140105>.

et de communication modernes sont utilisées ; des normes de données ouvertes sont proposées, avec un accès dans un format non propriétaire (Ubaldi, 2013 ; Emilsson et alii, 2020).

Pour appliquer ces principes, il faut aussi tenir compte de la grande diversité des profils et attentes des insee-nautes. Citons quelques exemples : une étudiante doit réaliser un exposé sur la comptabilité nationale et a juste besoin de consulter un tableau des grands agrégats comptables (PIB, valeur ajoutée, etc.) sur une page web ; un particulier loue son appartement et veut obtenir tous les ans l'indice de référence des loyers pour réévaluer le loyer ; une chercheuse souhaite analyser les migrations résidentielles entre communes et pour cela télécharger le fichier de données du recensement de la population à un niveau fin, etc. L'Insee choisit d'aller au-devant de tous ces publics et à ce titre, doit proposer différents modes d'accès aux données, en commençant par la datavisualisation (De Jonge et Ten Bosch, 2012).

► Des figures pour faciliter l'accès aux données

Pour connaître l'essentiel sur une thématique, l'Insee propose ses chiffres clés, souvent présentés sous forme d'infographies et de tableaux synthétiques : la datavisualisation, c'est-à-dire un ensemble d'indicateurs synthétiques sous forme de visuels simples et faciles à appréhender (Lagarenne et alii, 2023). Elle est privilégiée pour accompagner l'insee-naute dans sa lecture et lui permettre de s'approprier plus facilement les résultats d'une étude. Ainsi, une série chronologique⁶ représentée graphiquement sous forme de courbe selon les périodes disponibles satisfait les besoins de tous les publics sur la plupart des thématiques (indices des prix à la consommation ou de production, chiffres du chômage, emploi salarié, etc.).

Un autre exemple est le **Tableau de Bord de l'Économie Française (TBEF)**, service multi-thématique de datavisualisation sur le site insee.fr. Toutes les informations essentielles des différents domaines du débat public (économie, pouvoir d'achat, démographie, société, salaires, entreprises, développement durable, etc.) sont présentées et cela selon trois volets géographiques (Europe, France, territoires) (**figure 1**). L'institut de statistique du Danemark⁷ propose, quant à lui, une arborescence thématique dans la rubrique « Trouver des statistiques » de son site : une fois le domaine choisi, les données statistiques sont présentées sous forme de figures avec options de téléchargement de ces données et d'analyse plus fine du domaine.

► Télécharger des données pour les réutiliser

Sur le site web de l'Insee, les figures de datavisualisation comprennent systématiquement une option de téléchargement des données. Cela peut servir aux étudiants pour étayer leurs présentations ou aux professeurs d'économie ou de sciences sociales pour préparer des cours. Les journalistes s'intéressent également aux mises à jour de ces indicateurs ou

⁶ Par exemple cette page regroupant les principaux indices et séries chronologiques : <https://www.insee.fr/fr/statistiques/3530678>.

⁷ <https://www.dst.dk/en>.

► **Figure 1 - Vue du Tableau de Bord de l'Économie Française**



- 1 Entrée par thème
- 2 Mise en perspective territoriale, nationale et européenne
- 3 Résumé des indicateurs du thème, avec rubrique « En savoir plus » pour accéder à des données complémentaires
- 4 Pour pouvoir télécharger le tableau des données
- 5 Datavisualisation des indicateurs synthétiques

résultats d'enquête pour préparer un article ; les data journalistes téléchargent notamment les séries chronologiques pour analyser un ensemble de données en support ou complément d'un article de fond.

Au-delà de la datavisualisation, des fichiers avec des volumes de données plus importants sont disponibles, notamment au format XLSX. Ceux-ci portent sur des niveaux de détail plus fins ou regroupent toutes les informations disponibles sur un thème donné et pas seulement un extrait, comme c'est le cas pour une figure de datavisualisation. Cette offre de fichiers à télécharger s'adresse aux inseeuistes qui veulent exploiter directement les données pour leur propre analyse, comme les bureaux d'études, les chercheurs ou certains acteurs locaux. Ainsi, un Conseil régional peut étudier l'activité économique de sa région en utilisant



L'offre de fichiers à télécharger s'adresse aux insee-nautes qui exploitent les données pour leur propre analyse.



les fichiers des créations d'entreprises très finement agrégées selon la localisation géographique, l'activité, la taille et la catégorie juridique des entreprises. Le niveau de détail peut parfois aller jusqu'à des données individuelles comme sur les naissances, les mariages ou les décès issus de l'état civil. Le Conseil régional peut alors faire ses propres agrégations et évaluer les besoins d'installations d'équipement en fonction de la population de sa région.

► Organiser l'offre de fichiers de données

Compte tenu de la variété des thématiques et de la diversité des utilisations, l'Insee doit organiser au mieux son offre de fichiers, en commençant par définir leur contenu. Toute la difficulté est de créer des fichiers de données avec des **axes d'analyse** (aussi appelés variables) pertinents pour des insee-nautes aux profils différents. Par exemple, concernant la thématique des salaires, si un journaliste s'intéresse aux inégalités de genre, il compare les salaires en privilégiant le genre alors qu'une chargée d'études qui suit l'évolution des salaires tout au long de la carrière professionnelle privilégie plutôt l'âge. Il est donc pertinent de proposer un jeu de données sur les salaires moyens croisés selon les axes d'analyse « genre » et « âge » afin de satisfaire ces deux besoins.

La taille des fichiers de données est aussi un élément important de l'offre. Les fichiers ne doivent être ni trop gros (difficilement exploitables par les insee-nautes), ni trop petits (nécessité d'en consulter beaucoup pour analyser un sujet). Par exemple, un fichier de données issues du recensement de la population contenant toutes les informations de l'Insee sur la population française serait beaucoup trop gros et l'insee-naute s'y perdrait facilement. Il doit être fractionné selon des thématiques comme le logement, la famille ou la population étrangère et immigrée. Un découpage peut aussi se faire par le degré d'information : un fichier sur le logement avec les informations principales à connaître, complété par un fichier contenant des informations complémentaires, à destination d'insee-nautes plus spécialistes⁸.

► De la nécessité de normer les fichiers pour les utiliser facilement

Afin d'en faciliter l'exploitation, le format des fichiers de données est généralement normé. Les formats dits plats sont utilisés, au premier rang desquels le format CSV ou plus récemment Parquet (Dondon et Lamarche, 2023) car ils sont facilement lisibles dans un langage de programmation⁹, voire dans un tableur¹⁰ si le fichier n'est pas trop volumineux.

⁸ Un tel découpage reflète la manière dont le recensement de la population est conçu par l'Insee : une exploitation principale et une exploitation complémentaire.

