Décembre 2023

Courrier des statistiques

10

Insee
Mesurer pour comprendre

STATISTIQUE PUBLIQUE

Rédaction en chef
Catherine Fresson-Martinez

Contribution
Insee : Pascal Ardilly, Christine Lagarenne, Frédéric Minodier, Odile Samson
ANS : Johanna Bensoussan, Joël Bizingre, Nathalie Courvalin
DEPP : Séverine Bidet-Caulet et Christian Burel
Injep : Augustin Vicard
OED : Pierre Greffet

Directeur de la publication
Jean-Luc Tavernier

Directeur de la collection
Pascal Rivière

Rédaction
Catherine Fresson-Martinez, Pierre Glénat, Solenn Ily, Marine Le Roux, Pascal Rivière

Composition
Agence Efil
90 boulevard Heurteloup
37 000 Tours
02 47 47 03 20
www.efil.fr

Photo de couverture
Getty Images

Éditeur
Institut national de la statistique et des études économiques
88, avenue Verdier
92541 MONTROUGE CEDEX

www.insee.fr

Courrier des statistiques N10

SOMMAIRE

PRÉSENTATION DU NUMÉRO

Cinq ans déjà de publications pour le Courrier des statistiques nouvelle mouture, plus de 70 articles, et toujours une exigence de pédagogie, la volonté de s'ouvrir à de nouveaux sujets, de nouveaux services… avec désormais des numéros à deux chiffres, et une maquette modernisée.

Ce numéro N10 commence par une problématique jamais abordée dans le nouveau Courrier (mais à plusieurs reprises dans l'ancien), celle de la diffusion des statistiques. Il s'agit plus précisément de la visualisation des données, ou datavisualisation (dataviz pour les intimes). L'article de Christine Lagarenne, Frédéric Minodier et Odile Samson est riche et multiplie les exemples, y compris anciens (l'extraordinaire carte de Charles Joseph Minard sur la campagne napoléonienne de Russie) pour nous faire toucher du doigt toutes les facettes de cette activité. Cependant, il ne faudrait pas hâtivement la résumer à la réalisation de « jolis » graphiques. À la frontière entre diffusion et communication, la dataviz cherche à simplifier les messages pour une compréhension au premier coup d'œil mais doit aussi donner envie aux lecteurs de lire. Pour cela, elle fait appel à différentes techniques : infographie, défilement narratif, datavisualisation interactive… L'article soulève également des questions opérationnelles, techniques et d'organisation, sur un sujet désormais incontournable pour les statisticiennes et statisticiens.

Une préoccupation permanente du Courrier est de donner la parole à tous les services statistiques ministériels, afin de donner à voir la manière dont se décline la statistique dans différents secteurs d'activité. Jusqu'à présent, le Courrier a proposé des articles provenant des services statistiques ministériels (SSM) de la justice (N1), de l'intérieur (N2, voire N7), de la santé (N4), du SDES pour la partie logement (N4), de l'éducation (N5, N6, N7), de l'agriculture (N7), et des collectivités locales (N8). Dans le présent numéro, deux nouveaux domaines sont abordés, la défense et le sport.

Dans l'article sur les statistiques de la défense, Pierre Greffet aborde les spécificités du domaine, non soumis à des règlements européens, ne se fondant pas facilement dans les nomenclatures générales (NAF), et entretenant des liens réguliers et très formalisés avec le monde de la recherche. Il met aussi en évidence un paradoxe : de telles données sont souvent très confidentielles, avec une sensibilité qui dépasse le secret statistique… et dans le même temps il existe une volonté d'ouverture, notamment aux chercheurs. On peut résoudre cette contradiction apparente en soulignant que si certains sujets peuvent être sensibles (économie de défense), d'autres ne le sont pas (fréquentation des lieux de mémoire) ; ainsi, un projet pour organiser une ouverture très maîtrisée est en cours de réflexion avec le principe de *data room*.

Avec les statistiques sur le sport, on retrouve un sujet non encadré par une réglementation internationale… et peu encadré de façon générale. Augustin Vicard présente les sources disponibles dans ce domaine et met en évidence les limites des données administratives, les enjeux de caractérisation de la notion de pratique sportive ainsi que les difficultés

liées à la multiplicité des sports, certains plus rares. Enquêter sur le sujet requiert donc des choix forts, par exemple en se concentrant sur la pratique sportive régulière. Au-delà du processus de production, la question même des enjeux de ces statistiques se pose : envisage-t-on le sport comme fait social ? Comme activité physique, avec en particulier une approche sanitaire, et des politiques publiques qui vont avec ? L'auteur aborde enfin la question des données issues d'applications connectées, de capteurs, intéressantes mais qui ne permettent pas un suivi harmonisé de la pratique.

Il s'ensuit une série de deux articles à la structure très proche, portant cette fois sur les domaines de l'éducation et de la santé, mais qui ne sont pas *stricto sensu* des articles de SSM : en effet, ils ne portent pas sur les statistiques mais sur les répertoires. De ce point de vue, ils font écho au dossier sur les répertoires du numéro N8, qui contenait des articles sur le RNIPP, le SNGI, SIRUS, et la Base permanente des équipements. Dans le présent numéro, deux répertoires d'établissements sont décrits : le répertoire FINESS des établissements sanitaires et sociaux et le répertoire Ramsese des établissements du système éducatif. On ne saurait trop recommander aux lectrices et lecteurs de lire les deux articles « en miroir » pour en percevoir les points communs et les différences.

FINESS est un répertoire connu dont la création date de plus de 40 ans (1979) et qui joue un rôle fondamental dans la régulation, l'évaluation, le pilotage, le financement et l'identification des structures qui le constituent. Il se caractérise par un cadre formel très exigeant, tout enregistrement de données requérant l'existence d'actes juridiques ou administratifs. Chaque établissement possède un numéro d'identification. FINESS joue un rôle majeur dans l'écosystème des systèmes d'information de santé. Articulé avec deux autres référentiels du domaine, le RPPS et le ROR[1], il est aussi apparié avec Sirene. Il fait l'objet de nombreuses utilisations par les administrations centrales, le grand public et les établissements eux-mêmes. Son rôle central, sa diffusion large, induisent de fortes exigences de qualité. Pour pallier certaines limites, il fait l'objet d'une refonte pilotée par l'Agence du numérique en santé.

Créé en 1977, à la même époque que FINESS, le répertoire académique et ministériel sur les établissements du système éducatif Ramsese s'appuie sur une démarche très différente, profitant de l'organisation territoriale du système éducatif. Ainsi, la gestion déconcentrée du répertoire est-elle assurée par les services statistiques académiques (SSA), le travail des gestionnaires au niveau local étant déterminant pour garantir la qualité des données du répertoire. Ses usages sont très variés et permettent de répondre à des besoins statistiques, de pilotage, de gestion ou d'interopérabilité, chaque établissement disposant d'un numéro d'identification unique. Ramsese joue un rôle central dans le système éducatif et veille à la cohérence des données des structures publiques avec Sirene. Sa diffusion fondée sur des API facilite le partage des données dans le cadre des projets d'urbanisation des applicatifs. La visibilité de Ramsese s'étend, avec la mise à disposition en *Open data* d'une partie de ses données.

1 Répertoire partagé des professionnels de santé et répertoire opérationnel des ressources.

Ce numéro s'achève avec un thème *a priori* plus classique dans l'univers statistique, celui des sondages. Enfin, classique… jusqu'à un certain point, car Pascal Ardilly y aborde un sujet peu évoqué en statistique publique : les sondages empiriques. N'étant pas fondés sur une sélection aléatoire *a priori* d'un échantillon, ils constituent, en revanche, la norme dans les instituts de sondage. De façon pédagogique, l'auteur décline les différences entre sondages aléatoire et empirique, avec force schémas. Les sondages empiriques, efficaces pour maîtriser les coûts, le sont moins pour maîtriser les erreurs, avec en particulier un problème spécifique de biais. Mais ils possèdent eux aussi une véritable justification théorique. Pour finir, à travers deux exemples, l'article éclaire d'un jour nouveau l'utilisation du *big data*, en nous délivrant un message : attention, la quantité n'est pas gage de qualité…

Pascal Rivière
Directeur de la collection, Insee

Comment présenter nos données pour mieux communiquer ?

La datavisualisation : synthèse et simplicité

Christine Lagarenne[*], *Frédéric Minodier*[**] *et Odile Samson*[***]

L'image est un formidable vecteur de transmission : les messages passent quasi instantanément de l'œil au cerveau pour devenir information. L'image a été mise au service de la statistique, en particulier au XIXe siècle où cartes et tableaux ont été popularisés en France. Pour en augmenter l'impact, elle a été adossée à un récit qui en explicite les messages ; les technologies web ont complété l'éventail en facilitant l'accès à un très grand nombre de données et en les appréhendant à travers des supports dynamiques. La sémiologie graphique et le design sont complémentaires : la sémiologie améliore l'efficacité des illustrations pour la bonne compréhension et le design réduit l'effort de lecture. Dans la réalisation de la datavisualisation ou dataviz, les règles fondamentales du métier de statisticien doivent être respectées (métadonnées, rigueur et présentation du chiffre). Des solutions simples et génériques favorisent l'expérience utilisateur et permettent au statisticien de conserver la maîtrise technique dans un environnement hautement évolutif. Avec la dataviz comme vecteur de communication, la statistique publique continue à éclairer le débat public et à l'alimenter, avec toujours plus de données accessibles au plus grand nombre.