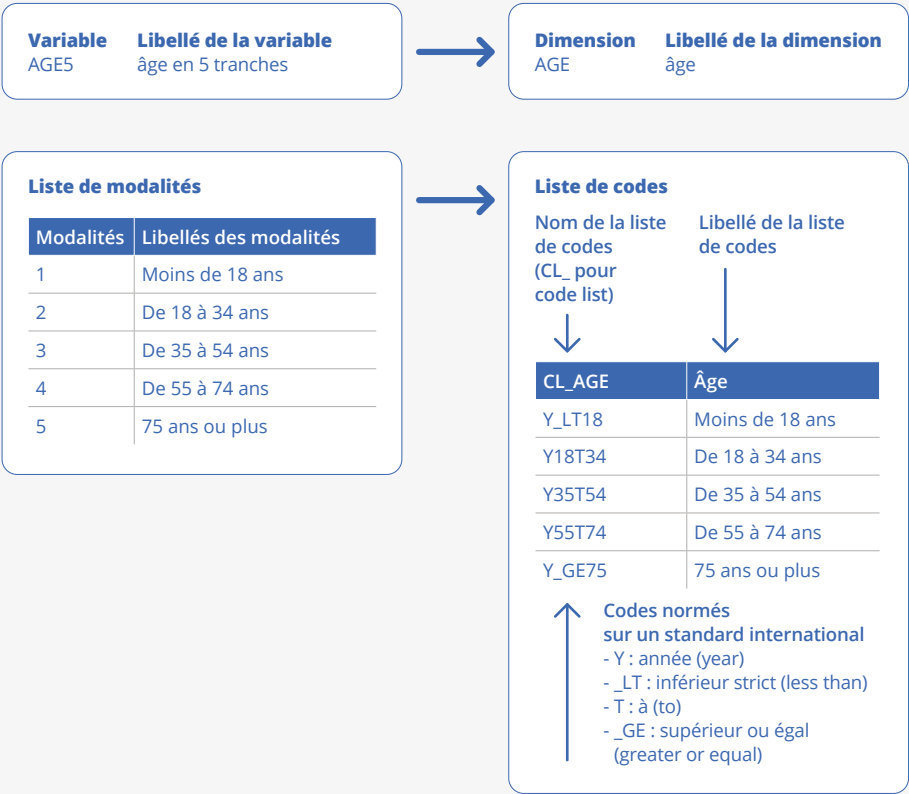
⁹ R ou Python.

¹⁰ Par exemple, Calc de la suite Libre Office.

Le contenu statistique des fichiers est également normé. D’une part, chaque colonne du fichier correspond à une variable déclinée selon ses modalités. Ensuite, les fichiers ne contiennent pas de libellés au niveau des titres de colonnes ou des lignes mais des codes, lesquels sont plus faciles à utiliser quand on veut exploiter le fichier : le titre de colonne est un code relatif à une variable (par exemple le code AGE pour l’âge) et chaque cellule de cette colonne est un code relatif à des modalités de cette variable (par exemple le code « Y35T39 » qui représente la tranche d’âge de 35 à 39 ans). Enfin, les valeurs dans chaque colonne sont dans un même format. Les principaux formats sont la date, la chaîne de caractère, ou le format numérique. Comme le format de chaque colonne est fixe, le contenu de cette dernière peut être exploité plus rapidement par des outils informatiques d’analyse de données.

En accompagnement du fichier, les codes de variables et de leurs modalités sont documentés dans un dictionnaire de codes où ils sont associés à des libellés et regroupés dans des listes de codes. Par exemple, le code de variable AGE a pour libellé Âge et possède une liste de codes formée de codes comme Y35T54 (de libellé « de 35 à 54 ans ») ou Y_GE75 (« 75 ans ou plus ») (*figure 2*). Les variables du fichier peuvent également être attachées à des concepts sémantiques bien définis. Dans l’exemple, la variable de code

► **Figure 2 - Modélisation de la variable âge**



AGE sera attachée à un concept d'âge qui précise s'il s'agit de l'âge en années révolues ou calendaires. De même, des informations générales des tableaux de données comme des précisions dans le titre, l'unité de mesure, le caractère provisoire ou révisé des données sont formalisées et regroupées dans des variables et des listes de codes.

Les variables doivent être comparables d'un jeu de données à un autre lorsqu'elles ont le même sens.

L'ensemble de ces descriptions des données, nommé métadonnées de structure, est essentiel pour comprendre les données (Bonnans, 2019). Lorsqu'un inseeur s'intéresse à un sujet, il souhaite généralement obtenir toutes les informations disponibles sur celui-ci. Il faut donc que les variables soient comparables d'un fichier de données à un autre lorsqu'elles ont le même

sens. D'où l'importance, la nécessité même, d'harmoniser les métadonnées identiques des différents fichiers de données, pour rendre cohérentes les données entre sources. Pour ce faire, une norme de description en conformité avec les standards internationaux est utilisée par l'Insee.

► Des données structurées sous la forme de cubes multidimensionnels

Une notion structurante de la diffusion est celle de « jeu de données » (ou « dataset » en anglais) qui renvoie aux informations contenues dans le fichier de données. Il convient de bien dissocier cette notion de celle de fichier, un même jeu de données pouvant se présenter dans plusieurs fichiers de format différent.

Les jeux de données statistiques sont structurés sous forme d'hypercubes.

Les jeux de données vont être structurés sous forme de « cubes multidimensionnels » ou « hypercubes » dont les dimensions sont les axes d'analyse. On dénombre plusieurs centaines de tels axes dans toute la diffusion de l'Insee ; les plus fréquents sont l'âge, le sexe, la catégorie socioprofessionnelle, le secteur d'activité ou la catégorie d'entreprise (d'un point de vue juridique ou selon la taille). Au croisement des

dimensions de ces cubes, on trouve les valeurs des indicateurs, comme le nombre d'habitants, le nombre d'entreprises ou le revenu.

Ces cubes multidimensionnels sont décrits via le standard international SDMX^{11 12}, et plus particulièrement son modèle d'information. Il est utilisé par l'Office statistique de l'Union européenne Eurostat dans ses échanges de données avec les États membres et par les

¹¹ SDMX signifie *Statistical Data and Metadata eXchange*. L'initiative SDMX, lancée en 2002, établit des normes pour faciliter l'échange de données statistiques et de métadonnées entre les organisations internationales et leurs pays membres, à l'aide des technologies modernes de l'information. Ce format est parrainé par sept organisations internationales : la Banque des règlements internationaux (BRI), la Banque centrale européenne (BCE), l'Office statistique de l'Union européenne (Eurostat), le Fonds monétaire international (FMI), l'Organisation de coopération et de développement économiques (OCDE), la Division de statistiques des Nations Unies (DSNU) et la Banque mondiale. Pour plus de détails, voir (SDMX, 2012).

¹² Un autre standard s'appuyant sur le modèle d'information du SDMX est le standard de web sémantique Datacube.

Nations Unies pour les indicateurs des objectifs de développement durable¹³. Les portails de ces sites explicitent distinctement cette norme descriptive et son format d'utilisation sur des pages dédiées de la rubrique sur les données. Le dictionnaire de codes du cube y est appelé Définition de Structure de Données (ou « *Data Structure Definition* » en anglais, abrégé en DSD). Les variables du cube sont de trois types : les mesures, les dimensions et les attributs. Ces composants sont définis comme suit :

- **les mesures**

Les mesures représentent un phénomène observé via une statistique (population, opérations comptables en statistique d'entreprise ou comptabilité nationale, nuitées dans les hôtels, indices de prix à la consommation ou à la production industrielle, etc.).

- **les dimensions**

Les dimensions correspondent aux axes d'analyse du phénomène observé. Si l'on s'intéresse à une population, il peut être intéressant de décliner cette mesure selon les dimensions telles que le genre, l'âge ou le statut d'emploi. Deux dimensions ont un statut particulier dans la diffusion : la période temporelle (typiquement l'année de référence des données) et le niveau géographique (la région par exemple).

- **les attributs**

Ils apportent des informations qui ne sont pas indispensables à la valeur mesurée mais nécessaires à la compréhension de ce qui est mesuré. Ils peuvent spécifier par exemple les unités de mesure (personnes physiques ou équivalent temps plein), les facteurs d'échelle (unités ou milliers) et le statut de la valeur (définitive ou provisoire).

Avec cette modélisation, un tableau de la population nantaise ventilée selon différents axes correspond à un cube multidimensionnel où la mesure est la population, où les dimensions sont le sexe, l'âge, la catégorie socioprofessionnelle, la commune, l'année et où l'attribut « nombre de personnes » indique que la population est mesurée en unités et non pas en milliers de personnes par exemple (*figure 3*¹⁴). Autre exemple : à partir du tableau « Chiffres-clés » sur l'enquête Cadre de vie et sécurité sur le **nombre de victimes d'agression ou de vol hors ménage selon l'âge et le sexe**¹⁵, la mesure est le nombre de victimes d'agression ou de vol hors ménage. Il est mesuré selon trois dimensions que sont le sexe, l'âge et le type de violences. Les unités de mesure (valeurs en milliers de personnes, taux de plainte en pourcentage) sont informatives et constituent donc des attributs.

Les dimensions et leurs listes de codes sont réutilisables d'un jeu de données à l'autre, ce qui permet de filtrer les jeux de données qui contiennent la dimension recherchée (par exemple l'âge), voire d'aller plus loin en filtrant plus précisément ceux qui contiennent tel code de cette dimension (concrètement une tranche d'âge particulière). C'est une fonction de recherche très utile pour un catalogue.

¹³ <https://unstats.un.org/sdgs/dataportal>.

¹⁴ On peut ici faire une représentation graphique du cube, car il n'a que trois dimensions.

¹⁵ <https://www.insee.fr/fr/statistiques/2525801>.

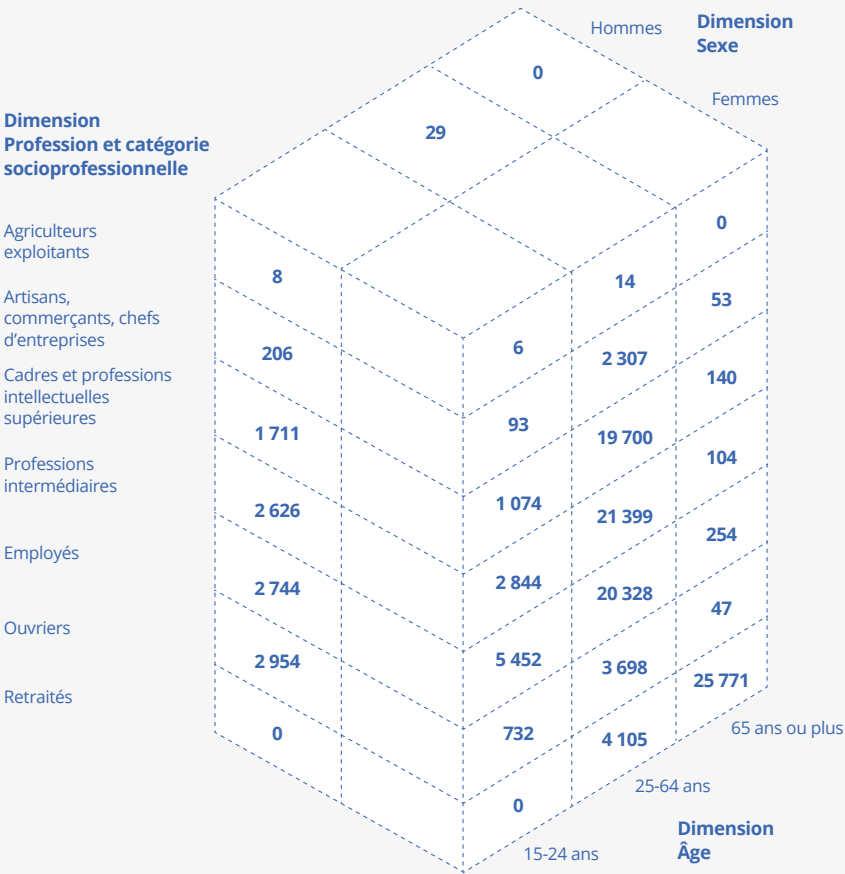
► **Figure 3 - Un cube de données sur la population active de 15 ans ou plus à Nantes en 2020 selon le sexe, l'âge et la catégorie socioprofessionnelle**

Mesure = Population

Dimension géographique = Nantes

Dimension temporelle = 2020

Unité = Nombre de personnes
= Attribut



Source : Insee, recensement de la population 2020.

Lecture : À Nantes en 2020, 732 femmes de 15 à 24 ans sont ouvrières et 29 hommes de 25 à 64 ans sont agriculteurs exploitants. Les valeurs sont affichées pour les croisements de dimensions visibles.

► Un catalogue pour découvrir les jeux de données

Concevoir et structurer les jeux de données n'est pas suffisant. Encore faut-il que l'inseenaute en connaisse l'existence ! Pour cela, ils sont généralement présentés dans un catalogue dédié. Grâce à cet outil, l'inseenaute recherche, selon différents critères, le jeu de données de son choix puis obtient des informations et accède aux données associées. Les critères de recherche sont très importants pour permettre de trouver les fichiers efficacement ; et plus la description des jeux de données est claire, plus le résultat de la recherche sera précis.