The image is a powerful vector of transmission: messages are passing quite instantly from the eye to the brain where they become information. That's why image has been used by statisticians, starting from 19th century when maps and charts were made popular (Rendgen, 2020). Storytelling was added to spell out the intended message, before web technologies completed the scope by enabling the general public to easily access a very large amount of data and by helping to comprehend them through dynamic media. Graphic semiology and design are complementary: semiology has improved the effectiveness of illustrations in terms of comprehension; design helps to reduce the reading effort. When it comes to producing a datavisualisation or dataviz, the statistician's basic rules must be kept in mind (metadata, rigour and presentation of figures). When implementing, simple and generic solutions should drive choices to enhance user experience and enable statisticians to keep the control for technology, especially in the highly evolving environment. Using dataviz, a mean of communication, official statistics continue to enlighten public debate and to keep it well informed with large amounts of highly accessible data.

* Cheffe du département de l'offre éditoriale, DDAR, Insee.
christine.lagarenne@insee.fr

** À la date de la rédaction, chef de la division édition et diffusion internet, DDAR, Insee.
frederic.minodier@insee.fr

*** Cheffe de la section graphique, DDAR, Insee.
odile.samson@insee.fr

La datavisualisation ou dataviz regroupe l'ensemble des représentations visuelles des données, depuis les diagrammes en bâtons utilisés de longue date jusqu'aux infographies et tableaux de bord dynamiques. Quel que soit le support, elle est vecteur d'information ou de communication, au format papier ou de plus en plus numérique. Elle permet de rendre compte de façon synthétique d'un ensemble de données mais aussi d'accompagner l'utilisateur dans sa lecture afin qu'il s'approprie plus facilement les résultats d'une étude.

> ***La datavisualisation ou dataviz regroupe l'ensemble des représentations visuelles des données.***

Dans le contexte de développement massif des données, la datavisualisation est une opportunité pour la statistique publique de les exploiter encore davantage. Elle permet de diffuser, d'une part, les données de manière plus élaborée et, d'autre part, les résultats des études statistiques et économiques de façon plus accessible. Elle participe à la stratégie de lutte contre les infox (*fake news*) en touchant un public large grâce à sa facilité de lecture.

L'évolution des différentes formes de dataviz est illustrée par quelques exemples de la littérature sur le sujet et des pratiques de l'Insee et de services statistiques ministériels. Les pré-requis pour la développer et rester à niveau tout en respectant les critères de diffusion de la statistique publique sont ensuite détaillés.

▶ L'image, un média très utile…

> ***« Une image vaut mille mots », a dit le philosophe Confucius.***

« Une image vaut mille mots », a dit le philosophe Confucius. Les données sont mieux perçues quand elles sont représentées sous un format visuel et non plus textuel. Non seulement le cerveau traite les images à très grande vitesse[1] mais en plus, le taux de reconnaissance d'une image après 3 jours est de 90 % (*Haber, 2013*). Il faut dire qu'en moyenne, on ne lit qu'une dizaine de caractères par fixation oculaire (*Brysbaert, 2019*). Les lecteurs ne lisent pas vraiment dans un premier temps. Ils butinent, ils scannent, ils ont une lecture sélective et associative.

Le cerveau cherche à mettre en forme, à donner une structure signifiante à ce qu'il perçoit, afin de le simplifier et de l'organiser devant la complexité de notre environnement. La théorie de la *Gestalt* (i.e. « forme » en allemand) inspirée notamment par Christian von Ehrenfels (*Dortier, 2012*) au début du XX^{e} siècle traite de la psychologie de la forme. Il existe six lois, dont trois vont être utiles à la dataviz. Tout d'abord, la loi de la bonne forme, loi principale dont les autres découlent : notre cerveau cherche à reconnaître en toute chose des formes simples et stables qui lui sont familières et l'utilisation directe de ces formes dans les représentations améliore notre rapidité de lecture et de compréhension. Ensuite, la loi de proximité regroupe les points les plus proches les uns des autres et est utilisée dans les histogrammes par exemple.

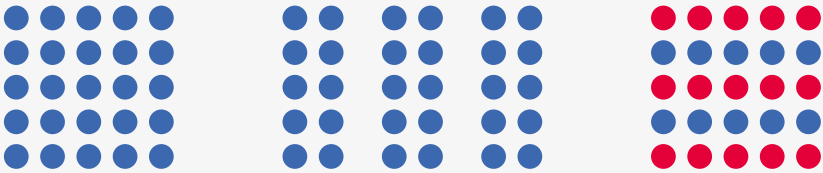
1 *https://news.mit.edu/2014/in-the-blink-of-an-eye-0116.*

> ***Notre cerveau cherche à reconnaître en toute chose des formes simples et stables qui lui sont familières.***

Enfin, la loi de similarité attribue un caractère commun (couleur, taille, etc.) à des éléments permettant de les regrouper visuellement même s'ils sont épars (***figure 1***).

Pour que le plus grand nombre de personnes puisse s'approprier les données, notamment statistiques, la forme graphique est un moyen essentiel, largement utilisé. Il permet de synthétiser mais aussi et surtout de distinguer les éléments les plus importants. La simplicité de la représentation (courbe, diagramme en bâtons) est primordiale. Des représentations plus techniques ont également fait leurs preuves pour accéder à l'information, en permettre une compréhension rapide et apporter un soutien à l'écrit[2]. Elles font désormais partie de notre vie de tous les jours (***encadré***), comme les cartes géographiques.

▶ Figure 1 - Trois lois utiles à la dataviz



Bonne forme (prégnance)
Notre cerveau cherche à reconnaître des formes simples et stables qui lui sont familières.

Proximité
Les éléments proches sont spontanément associés pour créer un groupe.

Similarité
Les formes similaires (forme, couleur, orientation, etc.) sont regroupées spontanément entre elles.

Combien y a-t-il de chiffres 3 ?

Grace à la loi de similarité, juste en coloriant une donnée celle-ci est immédiatement lisible.

2 Voir notamment le rapport « La culture statistique des Français : constats, enjeux et perspectives », INSEE N° 2023_14/DG75-B001, IGÉSR N° 21-22 316A – février 2023, IGAC N° 2023-05. *https://intranet.insee.fr/jcms/18876156_DBFileDocument/mi-2022-6-litteratie-statistique-rapport-ig-21-04-2023?details=true.*

▶ Encadré 1. La représentation cartographique, au service des voyageurs

La représentation cartographique s'est appuyée sur une connaissance de plus en plus fine de notre planète au fil des siècles. Les photos aériennes, les satellites dont le fameux *Global Positionning System* (GPS) ont permis de préciser les mappemondes de la Renaissance et les relevés des différentes expéditions scientifiques du siècle des Lumières. Pour autant, la précision de la carte n'est pas toujours nécessaire. C'est ainsi qu'est né le plan du métro londonien dont l'archétype date de 1931. Ce type de plan est maintenant la norme.

Plus proche de nos métiers, les horaires des transports en commun adoptent la forme d'un diagramme arborescent (*steam-and-leaf*). Pour concentrer l'information, l'axe des ordonnées représente les heures et celui des abscisses les minutes de l'horaire de départ du train.

Dans les deux cas, on prend de la distance par rapport à l'objet que l'on veut étudier : on est loin de l'exactitude géographique ou de la représentation linéaire du temps. « La carte n'est pas le territoire » (*Korzybski, 1933*).

▶ … qui illustre la statistique depuis le XIXe siècle

Le développement de la forme graphique s'est fait progressivement. À titre d'exemple, les évolutions temporelles sont représentées sous forme de courbe depuis le Xe siècle (*Andry et alii, 2022*), où elles servaient à décrire le mouvement des planètes. L'utilisation pour la statistique est plus récente.

> ***Cette démarche cherche à valoriser les données chiffrées. L'usage de ces représentations graphiques simples est intentionnel pour permettre une compréhension au premier coup d'œil.***

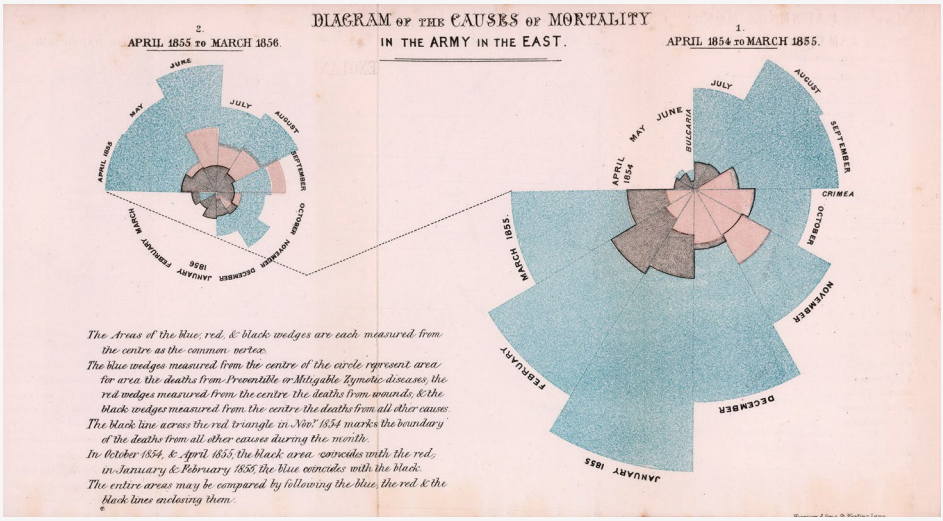
Ainsi, à la fin du XVIIIe siècle, William Playfair a pour la première fois utilisé les trois modes fondamentaux de représentation statistique, à savoir la courbe pour des séries temporelles, le diagramme en bâtons pour des effectifs ou des proportions, et le camembert pour des proportions. Cette démarche cherche à valoriser les données chiffrées. L'usage de ces représentations graphiques simples est intentionnel pour permettre une compréhension au premier coup d'œil. Au XIXe siècle, l'utilisation des premiers graphiques permet d'identifier des corrélations, et de nouvelles représentations plus complexes apparaissent. Ainsi, en 1854, John Snow[3] met en regard les décès et les points d'eau sur une carte, ce qui lui permet de trouver l'origine de l'épidémie de choléra à Londres : une pompe infectée.