Pour bien les décrire, des standards internationaux sont mobilisables comme DCAT¹⁶. Ce standard décrit les métadonnées de catalogage, c'est-à-dire les champs pertinents d'un jeu de données qui sont autant de critères de recherche possibles : par exemple, sa date de création, son thème, son millésime, sa maille géographique (commune, département, région, etc.) ou encore sa source¹⁷. Le standard DCAT aide notamment à assurer la comparaison internationale entre les jeux des différents instituts nationaux de statistique (INS). Au final, un jeu aura donc deux types de métadonnées : ses métadonnées de catalogage et ses métadonnées de structure (*figure 4*).

Une fois les jeux de données décrits, ils peuvent être présentés dans une interface Web de catalogue afin d'y accéder facilement. Cette interface présente l'ensemble des jeux de données et permet à l'internaute de les filtrer selon les critères de recherche. Elle affiche également des informations supplémentaires sur chaque jeu (résumé ou couverture temporelle des données).

En pratique, les catalogues disponibles sur les sites internet de statistiques publiques organisent majoritairement leurs jeux de données par une entrée thématique (démographie, emploi, etc.) puis une arborescence plus fine des thèmes pour obtenir le jeu de données souhaité. Un catalogue est disponible sur le site d'Eurostat¹⁸ pour consulter les différents jeux des données statistiques européennes¹⁹.

C'est également le cas de l'institut de statistique allemand Destatis²⁰ qui met à disposition ses données statistiques via son catalogue Genesis. Comme souvent, ce site sépare le catalogue des autres informations statistiques (tableaux, publications, etc.). Les jeux sont accessibles via la déclinaison de chaque thème. En sélectionner un permet de le visualiser avant de le télécharger. Il en est de même pour le site Agreste²¹ du service statistique du ministère de l'Agriculture qui propose dans la rubrique « Chiffres et analyses », l'accès aux tableaux interactifs par une arborescence thématique.

¹⁶ DCAT signifie Data Catalog Vocabulary. La Commission européenne s'est attachée à décrire un cadre mutualisé pour cataloguer les informations, dans le cas des catalogues de données. Ces derniers peuvent tout autant être ceux de fournisseurs de données (instituts statistiques, administrations publiques, opérateurs) que des portails d'agrégation proposant des regroupements d'information.

¹⁷ <https://www.insee.fr/fr/metadonnees/sources>.

¹⁸ https://ec.europa.eu/eurostat/databrowser/explore/all/all_themes?subtheme=demo&display=list&sort=category.

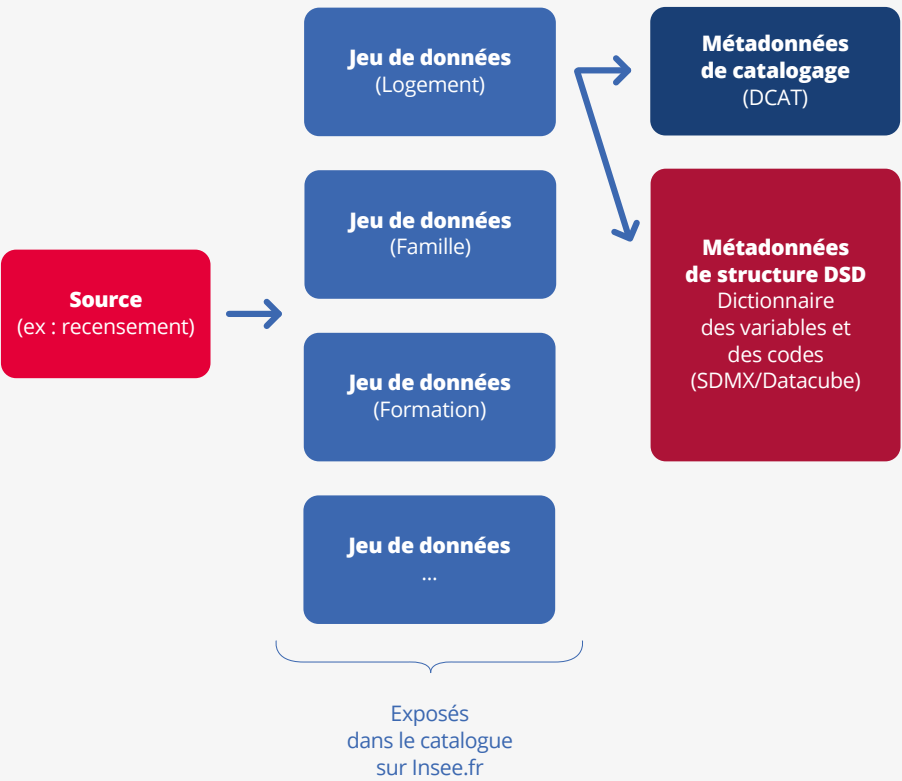
¹⁹ Un autre exemple d'accès est la plateforme ouverte des données publiques françaises : <https://www.data.gouv.fr/fr/pages/thematiques-a-la-une/>. La sélection d'un thème permet d'accéder à la documentation détaillant le sujet et aux jeux de données. L'accès peut se faire également via une arborescence restreinte de thèmes plus précis.

²⁰ https://www.destatis.de/EN/Home/_node.html.

²¹ <https://agreste.agriculture.gouv.fr/agreste-web/disaron/?searchurl/4b54e171-2bf3-4c8b-93b9-06e41472066c:cda8b080-3e9e-4368-b41d-7a29c1da0be6/search/>.

L'Insee dispose d'un tel catalogue²² qui permet un accès plus direct à la donnée via toutes les facettes de recherche (*figure 5*). Conçu dans le cadre d'un projet de modernisation de la diffusion, ce catalogue a vocation à intégrer l'ensemble des données statistiques de l'Insee, tout en répondant aux deux critères d'accessibilité et de clarté du Code de bonnes pratiques de la statistique européenne.

► **Figure 4 - Les métadonnées d'un jeu de données**



22 <https://catalogue-donnees.insee.fr>.

► **Figure 5 - Catalogue des jeux de données de l'Insee en ligne**
(<https://catalogue-donnees.insee.fr>)

Catalogue de données - version beta

L'application Melodi (Mon Espace de Livraison en-Open Data de l'Insee) met à disposition des jeux de données statistiques en open data.

Faites nous part de vos retours

Mes critères

Supprimer tous

Thème

Économie - Conjoncture - Comptes nationaux (10)

Démographie (3)

Revenus - Pouvoir d'achat - Consommation (9)

Conditions de vie - Société (3)

Marché du travail - Salaires (11)

Entreprises (2)

Secteurs d'activité (7)

Territoires, villes et quartiers (10)

Développement durable - Environnement (1)

Rechercher par mot-clé dans le catalogue

38 Résultat(s) de recherche

Afficher par 20 1-20 / 38

Déplacements domicile-travail

Explication principale

5 février 2024

Période temporelle : 2018

Les caractéristiques des navetteurs sont détaillées par communes selon l'âge et le mode de transport pour aller travailler.