Infirmière lors de la guerre de Crimée durant l'hiver 1854-1855, Florence Nightingale parvient à attirer l'attention du politique sur le fait qu'on ne meurt pas uniquement sur le champ de bataille, mais surtout des suites de ses blessures en raison de conditions sanitaires déplorables, grâce à un diagramme en crête de coq (***figure 2***). Il s'agit d'un diagramme polaire ; chaque secteur, correspondant à un mois donné est découpé en trois aires concentriques selon trois causes de décès (mort au combat, maladie infectieuse, autre cause).

Qu'il s'agisse de proportions ou d'évolutions temporelles, le statisticien dispose de plusieurs types de représentations graphiques adaptées aux données pour faire parler les chiffres et mettre en avant des résultats, pour établir les faits et convaincre. D'où l'importance du choix de la visualisation pour identifier et porter les messages.

3 "On the Mode of Communication of Chol era" by John Snow, originally published in 1854 by C.F. Cheffins, Lith, Southhampton Buildings, London, England. *https://archive.org/details/b28985266/page/n57/mode/2up*.

Figure 2 - Diagramme des causes de mortalité dans l'armée



Source : Harrison and sons, Saint Martin's Lane

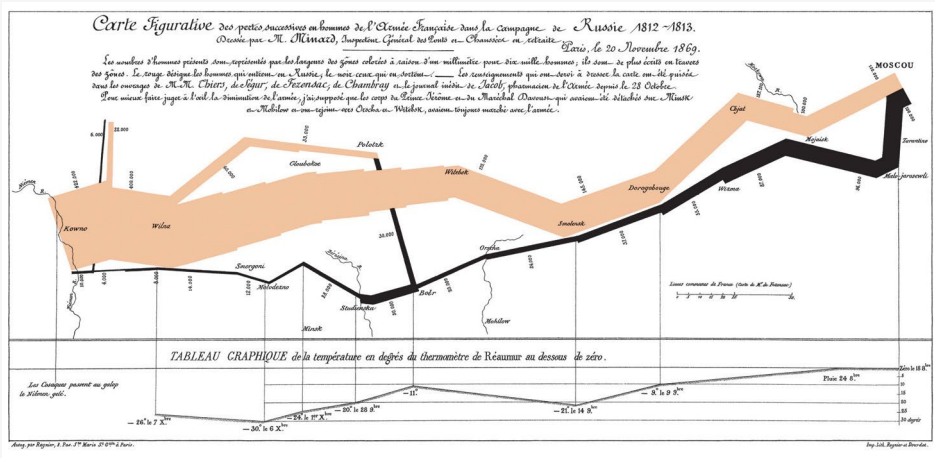
Raconter des histoires : le *storytelling*

Les objectifs d'une dataviz sont non seulement d'accompagner et de faire comprendre les données mais aussi de communiquer, d'illustrer des messages et de donner envie aux lecteurs de lire (accrocher le lecteur). Une datavisualisation est à la frontière entre diffusion et communication.

Pour intéresser le lecteur, raconter une histoire à l'aide de visuels permet de capter son attention : en anglais, on parle de *storytelling*. Il peut prendre des formes variées : la carte de la campagne de Russie réalisée par Charles Joseph Minard (1869), une des premières réalisations de datavisualisation multidimensionnelle, s'apparente au *storytelling* d'avant le numérique. L'auteur réussit à raconter sous forme graphique l'évolution de l'armée française entre juin et décembre 1812 (***figure 3***). Dans un même espace, il décrit simultanément son itinéraire, en indiquant topographiquement les lieux (cours d'eau, villes) et sa direction (en brun pour la campagne jusqu'à Moscou, en noir pour la retraite) ; la dimension statistique est ajoutée par la largeur de la coloration qui représente la taille de l'armée à chaque moment, ainsi que l'évolution de la température de l'air pendant la retraite. Minard a développé une cartographie originale dans le but de « faire apprécier immédiatement par l'œil, autant que possible, les proportions des résultats numériques » (*Palsky, 1996*).

> ***L'auteur réussit à raconter sous forme graphique l'évolution de l'armée française entre juin et décembre 1812.***

▶ Figure 3 - Carte figurative de la campagne de Russie 1812-1813



Source : Charles Joseph Minard, 1869.
Reproduction provenant de l'École nationale des Ponts et Chaussées, Paris.

Plus simplement, faire ressortir une information sur un graphique complexe constitue déjà en soi du *data storytelling*. Par exemple, distinguer une courbe parmi plusieurs, isoler une part dans un camembert, fournit aux lecteurs une clé de lecture implicite, un message.

▶ Vers l'infographie et le défilement narratif (*scrollytelling*), deux formes de *storytelling*

Une infographie regroupe un graphique, un diagramme ou toute autre image visuelle, comme un pictogramme ; elle est destinée à présenter des informations complexes sous une forme facilement compréhensible[4], adaptée à la diffusion sur support numérique ou papier. À titre d'exemple, une infographie sur l'étude *Femmes et Hommes : une lente décrue des inégalités* met en lumière les orientations post-bac des femmes. Les comparaisons femmes-hommes sont illustrées par deux podiums (6 chiffres extraits de deux figures) (***figure 4***).

À l'ère du numérique, une autre forme de scénarisation se développe : le *scrollytelling* qui est une animation de dataviz dans une page *web*. Ce terme provient de la contraction entre *storytelling* et *scroll*, le défilement haut-bas sur un écran. Elle incorpore tous les éléments multimédias (sons, textes, vidéos, infographies, animations, photos ou dessins) de façon fluide grâce à la parallaxe[5]. Elle est plus ou moins interactive et peut orienter l'utilisateur vers différents scénarios.

4 Selon le Dictionnaire Collins en langue anglaise : *infographics, a graph, diagram, or other visual image designed to present complex information in an easily understandable form*. Les dictionnaires Robert et Larousse donnent une définition plus générique de l'infographie (image *via* un ordinateur).

5 La parallaxe est un effet visuel lié à une vitesse de défilement différente entre le premier et l'arrière-plan lors du *scroll* (défilement).

▶ Figure 4 - Part des femmes dans les différentes formations d'enseignement supérieur en 2020-2021

Part des femmes dans les différentes formations d'enseignement supérieur en 2020-2021

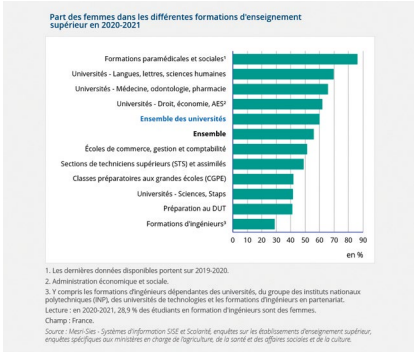
en %

Formations paramédicales et sociales[1]	86,0
Universités - Langues, lettres, sciences humaines	69,7
Universités - Médecine, odontologie, pharmacie	65,6
Universités - Droit, économie, AES[2]	61,7
Ensemble des universités	59,7
Ensemble	**55,9**
Écoles de commerce, gestion et comptabilité	51,1
Sections de techniciens supérieurs (STS) et assimilés	48,9
Classes préparatoires aux grandes écoles (CGPE)	41,9
Universités - Sciences, Staps	41,4
Préparation au DUT	40,9
Formations d'ingénieurs[3]	28,9

1. Les dernières données disponibles portent sur 2019-2020.
2. Administration économique et sociale.
3. Y compris les formations d'ingénieurs dépendantes des universités, du groupe des instituts nationaux polytechniques (INP), des universités de technologies et les formations d'ingénieurs en partenariat.
Lecture : en 2020-2021, 28,9 % des étudiants en formation d'ingénieurs sont des femmes.
Champ : France.
Source : Mesri-Sies - Systèmes d'information SISE et Scolarité, enquêtes sur les établissements d'enseignement supérieur, enquêtes spécifiques aux ministères en charge de l'agriculture, de la santé et des affaires sociales et de la culture.

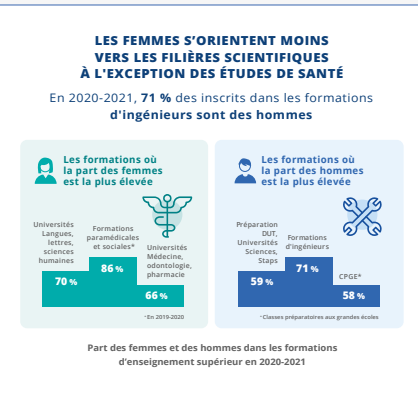
Tableau de données

- Les données sont difficiles à appréhender ;
- Pas d'éléments saillants ;
- Il faut lire, ce qui nécessite du temps.



Datavisualisation

- Mise en évidence des éléments faibles ou forts par la longueur des barres et le jeu des couleurs ;
- Lecture rapide des données ;
- S'accompagne souvent d'un texte explicatif.



Infographie

- Visualisation immédiate des éléments à retenir d'une étude : texte + chiffres ;
- Nécessite pour le producteur un tri de l'information ;
- Jeux de textes, graphiques, chiffres, pictos ;
- Permet une compréhension facile et une meilleure communication.

Collection : Insee Références
Femmes et hommes, l'égalité en question (Édition 2022)

Le service statistique du ministère chargé de l'agriculture a présenté de cette manière les résultats du recensement agricole de 2020 (*Le Grand, 2022*) ; de même, l'office national de statistique du Royaume-Uni utilise cette façon de faire pour diffuser les résultats du recensement de la population (*How the population changed where you live: Census 2021*). Le *scrollytelling* est engageant pour le lecteur mais nécessite toutefois un grand effort de scénarisation, comme les infographies associées à des études. La démarche consiste à hiérarchiser les résultats, les présenter par ordre d'importance décroissante selon la technique de la pyramide inversée, technique rédactionnelle journalistique (*Angel, 2009*).

▶ Datavisualisation interactive

Avec le *scrollytelling*, l'utilisateur est actif, il fait défiler les pages, mais la datavisualisation interactive va plus loin dans sa relation avec les chiffres : elle autorise la libre exploration des données par tout un chacun. Elle offre la possibilité à l'utilisateur d'explorer les données, à l'image du statisticien qui les examine avant de les diffuser (***encadré 2***).

Le développement de l'interactivité et la multiplicité des représentations graphiques possibles autorise la prise en main, par le plus grand nombre, d'ensembles de données de grande dimension. Les pyramides des âges interactives ont été l'un des premiers exercices de ce type : l'outil permet par le biais d'une animation de faire défiler les résultats sur plus de 70 ans. Plus proche de nous, l'ouvrage de référence « Tableaux de l'économie française », a été entièrement refondu et repensé dans une optique utilisateur : si la version papier annuelle a été conservée, le produit est aujourd'hui modernisé, actualisé en continu avec les dernières informations, fournissant un meilleur service aux internautes avec différents niveaux de lecture, tout en préservant une simplicité d'alimentation pour les gestionnaires. La définition de cet ouvrage a nécessité des compromis dans le choix des indicateurs et leur mise en valeur, en évaluant leur caractère structurant, leur actualité, leur capacité d'appréhension par le public. La nomenclature des thèmes a été calquée sur celle d'« insee.fr » pour que l'internaute se retrouve facilement dans les rubriques. Dernier en date, l'outil interactif sur *les espérances de vie* rassemble dans une même enveloppe toutes les espérances de vie calculées par l'Insee depuis 1946, déclinées par sexe, âge, territoire, diplôme, niveau de vie et catégorie sociale. L'utilisateur peut s'approprier cette mine d'informations par une vidéo didactique et modifier les différents critères par le jeu de figures interactives.

Cependant, le parcours utilisateur doit être suffisamment réfléchi pour que l'internaute s'y retrouve. L'expérience utilisateur (UX, acronyme pour l'anglais *User eXperience*) doit être pensée grâce à des tests, notamment avec des utilisateurs externes.

▶ Diffuser massivement…

Le statisticien conserve son rôle d'accompagnement dans la mise à disposition des données. Tout d'abord, le choix des croisements doit être pertinent (une simple liste à la Prévert des variables ne suffit pas pour une diffusion large) tout comme la mise en forme des données. Les métadonnées (concepts, variables et modalités) sont essentielles pour permettre la bonne appréhension des données.

> ***Les métadonnées (concepts, variables et modalités) sont essentielles pour permettre la bonne appréhension des données.***

La statistique publique diffuse de plus en plus massivement des données structurées, nécessaires à la datavisualisation. Elle met ainsi à disposition des outils[6] permettant à la fois de mieux découvrir (catalogue de jeux de données) et mieux explorer les données (tableaux dynamiques) ; les métadonnées sont essentielles pour ces usages. Les outils de gestion des données (*datastewardship* ou *datamanagement*[7]) permettent en effet d'élargir la réutilisation des données ; d'autres acteurs peuvent s'en emparer pour proposer d'autres valorisations des statistiques publiques.

▶ Encadré 2. La dataviz au service de la qualité

L'aspect exploratoire des données renvoie au travail du statisticien, dans son appropriation initiale des données collectées, avant leur diffusion. Ainsi, « l'analyse exploratoire des données, c'est donc, pêle-mêle, des instruments efficaces et simples de mise en œuvre, des logiciels puissants et ergonomiques, la réhabilitation des graphiques comme outils d'analyse, mais aussi et surtout une attitude du statisticien face à son problème et ses données. » (*Destandau et alii, 1999*).

Le statisticien mobilise différentes représentations des données pour les analyser, repérer des sous-populations particulières, identifier les erreurs ; sur le *web*, les possibilités sont variées, le recours aux outils de datavisualisation contribue à l'amélioration de la qualité des données en multipliant les possibilités de contrôles. À titre d'exemple, une carte de déplacements domicile-études a ainsi permis de corriger des données aberrantes dans la diffusion du recensement de la population. La première datavisualisation interactive sur les salaires est née des outils de validation de ces données. Cet outil interactif (*Dataviz salaires*) permet une exploration en profondeur de la distribution des salaires, avec différentes dimensions d'analyse (âge, secteur d'activité, géographie, etc.). Ce type d'outil est un moyen d'analyse des données puissant : il permet de repérer, pour une sous-population particulière, certains résultats qui n'auraient pas été mis en évidence ou n'auraient pas trouvé leur place dans les publications à vocation large.

Mis à disposition sur le site de l'Insee, l'outil permet aux internautes d'accéder à des données extrêmement détaillées de façon simple et rapide (*user-friendly*). C'est un exemple d'application du concept de réutilisation dans l'acronyme FAIR*.

* *L'acronyme FAIR est utilisé pour désigner les propriétés de l'open data :* Findability, Accessibility, Interoperability, and Reuse *ou Facilité de recherche, Accessibilité, Interopérabilité et Réutilisation.*

▶ ... en gardant la rigueur du statisticien

Pour viser le plus grand nombre, il est conseillé d'utiliser des mots simples[8] sans renoncer à la rigueur scientifique ni au recours aux concepts adéquats. Par exemple, il faudra veiller à ce que les définitions soient aisément accessibles sur le site de diffusion.

6 Le projet Melodi développe ces outils pour l'Insee. Les services statistiques des ministères en charge de l'enseignement supérieur, de la santé et de la culture ont déjà des solutions opérationnelles.

7 *Datastewardship* et *datamanagement* sont deux termes utilisés pour désigner les processus, outils et techniques de gestion des données stockées dans une entreprise, dans un triple objectif de cohérence, qualité et sécurité. *https://www.oracle.com/fr/database/definition-data-steward/*. *https://www.lebigdata.fr/data-management*.

8 Pour rendre accessibles à un public élargi certains de ses résultats d'études, la DREES (Direction de la Recherche, des Études, de l'Évaluation et des Statistiques), service statistique du ministère chargé de la santé et des solidarités, publie des transcriptions selon la méthode Facile à lire et à comprendre (FALC). *https://www.culture.gouv.fr/Thematiques/Developpement-culturel/Culture-et-handicap/Facile-a-lire-et-a-comprendre-FALC-une-methode-utile*.