Caractéristiques de l'emploi

Explication complémentaire

2 février 2024

Période temporelle : 2018

Les caractéristiques des actifs, des salariés et non-salariés sont fournies pour l'ensemble des communes de France (hors Mayotte).

Famille

Produit Intérieur Brut (PIB) et grands agrégats économiques

melodi

Description

Structure

Données

Titre : Produit Intérieur Brut (PIB) et grands agrégats économiques

Sous-titre : Comptes nationaux annuels - Base 2014

Résumé :

Le produit intérieur brut (PIB) est le principal agrégat mesurant l'activité économique. Il correspond à la somme des valeurs ajoutées brutes nouvellement créées par les unités productrices résidentes une année donnée, évaluées au prix du marché. Il donne une mesure des richesses nouvelles créées chaque année par le système productif et permet des comparaisons internationales. Le produit intérieur brut est publié à prix courants et en volume aux prix de l'année précédente chûnés. Son évolution en volume (c'est-à-dire hors effet de prix) mesure la croissance économique.

Les grands agrégats économiques associés au PIB sont le revenu national brut (RNB), la capacité ou le besoin de financement de la Nation, les grandes composantes de l'équilibre entre les éléments de l'offre (PIB, importations) et de la demande (consommation, investissement, exportations), la ventilation des facteurs de production (emploi, stock de capital) par secteurs institutionnels (entreprises, ménages, administrations publiques considérés comme producteurs de richesses) et la valeur ajoutée brute qu'ils génèrent.

Description :

Identifiant du jeu de données : ID_CNA_AGREGATS_BETA

Source : Comptes nationaux annuels (base 2014)

Thème : Comptes nationaux annuels

Thème (Domaine Statistique de l'ONU) : 2.2 Comptes nationaux

Date de création : 15 décembre 2023

Date de dernière modification : 15 décembre 2023

Fréquence de publication : Annuel

Période temporelle : 1949 - 2022

Fréquence : Annuel

Lecture : Les facettes à gauche permettent de filtrer selon différents critères. Les jeux de données sont alors affichés à droite. En sélectionnant le jeu souhaité, on obtient sa description. On peut télécharger le jeu de données au format CSV, et parfois sous forme de fichiers XLSX.

42

Faciliter l'accès aux données de l'Insee
Cubes, catalogue et métadonnées

► Naviguer dans les cubes pour analyser les données

Une fois le jeu de données choisi dans le catalogue, il est intéressant de l'explorer dynamiquement et de construire ses propres extractions de tableaux. Les INS sont nombreux à proposer ce moyen souple d'exploration. Sur le site de l'Institut italien Istat²³, l'internaute choisit un jeu de données, le visualise directement et accède à la documentation de chaque variable et modalité en cliquant sur les multiples points d'information. L'internaute personnalise ensuite le jeu de données par une sélection des variables et/ou modalités. À l'inverse, le site de l'institut néo-zélandais²⁴ propose tout de suite de construire son tableau avant de le prévisualiser et de l'exporter. Les explorateurs de ces deux sites sont particulièrement complets concernant le choix des caractéristiques des données (unités, présence ou non des lignes ou colonnes vides de valeurs pour les modalités sélectionnées, etc.) et proposent plusieurs formats d'export des données sélectionnées, pouvant inclure des informations sur les données (provisoire, révisée, etc.).

En France, le site Agreste du ministère de l'Agriculture présente ses cubes en ligne et permet, par exemple, de consulter les cubes issus du recensement agricole sur les exploitations²⁵. De même, l'explorateur attaché au catalogue de données sur insee.fr est similaire et permet d'extraire une partie d'un jeu de données en sélectionnant les modalités pertinentes des différents axes. Par exemple, une chargée d'études d'une mairie étudiant les logements locaux pourra filtrer les données de recensement de la population sur sa commune et les communes avoisinantes.

Ces services permettent différents modes d'exploration de cubes multidimensionnels qui peuvent se résumer comme suit²⁶ :

- **le découpage en tranches** : on fige une dimension à une valeur (en anglais « *slice* » pour tranche) en laissant varier les autres dimensions. Dans l'exemple des salaires moyens selon le sexe, l'âge et la catégorie socioprofessionnelle²⁷ (*figure 6a*), on s'intéresse aux données spécifiquement sur les personnes de 50 à 59 ans : on tranche ici selon l'âge en figeant la dimension AGE à la modalité « De 50 à 59 ans ». On obtient alors la ventilation des salaires des personnes de 50 à 59 ans selon leur catégorie socioprofessionnelle (*figure 6b*). Si l'on souhaite regarder les écarts de salaires entre les hommes et les femmes, on constitue une tranche plus fine selon le sexe en fixant la dimension SEXE à la modalité « Femme » pour obtenir un cube sur les salaires moyens des femmes de 50 à 59 ans (*figure 6c*).
- **le découpage en sous-cubes** : on croise cette fois-ci plusieurs dimensions entre elles selon certaines modalités (« *dice* » en anglais), pour obtenir un sous-ensemble de données du cube. Sur ce même exemple, on extrait le salaire moyen des femmes ouvrières de 50 à 59 ans.

²³ <https://www.istat.it/en/analysis-and-products/databases/statbase>.