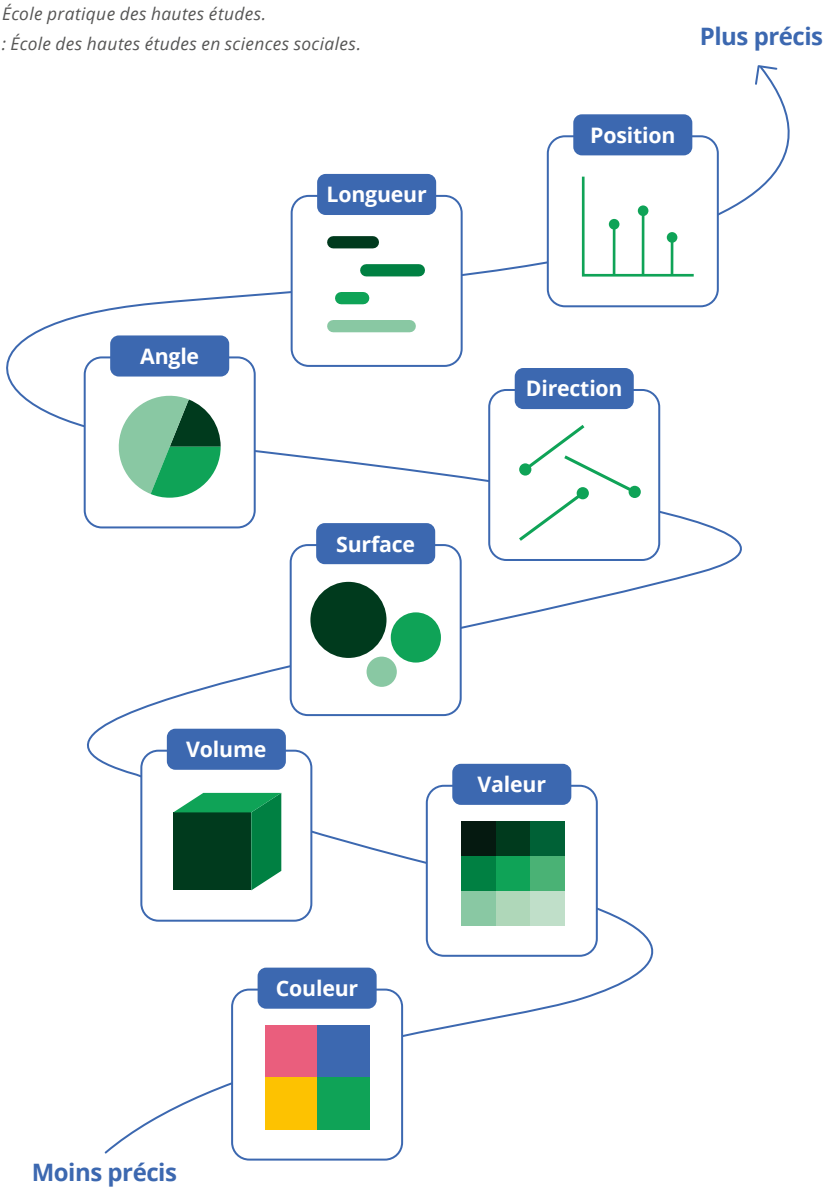
▶ Encadré 3. La théorie autour de la sémiologie graphique

Les variables visuelles clés d'un graphique sont la taille, la valeur, le grain, la couleur, l'orientation et la forme (*Bertin, 1967*). Selon la nature des caractéristiques des données à représenter, Jacques Bertin a déterminé la combinaison de variables visuelles la plus adaptée : « *… si pour obtenir une réponse correcte et complète à une question donnée, […] une construction requiert un temps d'observation plus court qu'une autre construction, on dira qu'elle est plus efficace pour cette question* ». Ce corpus théorique a été construit grâce à l'expérience qu'il a accumulée au cours de sa carrière au laboratoire de cartographie de l'EPHE* (aujourd'hui EHESS**).

Quelques années plus tard, les fondements d'une théorie scientifique ont été posés pour étayer les principes sémiologiques (*Cleveland et Mac Gill, 1984*), en s'appuyant sur des tests de perception des valeurs par les différents éléments d'un graphique. Le système précédent de variables visuelles a été raffiné. *In fine*, ajouter une échelle (ou plusieurs) est préférable aux autres façons d'indiquer une valeur, l'ombre et la couleur n'arrivant qu'en dernier.

* *EPHE : École pratique des hautes études.*

** *EHESS : École des hautes études en sciences sociales.*



Plus généralement, pour garder une diffusion cohérente, il est essentiel de respecter une charte graphique et des règles éditoriales, en particulier les règles de la statistique publique sur les outils de dataviz : unité statistique, échelles de données clairement affichées, titre, date, champ pour chaque illustration, affichage des sources, données accessibles en téléchargement, respect du secret statistique (*Darriau, 2020*). Cela participe à l'image de marque de l'institut de statistique.

Comme pour une publication de quatre pages ou plus, la réalisation d'une datavisualisation et notamment d'une infographie nécessite une sélection drastique de l'information à présenter, une focalisation sur les résultats les plus significatifs. Il faut renoncer à tout dire, ce qui peut être difficile pour le chargé d'études. Gardons toutefois en tête que, dans l'univers statistique, la dataviz reste le plus souvent complémentaire au contenu détaillé sous forme d'études, de documents de travail, etc. Elle ne remplace ni les données ni les publications et la méthodologie associée. Elle joue un rôle de produit d'appel, d'incitation à lire l'étude.

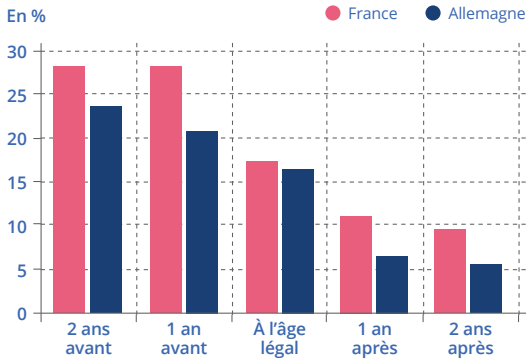
▶ Avec des représentations graphiques soignées…

Une datavisualisation a 30 fois plus de chances d'être vue et lue qu'un simple texte ou tableau. Mais comment faire pour que la dataviz transcrive au mieux les données ? La sémiologie répond à cette question (***encadré 3***) : c'est « l'ensemble des règles d'un système graphique de signes pour la transmission d'une information » (*Bertin, 1967*) qui vise l'efficacité. L'efficacité est renforcée par l'idée du minimalisme (*Tufte, 1983*) : on vise alors « l'excellence graphique ». Cette dernière est atteinte lorsque la quantité d'informations est transmise au lecteur dans un temps minimal et avec le moins d'encre possible. La notion de taux d'encrage (*data-ink ratio*) est introduite (***figure 5***) : il s'agit du rapport entre la partie – minimale et nécessaire – du graphique représentant des données qu'on ne peut pas effacer sans réduire les informations diffusées, et le total de l'encre imprimée du graphique. Ainsi, sur un quadrillage, on évitera de mettre une grille apparente qui relègue les données au second plan ou *a minima* gêne la lecture. L'idée est d'éliminer autant que faire se peut tout ce que l'on peut enlever à un graphique sans perdre de sens pour la visualisation (*chart junk*).

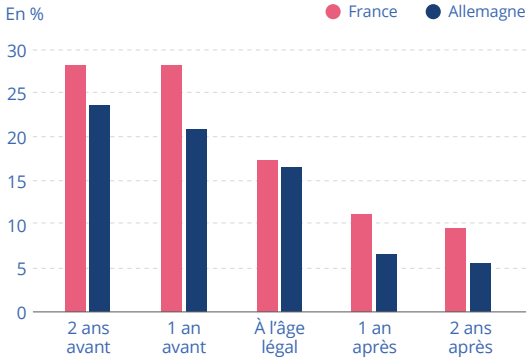
Les couleurs participent également à une meilleure compréhension des graphiques. Mais trop de couleurs rendent un graphique illisible ! Elles doivent être les plus épurées possible avec une palette restreinte et harmonieuse (***figure 6***). Par exemple, il faut attribuer une seule couleur à un seul et même type de données. Enfin, tous les graphiques doivent suivre les normes d'accessibilité par rapport aux différentes déficiences visuelles (daltoniens, malvoyants, etc.).