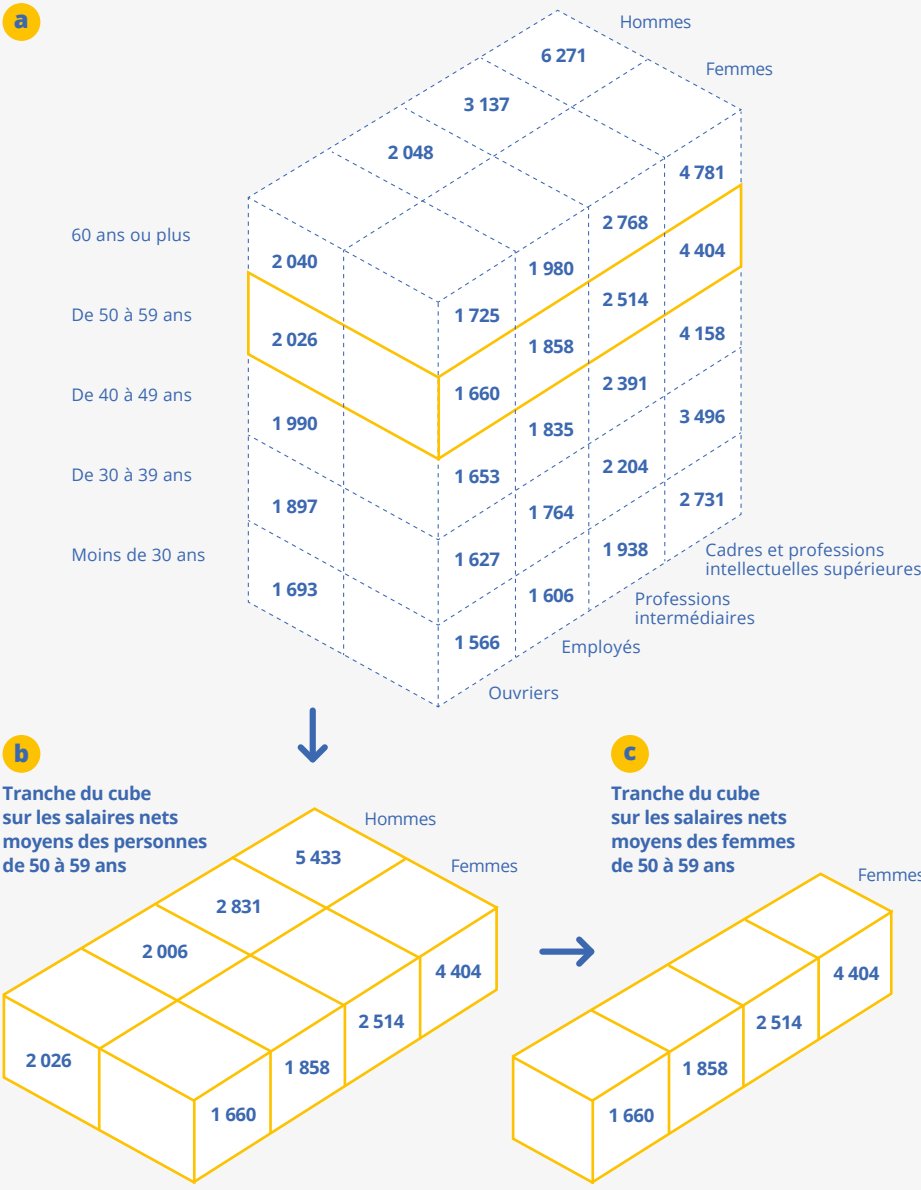
²⁴ <https://infoshare.stats.govt.nz/>.

²⁵ Par exemple, le cube sur les exploitations par taille économique et par orientation : https://agreste.agriculture.gouv.fr/agreste-web/disaron/RA2020_001/detail/.

²⁶ Ceci est expliqué dans la littérature des architectures de données sous le terme de OLAP, acronyme d'« Online Analytical Processing ». C'est une technologie de base de données optimisée pour les requêtes et les rapports, plutôt que pour le traitement des transactions (Codd et alii, 1993).

²⁷ <https://www.insee.fr/fr/outil-interactif/5369554/index.html>.

► **Figure 6 - Cube et tranche de cubes sur les salaires nets moyens en 2021 selon le sexe, l'âge et la catégorie socioprofessionnelle**



Source : Base Tous salariés 2021.
Lecture : À partir du cube sur les salaires nets moyens en 2021, on peut extraire une tranche de cube sur les salaires nets moyens des personnes de 50 à 59 ans ; on peut ensuite cibler une tranche de cube sur les salaires nets moyens des femmes de 50 à 59 ans.

- **le forage vers le haut ou vers le bas** : il est possible de zoomer et dézoomer (en anglais « *drill up* » et « *drill down* ») sur la donnée. Ceci est particulièrement intéressant dans le cas de niveaux d'agrégation emboîtés, notamment pour des nomenclatures, afin d'étudier les données sur des granularités différentes. Ainsi, sur la population de tranche d'âges 50 à 59 ans, on zoome sur cette tranche d'âges pour cibler les populations sur les deux tranches 50 à 54 ans et 55 à 59 ans, voire pour chaque année de 50 à 59 ans. Autre exemple : on dézoome selon des niveaux géographiques allant de la commune au pays.

Ces services d'exploration de données sont destinés à tous les publics, des particuliers pour rechercher des informations à titre personnel aux professionnels traitant de nombreuses données. Le plus souvent, les personnes intéressées ont toutefois un profil de statisticien ou d'économiste, et les professionnels qui exploitent les données de manière automatique et régulière ont besoin d'autres moyens plus techniques.

► Le moissonnage des données par les machines

La consommation de données se fait de plus en plus de machine à machine via des traitements automatisés. C'est le cas, en particulier, des sociétés qui veulent intégrer directement les données de l'Insee dans leur propre système d'information. L'Insee met ses données à disposition via une API, service web pouvant alimenter des applications clientes directement à partir de ses bases de données. Le principe de fonctionnement est le suivant : l'application cliente de l'API est programmée pour interroger régulièrement l'API afin de détecter les mises à jour de données, et le cas échéant, récupérer les dernières informations via une requête (**encadré**). Ce mode de consommation est particulièrement intéressant, car il évite de télécharger manuellement des fichiers sur le site insee.fr et permet grâce au paramétrage de la requête de récupérer uniquement les données d'intérêt (Jacobson et alii, 2011). On parle d'interface machine-machine, car la récupération se fait automatiquement par le programme client, sans aucune intervention manuelle.

De nombreux organismes proposent des APIs (Boyd et alii, 2020). Par exemple, la Cnav (Caisse Nationale d'Assurance Vieillesse) propose une API²⁸ qui permet de lire des données mises à disposition, comme le nombre de retraites au 31 décembre selon le genre, le montant global de la retraite au 31 décembre par type de droit ou le montant mensuel moyen de la retraite.

C'est le cas aussi de l'OCDE²⁹ ou de l'institut canadien StatCan³⁰. De la même manière, l'Insee offre déjà aujourd'hui des APIs pour différents domaines comme la Banque de données macroéconomiques (BDM) pour les séries macroéconomiques ou la Diffusion de Données Locales (DDL) pour les données locales : elles seront remplacées par une unique API appelée Melodi³¹ grâce à une modernisation de la diffusion à l'Insee.

²⁸ <https://data.cnnav.fr/api/explore/v2.1/console>. Cet exemple illustre le « *swagger* », cette page internet présentant l'ensemble des requêtes possibles et le format du résultat des requêtes.

²⁹ <https://data.oecd.org/fr/api/>.

³⁰ <https://www.statcan.gc.ca/fr/debut>.

³¹ Mon Espace de Livraison des données en Open Data de l'Insee.

► Encadré. Comment utiliser une API ?

L'exploration par API consiste à utiliser des adresses internet appelées aussi URL* pour interroger le jeu de données. L'API envoie directement le contenu (dans la page du navigateur internet ou dans l'application cliente) sous un format de fichier standard appelé JSON**.

La structure de l'URL est normalisée comme suit :

Nom de l'API / Méthode / Nom / Filtre de la requête.

Les « méthodes » usuelles sont DATA (pour indiquer qu'on récupère des données) et STRUCTURE (pour avoir le détail des métadonnées). Le nom est ensuite celui du jeu de données (pour la méthode DATA) ou de sa métadonnée de structure (pour la méthode STRUCTURE).

Par exemple, le jeu de données DS_TICM*** sur l'équipement des ménages en technologies de l'information et de la communication propose le taux d'équipement internet à domicile et la part des personnes ayant le haut débit fixe ou mobile à domicile. Cette information est recherchée par une entreprise pour évaluer le marché de production de matériel électronique.

* Sigle de l'anglais « uniform resource locator », localisateur universel de ressources. Adresse qui précise la localisation d'une ressource Internet en indiquant le protocole à adopter, le nom de la machine, le chemin d'accès et le nom du fichier : <https://www.insee.fr/fr/accueil> est une URL.

** JavaScript Object Notation (JSON) est un format de données textuel dérivé de la notation des objets du langage JavaScript.

*** DS pour dataset et TICM pour l'enquête TIC ménages.

**** La future API unique de l'Insee sera <http://api-diffusion-catalogue-donnees-externe.insee.fr>.

Par simplification, supposons que le début de l'URL soit `insee.api****`. L'entreprise collecte l'ensemble des données du jeu de données dans son navigateur Internet à l'URL suivante : `insee.api/DATA/DS_TICM`.

L'entreprise peut aussi extraire une partie du jeu de données en filtrant sur les dimensions de celui-ci. Si elle recherche uniquement les taux d'équipement Internet des femmes en 2022, elle ajoutera le filtre correspondant dans la requête API :

`insee.api/DATA/DS_TICM?MESURE=EQUIP_INT&SEXE='F'&ANNEE=2022`.

MESURE est la dimension de mesure figée au code EQUIP_INT (taux d'équipement) ; SEXE est la dimension du sexe figée au code F (femme) ; ANNEE est la dimension période temporelle figée à 2022.

À noter enfin que seule la récupération des données est possible, tout calcul doit se faire chez le client à partir des données obtenues par l'API.

**Les APIs
démultiplient
le potentiel
de réutilisation
des données
statistiques.**

Cette forme de mise à disposition de données est particulièrement intéressante pour leur diffusion, car elle démultiplie le potentiel de réutilisation des données statistiques. En effet, les outils de datavisualisation s'appuient généralement sur les APIs. Ainsi, l'outil de datavisualisation des salaires sur [insee.fr](https://www.insee.fr)³² permet d'interroger les données de salaires sous différents angles tels que le métier, la catégorie socioprofessionnelle ou encore le sexe. Lorsque l'internaute choisit une profession pour en connaître le salaire moyen, une requête est faite à l'API de l'Insee qui trouve le chiffre

recherché dans la base de données de diffusion de l'Insee et l'envoie à l'outil, qui l'affiche. L'API envoie la valeur disponible la plus fraîche possible puisqu'il a accès directement à la base de données de diffusion. Cet outil sur les salaires reçoit la valeur et l'affiche.

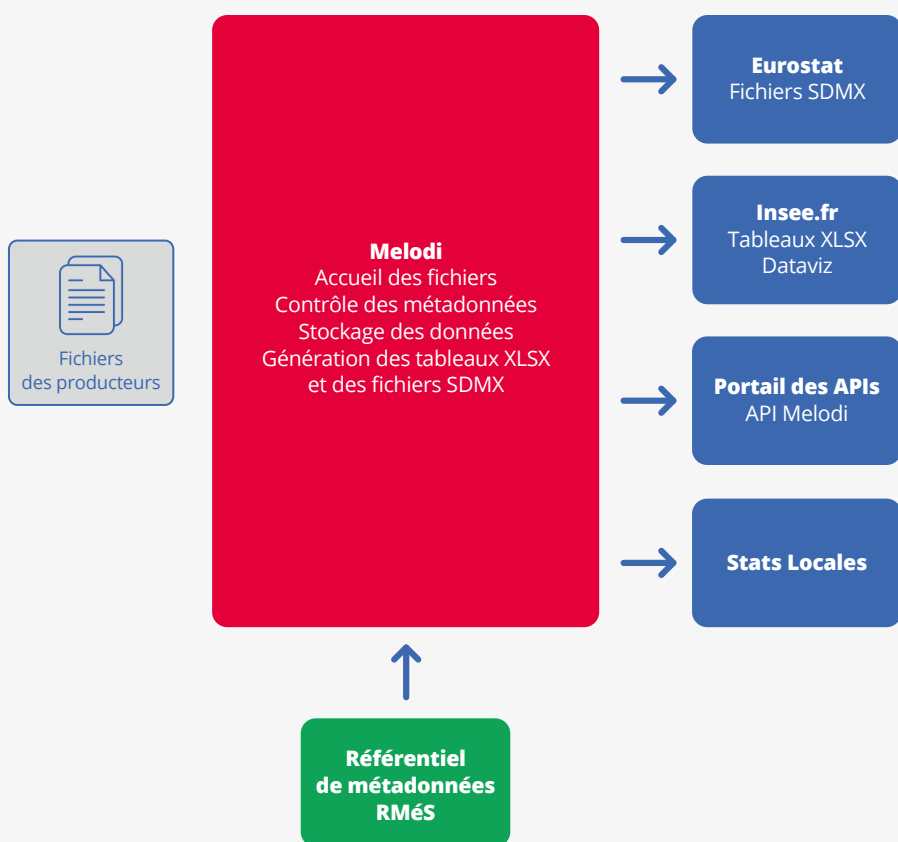
³² <https://www.insee.fr/fr/outil-interactif/5369554/index.html>.

► La nouvelle offre de diffusion de l'Insee

Pour offrir les services décrits précédemment (catalogue, description des cubes, explorateur de données, API), l'Insee s'est engagé dans un projet de modernisation appelé Melodi³³. Ce vaste projet de transformation numérique, qui conduit à une standardisation des données diffusées à l'Insee, repose sur plusieurs principes :

- le premier est de centraliser toutes les données à diffuser dans un même espace appelé entrepôt statistique³⁴ de données et de réaliser tous les produits de données en ligne (fichiers téléchargeables, tableaux web, fichiers envoyés à Eurostat) à partir des données de cet entrepôt, via des outils mutualisés pour l'ensemble de la diffusion (*figure 7*).

► **Figure 7 - Le nouveau processus de diffusion des données à l'Insee**



³³ Mon Espace de Livraison des données en Open Data de l'Insee.

³⁴ Voir (Goossens, 2012) pour une présentation détaillée d'un entrepôt statistique de données.

- le second principe est de décrire ces données selon des métadonnées standardisées (modèle d'information du SDMX/Datacube pour les métadonnées de structure et DCAT pour les métadonnées de catalogage). À ce titre, le processus Melodi s'appuie sur le référentiel de métadonnées statistiques de l'Insee, appelé RMÉS. Cette organisation a de fortes implications pour les équipes de production de données de l'Insee qui construisent les données à diffuser et assurent leur livraison dans l'entrepôt Melodi. Elles doivent fournir des fichiers au format attendu³⁵ et conformes aux métadonnées préalablement décrites dans le référentiel RMÉS.
- un troisième principe, « Dites-le-nous une fois », évite que les équipes de production livrent les mêmes données dans différents canaux de diffusion et réduit fortement le risque d'incohérence de données.