▶ **Figure 5 - Les taux d'encrage**

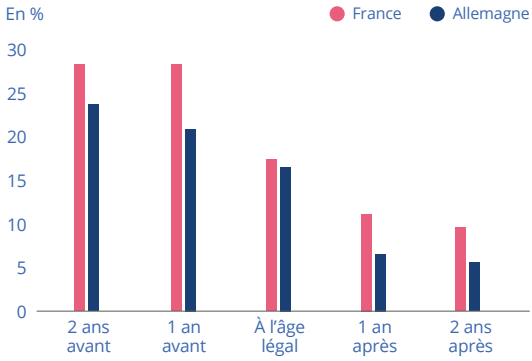
Taux d'encrage important



Taux d'encrage moyen

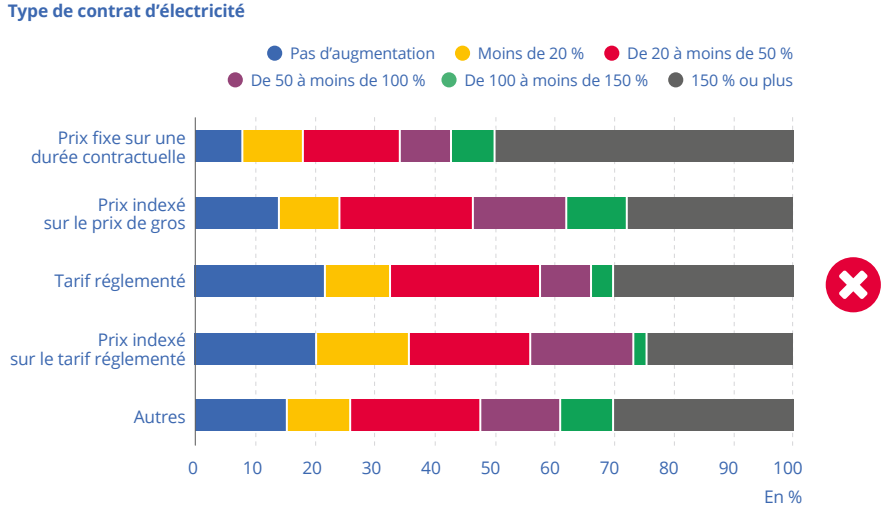


Taux d'encrage bas

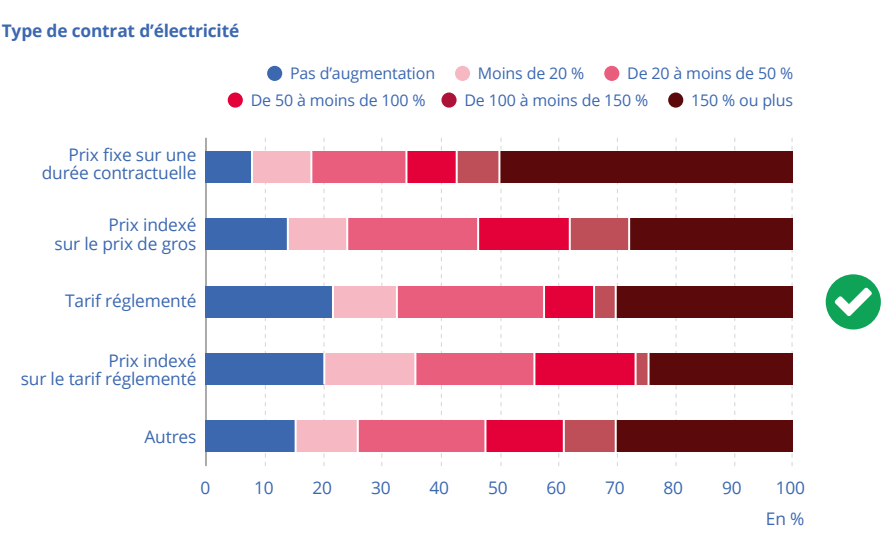


▶ Figure 6 - Exemple : Prix de l'électricité

Trop de couleurs



Palette de couleurs restreinte

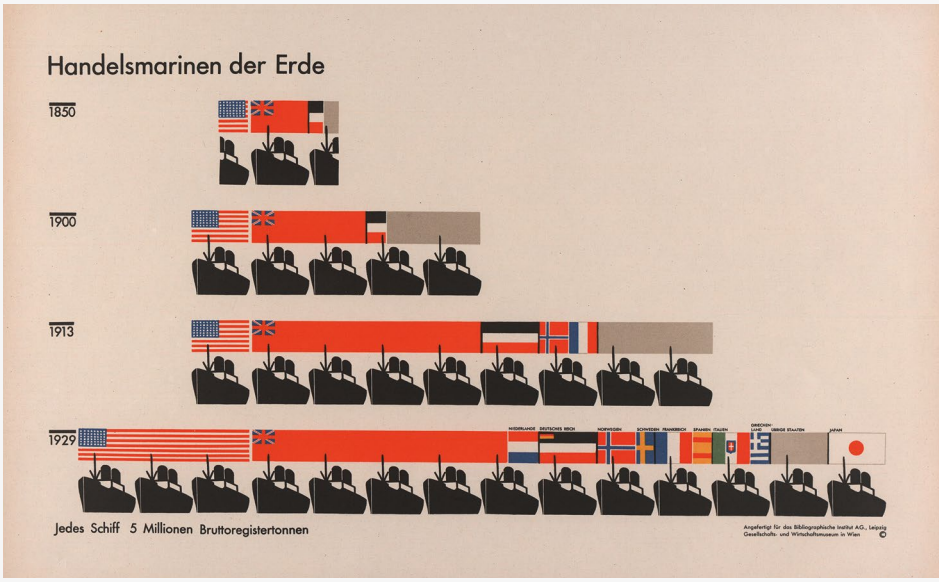


▶ ... et des illustrations parlantes

> ***Une meilleure accessibilité à l'information pour la population la plus large.***

Comme pour le texte, un recours au langage courant facilite la transmission du message à l'exemple de la convention d'usage : rouge pour le négatif et vert pour le positif. Ainsi, en Allemagne, les isotypes (*International System Of Typographic Picture Education*) sont créés (*Neurath et alii, 1925*) formulant un langage visuel, simple et universel influencé par l'esthétique du *Bauhaus*[9] et permettant une meilleure accessibilité à l'information pour la population la plus large. Ce sont les ancêtres de nos pictogrammes (***figure 7***) que l'on retrouve dans la dataviz, et surtout dans les infographies.

▶ Figure 7 - Les marines marchandes du monde



Lecture : Chaque navire représente 5 millions de tonnes brutes de marchandises.

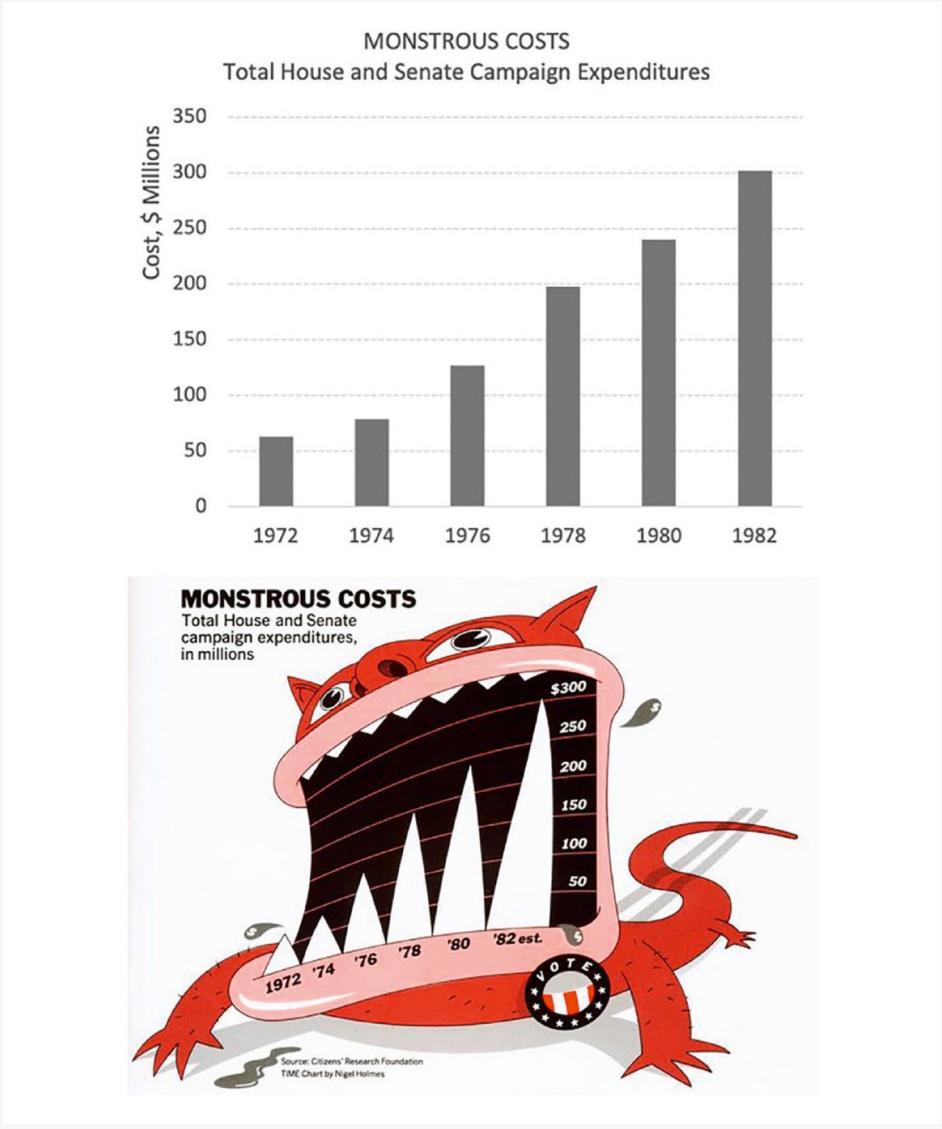
Source : 'Gesellschaft und Wirtschaft' page 55, Verlag des Bibliographischen Instituts AG. Leipzig, 1930.

Les journaux, à grand tirage notamment, utilisent fréquemment des illustrations qui vont au-delà de l'approche minimaliste pour mieux communiquer, comme l'illustre la représentation de coûts jugés monstrueux par un diagramme dans la gueule d'un monstre (***figure 8***). Des « embellissements » peuvent faciliter la transmission d'un message et sa mémorisation (*Bateman et alii, 2010*).

9 Courant artistique né en Allemagne à Weimar, 1919, concernant notamment l'architecture et le design.

Ces auteurs montrent, au travers d'une expérience sur deux séries de graphiques (avec et sans embellissements), que la lecture des graphiques n'est pas déformée par les ajouts, que la mémorisation à court terme (5 minutes) n'est pas significativement différente, mais qu'après deux ou trois semaines, les graphiques embellis sont mieux décrits que les autres.

▶ Figure 8 - Coûts monstrueux



Source : Holmes, N. Designer's Guide to Creating Charts and Diagrams, Watson-Guptill Publications, 1984.