Par ailleurs, la mise en place de Melodi constitue une réelle opportunité de revoir l'offre actuelle de données. Tout d'abord, cela conduit à revoir le contenu statistique de la diffusion : décider si des fichiers très peu téléchargés sont maintenus et à l'inverse développer des thèmes très demandés ou nouveaux. Ensuite, il s'agit de redessiner l'offre autour du catalogue de jeux de données, qui constitue un point d'accès central, et de son explorateur. On peut réduire l'offre de fichiers XLSX en la recentrant sur les indicateurs les plus demandés, et inviter les insee-nautes qui cherchent des données plus spécifiques ou plus détaillées à consulter l'explorateur pour construire leurs propres tableaux ou alors télécharger les fichiers contenant l'ensemble du jeu de données.

► Et demain ?

Ce besoin d'utilisation massive de données statistiques s'avère de plus en plus important et nécessite de mener à bien des évolutions conceptuelles et techniques pour y répondre. On pense notamment à la technologie données ouvertes connectées (*Linked Open Data* ou LOD). Le principe est de structurer les données autour de métadonnées qui sont des ressources universellement utilisées. Par exemple, la région Nouvelle-Aquitaine serait référencée sous la forme d'une « ressource » Internet unique et toute donnée portant sur cette région pointerait vers cette ressource. À la différence d'aujourd'hui où chaque producteur de données est libre de codifier cette région comme il veut, à l'avenir il devrait faire référence à cette codification universelle. Ce recours à des métadonnées universelles permettrait d'assurer une comparabilité entre jeux de données.

Au-delà de la sphère de la statistique publique, l'intelligence artificielle (IA) ouvre la voie à de nouveaux services d'interrogation de la donnée pour la rendre encore plus accessible. La description des métadonnées associées aux données facilite grandement leur compréhension par des algorithmes d'intelligence artificielle. C'est particulièrement utile pour des outils de type chatbot/statbot où l'internaute pose une question – par exemple quel est le dernier taux de chômage ? – question ensuite interprétée par un algorithme d'IA pour interroger la base de données et envoyer la réponse ; la qualité de la description des données sera alors un facteur déterminant dans la capacité de l'IA à répondre de manière pertinente.

³⁵ Beaucoup de livraisons des producteurs se font en SAS ou XLSX dans les processus actuels. Melodi impose des formats dits plats comme le CSV ou le Parquet, adaptés aux fichiers très volumineux.

► 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 6 février 2024]. Disponible à l'adresse : <https://www.insee.fr/fr/information/4168396?sommaire=4168411>.
- BOYD, Mark, GATTWINKEL, Dietmar, POSADA, Monica et VACCARI, Lorenzino, 2020. An Application Programming Interface (API) framework for digital government. In : *Publications Office of the European Union, Luxembourg*. ISBN : 978-92-76-18980-0. [en ligne]. [Consulté le 6 février 2024]. Disponible à l'adresse : https://publications.jrc.ec.europa.eu/repository/bitstream/JRC120715/final_version_pdf_version.pdf.
- CODD, Edgar Franck, CODD, Sharon B. et SALLEY, Clynch T., 1993. Providing OLAP to User-Analysts: An IT Mandate. In : *E. F. Codd & Associates*. [en ligne]. [Consulté le 6 février 2024]. Disponible à l'adresse : http://www.estgv.ipv.pt/paginaspeessoais/jloureiro/esj_aid2007_2008/fichas/codd.pdf.
- DE JONGE, Edwin et TEN BOSCH, Olav, 2012. Visualising official statistics. In : *Site de Statistics Netherlands*. [en ligne]. [Consulté le 6 février 2024]. Disponible à l'adresse : https://www.researchgate.net/publication/267856653_Visualising_official_statistics.
- DONDON, Alexis et LAMARCHE, Pierre, 2023. Quels formats pour quelles données ? In : *Courrier des statistiques*. [en ligne]. 30 juin 2023. Insee. N° N9, pp. 86-103. [Consulté le 6 février 2024]. Disponible à l'adresse : <https://www.insee.fr/fr/information/7635827?sommaire=7635842>.
- EMILSSON, Cecilia, RIVERA PÉREZ, Jacob A. et UBALDI, Barbara-Chiara, 2020. OECD Open, Useful and Re-usable data (OURdata) Index: 2019. In : *Site de l'OCDE*. [en ligne]. [Consulté le 6 février 2024]. Disponible à l'adresse : <https://web.archive.oecd.org/2020-03-10/547558-ourdata-index-policy-paper-2020.pdf>.
- EUROPEAN COMMISSION, 2015. Creating Value through Open Data. In : *Portail officiel des données européennes*. [en ligne]. Novembre 2015. [Consulté le 6 février 2024]. Disponible à l'adresse : https://data.europa.eu/sites/default/files/edp_creating_value_through_open_data_0.pdf.
- GOOSSENS, Harry, 2012. The statistical data warehouse: a central data hub, integrating new data sources and statistical output – Contributed Paper at the UNECE Conference of European Statisticians. In : *Site de l'UNECE*. [en ligne]. 8 octobre 2012. [Consulté le 6 février 2024]. Disponible à l'adresse : <https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/mtg2/WP18.pdf>.
- JACOBSON, Daniel, BRAIL, Greg et WOODS, Dan, 2011. APIs: A Strategy Guide. In : O'Reilly Media, Inc. ISBN : 9781449308926.

- LAGARENNE, Christine, MINODIER, Frédéric et SAMSON, Odile, 2023. Comment présenter nos données pour mieux communiquer ? – La datavisualisation : synthèse et simplicité. In : *Courrier des statistiques*. [en ligne]. 11 décembre 2023. Insee. N° N10, pp. 7-29. [Consulté le 6 février 2024]. Disponible à l'adresse : <https://www.insee.fr/fr/information/7722079?sommaire=7722116>.
- SDMX, 2012. SDMX 2.1 User Guide. In : *Site de SDMX*. [en ligne]. 19 septembre 2012. [Consulté le 6 février 2024]. Disponible à l'adresse : https://sdmx.org/wp-content/uploads/SDMX_2-1_User_Guide_draft_0-1.pdf.
- UBALDI, Barbara, 2013. Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives. In : *OECD Working Papers on Public Governance*, N° 22, OECD Publishing, Paris. [en ligne]. Mai 2013. [Consulté le 6 février 2024]. Disponible à l'adresse : <https://sharingcitiesalliance.knowledgeowl.com/help/open-government-data-towards-empirical-analysis-of-open-government-data-initiatives>.