On parle alors de design émotionnel (ou science cognitive, *Norman, 2012*) selon lequel les images attrayantes se mémorisent mieux parce qu'elles stimulent un plus grand plaisir d'utilisation. Les définitions de l'efficacité – uniquement centrées sur des attributs visuels et fonctionnels des visualisations de données – doivent être élargies, car elles ne prennent pas en compte l'individu et les caractéristiques qui lui sont propres : les émotions (*Kirk, 2016*).

De quoi faire réfléchir un statisticien, non ? Finalement, un juste équilibre doit être trouvé et, s'il est intéressant qu'un graphique marque son lecteur et lui permette de mémoriser le message d'un article, il ne doit pas sombrer dans une approche uniquement séductrice : le message et la rigueur doivent conserver la primauté.

L'auteur veillera à rester neutre, à identifier le vrai message et à ne pas faire mentir les statistiques, comme lorsque le choix de l'échelle masque le phénomène réel ; la dataviz véhicule alors un message significativement différent (*Huff, 1954*). Ce risque de « manipuler » les statistiques n'est pas nouveau, ni propre à leur diffusion de façon visuelle, mais il demeure voire est amplifié. Et ce d'autant plus que les nuances sont difficiles à transcrire ; la dataviz n'est ainsi pas forcément adaptée à tout type de données.

Les principes ci-dessus doivent s'appliquer quels que soient les supports de diffusion et de communication, dont la variété s'est considérablement étendue ces dernières années.

▶ Une variété de supports formant un écosystème

Au-delà du domaine des publications papier ou *web*, supports classiques, la statistique publique, notamment européenne au travers du programme DIGICOM[10], s'est ouverte dans les années 2010 aux nouveaux supports et canaux de communication : la vidéo avec animation graphique (*motion design*), les réseaux sociaux, notamment sur le smartphone, particulièrement consommateurs d'images. La datavisualisation est donc adaptée à cet écosystème. La vidéo renforce la scénarisation des résultats statistiques, elle se rapproche fortement du *datastorytelling* puisqu'elle intègre un flot de narration (***Encadré 4***).

▶ Encadré 4. Webographie

Parce que la dataviz n'est pas un ensemble de règles figées mais un corpus en perpétuelle évolution avec une forte part d'innovation et qu'une bonne partie se développe sur internet, les sites ci-dessous permettront de découvrir différentes réalisations. On y trouve à la fois des sites généralistes et des exemples d'applications interactives spécifiques, à consommer sans modération.*

https://informationisbeautiful.net/.

https://www.awwwards.com/.

https://www.visualcapitalist.com/.

https://www.dataviz-inspiration.com/.

https://visualisingdata.com/.

https://www.data-to-art.com/.

Exemples de *datascrolling* :

https://www.spiegel.de/wissenschaft/zirkel-der-genies-a-90c50289-30ac-4a4b-bc49-348676ce6687.

https://vizagreste.agriculture.gouv.fr/age-et-devenir-des-exploitations-agricoles.html.

* *Sites en ligne [consultés le 1er décembre 2023].*

10 DIGICOM est un programme européen sur la communication numérique, actif entre 2018 et 2022.

Par ailleurs, les images favorisent la viralité de l'information *via* les médias ou les réseaux sociaux. À titre d'exemple, un tweet sur une publication aura d'autant plus d'impact qu'il sera associé à une image (graphique de la publication ou infographie spécifique). Les principaux résultats chiffrés de l'étude toucheront ainsi un grand nombre de lecteurs qui les garderont en mémoire.

Le recours à ces nouveaux médias engendre des étapes supplémentaires à la diffusion de l'étude ; pour la vidéo, la gestion (à la seconde près) du temps, le choix entre la captation de vue réelle ou le *motion design* doivent être réfléchis et pour les réseaux sociaux, un tweet de 280 caractères issu du chapô doit être rédigé.

▶ Utiliser des composants les plus génériques possibles…

Encore plus complexes techniquement, les outils interactifs se sont progressivement développés sur insee.fr. Le premier outil a été le simulateur d'inflation, puis les premières pyramides des âges, adaptés tous deux d'une expérience de l'office allemand de la statistique (Destatis). Il s'agit de véritables applications informatiques conçues le plus souvent autour d'une thématique ciblée.

L'expérience d'Eurostat en la matière est intéressante : dans le cadre du projet DIGICOM, un « *work package* » avait pour objectif de préparer une publication interactive sur l'ensemble du territoire européen, intégrant la traduction dans les différentes langues. Le 20 octobre 2017, jour de la statistique européenne, le projet a abouti à la publication « La vie des femmes et des hommes en Europe », reprise sur le site de l'Insee[11]. Cet exercice a été reconduit sur plusieurs thématiques, dont plusieurs ont été traduites et diffusées sur insee.fr. En effet, l'architecture technique définie, à base de briques graphiques imbriquées avec du texte, était suffisamment souple pour s'adapter à d'autres thématiques et permettait de faire évoluer les fonctionnalités du produit : il suffisait de développer un nouveau type de visualisation pour qu'il puisse être utilisé dans d'autres publications. Cette évolutivité a permis à Eurostat de mettre en ligne une dizaine d'outils en quelques années, en travaillant essentiellement sur les aspects éditoriaux et non plus sur les briques techniques.

> ***Il suffisait de développer un nouveau type de visualisation pour qu'il puisse être utilisé dans d'autres publications.***

▶ … simples et robustes dans le temps

À l'Insee, le choix privilégié a été un développement interne du site *web* : c'est une volonté liée à la place de la diffusion dans le cœur de métier de l'Institut. La bibliothèque graphique du site insee.fr a été conservée dans le cadre du projet Web4G (2014-2016) : conçue par des statisticiens pour les statisticiens, elle remplissait les fonctions attendues et utilisait une grammaire maîtrisée par les différentes équipes d'alimentation du site. Néanmoins, pour pallier son obsolescence, l'intégration de nouvelles bibliothèques est en cours, permettant de proposer 2 500 nouveaux produits chaque année, à l'état de l'art, tout en conservant la profondeur historique (dix ans) qui fait la richesse du site.

11 *https://www.insee.fr/fr/outil-interactif/3142332/index.html.*

Pour les nouveaux outils interactifs, le retour sur investissement a été jugé meilleur pour les produits pérennes (dont les données sont mises à jour), comme le tableau de bord de l'économie française, que pour d'autres produits *ad hoc*. Cet outil a été conçu pour être alimenté en continu par des équipes internes, en intégrant d'emblée des capacités d'évolution pour le volet territorial. Cette façon de procéder permet de capitaliser en l'adaptant à différentes thématiques et de le maintenir dans la durée.

En effet, il est délicat, dans le cadre de la statistique publique qui porte sur du long terme, de développer des outils au cycle de vie court, au coût de maintenance élevé, avec des données non actualisées.

On est donc loin des solutions sophistiquées, même si de nos jours, avec les dernières innovations, il n'est plus nécessaire d'être un informaticien chevronné pour créer une application et la mettre en ligne sur le *web*.

▶ Maîtriser un environnement évolutif…

Ceci est rendu possible grâce à l'évolution technologique de ces vingt dernières années. Deux grands axes de transformation ont été mis en œuvre, avec d'une part l'apparition d'un écosystème mondial de la donnée et d'autre part une démocratisation du traitement de la donnée. On constate :

- l'explosion du volume de données diffusées dans les années 2000, avec le stockage sur CD puis DVD avant la généralisation de l'accès en ligne, avec les formats de données XML, CSV, JSON (*Dondon et alii, 2023*) ;
- le développement de l'*open data*, avec la directive Inspire[12] qui ouvre la voie en 2007, l'ouverture de data.gov en 2010 outre-Manche puis le changement de paradigme juridique avec la loi pour une République numérique et le service public de la donnée, également déployé au niveau européen avec la mise en place des jeux de données à forte valeur (*High Value Datasets*) : la mise à disposition gratuite de la donnée, hier encouragée, est aujourd'hui la norme[13] ;
- le développement des outils de traitement (R v1 en 2000 et apparition de Rstudio en 2011) et notamment de datavisualisation (création de D3.js en 2011) (***Encadré 5***) ;
- le développement de l'accès aux données *via* internet avec notamment les API REST, qui permettent d'échanger facilement les données, et seulement les données. Les services *web*, comme ceux offerts sur le portail api.insee.fr, sont devenus la porte d'entrée vers les données, permettant de filtrer le champ d'intérêt ou les variables utiles.

> ***Des mouvements à la fois rapides et profonds, qui défient le statisticien public.***

Ces évolutions s'inscrivent dans des mouvements à la fois rapides et profonds, qui défient le statisticien public. Elles sont en effet sources de contraintes et d'opportunités. L'accès par API permet une mise à jour indépendante de la donnée et de sa mise en forme : une simple mise à jour des données et la nouvelle version de l'outil est prête ; ou encore, un simple changement d'interface et l'ensemble

12 La directive européenne du 14 mars 2007, dite directive Inspire, vise à établir une infrastructure d'information géographique pour favoriser la protection de l'environnement.
13 Loi pour une République numérique du 7 octobre 2016 (article 1).

Encadré 5. Un éventail d'outils logiciels sans cesse renouvelé pour la datavisualisation

Dans le prolongement de l'informatique décisionnelle, la promesse de l'outil « Tableau »* est de créer un flot de narration. Comme toute solution propriétaire pour laquelle le passage à l'échelle demande un investissement financier conséquent, « Tableau » n'est aujourd'hui pas répandu dans le service statistique public (SSP). En France, le logiciel est utilisé notamment par la DGFIP**. D'autres solutions peuvent être citées : *Qlik*, un acteur relativement ancien, *Microsoft PowerBI* et des solutions utilisées dans le monde journalistique comme *Datawrapper* ou *Infogram*. La logique de ces produits est de définir un produit de datavisualisation comme on le ferait avec un outil de bureautique ; la reproductibilité n'est pas en soi un objectif.

Derrière ces produits clé en main, d'autres outils accessibles nécessitent une prise en main plus complexe mais offrent des services complémentaires utiles pour le statisticien. Des bibliothèques (comme *Gapminder* et *Highcharts*) offrent des ensembles de graphiques simples pour les pages *web*. La présentation d'Hans Rosling*** sur l'évolution sur 200 ans de l'espérance de vie et du revenu de chaque pays est une référence en matière de datavisualisation : elle montre en particulier que, si la réalisation technique est importante, la scénarisation l'est encore plus pour engager le spectateur.

Le spectre est complété par la bibliothèque Javascript « D3.js », dont la richesse fonctionnelle est illustrée par la page d'accueil de son site. Cette richesse a pour contrepartie une complexité d'utilisation bien plus forte. L'importation et le traitement des données, leur affichage et leur présentation requiert un paramétrage confinant à l'informatique.

Pour le statisticien, les choses sont encore facilitées : nombre de bibliothèques « Javascript » sont portées sous R. Un simple « Enregistrer sous… » avec le package *htmlwidgets* permet alors de disposer d'un code directement lisible par un navigateur.

Le service statistique public a expérimenté différentes voies. La DREES a ainsi développé des datavisualisations sous *Rshiny*, mises sur internet par le biais du site *shinyapps.io*.

Dans la lignée des *notebooks Jupyter* utilisés pour générer des séquences reproductibles à partir des données, la technologie *Observable*****, développée par Mike Bostock, créateur de D3.js, permet de présenter l'ensemble d'une étude sous la forme d'une suite de morceaux (*chunks*) de texte, code et résultats. Chaque utilisateur peut personnaliser les résultats en les modifiant pour s'attacher à un axe d'analyse particulier ou/et en les exécutant pour dérouler le reste de l'analyse.

* *https://www.tableau.com/fr-fr.*

** *DGFIP : La direction générale des Finances publiques est une direction de l'administration publique centrale française qui dépend du ministère chargé de l'économie.*

*** *https://www.presentationzen.com/presentationzen/2010/07/hans-rosling-tips-on-presenting-data.html.*

**** *https://observablehq.com/.*

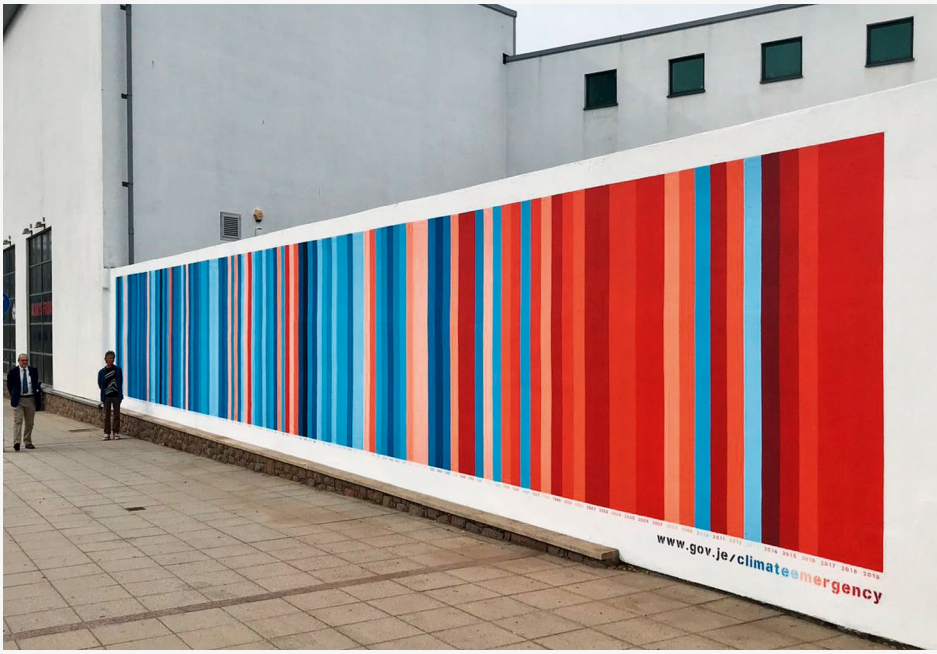
de l'historique du jeu de données est accessible. Mais attention, la pression pour une mise à disposition toujours plus rapide risque de nuire à la qualité des données. D'où l'importance d'une automatisation des contrôles sur les données initiales.

Le mouvement *open data* oblige également à une diffusion de plus en plus large, butant potentiellement sur les contraintes de confidentialité et demandant une analyse en profondeur de l'arbitrage entre utilité et protection de la donnée diffusée ; inversement, il donne ou *a minima* facilite l'accès à de nouvelles sources de données, permettant ainsi des croisements plus nombreux, apportant plus d'informations, qu'il faut alors expertiser puis mettre en valeur avec la datavisualisation.

… au bénéfice (mesuré) de l'utilisateur ?

Pour mesurer l'impact de la datavisualisation dans la statistique publique, on suit les consultations sur internet des produits concernés, le nombre de vues des vidéos, le nombre de tweets, etc. Le résultat de nos actions devient mesurable et est mesuré en continu.

▶ Figure 9 - Climate stripes



Lecture : L'artiste local Ian Rolls a créé une nouvelle fresque murale montrant l'augmentation moyenne des températures de l'air à Jersey, afin d'attirer l'attention sur le changement climatique.

Source : https://www.bailiwickexpress.com/jsy/news/126-reasons-be-green/. © Copyright Bailiwick Publishing 2023

> ***La datavisualisation doit rester une opportunité pour mieux communiquer et porter une image de marque forte de la statistique publique en accord avec son temps.***

L'objectif est non seulement de capter l'attention de l'internaute et de maximiser son temps de lecture mais également de susciter un retour de l'utilisateur (*feedback*) pour vérifier que l'on va bien au-devant de tous les publics et que les chiffres parlent (d'eux-mêmes). Une bonne dataviz peut être largement relayée, augmentant ainsi son impact dans le débat public. À titre d'exemple, l'illustration de l'évolution de la température mondiale annuelle depuis 1850 est partagée et désormais connue (***figure 9***).

La déontologie joue alors un rôle majeur, pour rester neutre et conserver le recul nécessaire à une prise en main objective du chiffre. La datavisualisation doit rester une opportunité pour mieux communiquer et porter une image de marque forte de la statistique publique en accord avec son temps.

▶ Aller plus loin dans la dataviz pour communiquer

En définitive, la statistique publique a pris le tournant de la datavisualisation : celle-ci a été diffusée à la fois dans une version statique, permettant une meilleure appréhension des publications par l'image et la scénarisation, et dans une version dynamique, offrant à voir des monceaux de données. Les problématiques sur ce dernier volet sont importantes et renvoient à l'identité même du statisticien : jusqu'où doit-on donner à voir les données et quel est le niveau de qualité nécessaire pour la diffusion ?

Des ressources spécifiques sont le plus souvent nécessaires (design, montage) : le besoin d'un « designer talentueux » est identifié (*Bateman et alii, 2010*). Le mariage des compétences du statisticien et du designer graphique est nécessaire pour réaliser des produits esthétiques, engageants mais aussi rigoureux et informatifs.

La large mise à disposition de données permet à d'autres acteurs de s'emparer de celles-ci et de réaliser leurs propres datavisualisations. Les *data*-journalistes relayent ainsi l'information diffusée par la statistique publique de plus en plus souvent sous forme de dataviz, ce qui redonne de la valeur aux données produites.

De manière plus légère, la statistique publique doit-elle s'engager dans le *data-art* qui pousse l'esthétique jusqu'à l'œuvre d'art dans la représentation des données ? L'engagement du « lecteur » est une question essentielle et le chiffre doit marquer les esprits. Lors de l'exposition en partenariat avec la SNCF pour les 75 ans de l'Institut, nos travaux ont été présentés sous la forme de grands panneaux pédagogiques. Au-delà de la qualité de nos données, la présence de l'Insee hors de ses murs est essentielle pour porter ses messages.

> ***Un moyen extrêmement efficace pour partager et communiquer autour des statistiques.***

À l'ère du *web*, avec toujours plus de données et de moins en moins de temps pour les lire, la dataviz s'impose comme un moyen extrêmement efficace pour partager et communiquer autour des statistiques. Elle se développe de plus en plus et de façon de plus en plus sophistiquée et la tendance ne semble pas près de s'arrêter. Elle a dès lors un vrai rôle à jouer dans le débat public, et un rôle à part entière dans la communication.

▶ Bibliographie

- ANDRY, Tiffany, KIEFFER Suzanne et LAMBOTTE François, 2022. De Boeck Supérieur. ISBN 978 2807341579

- ANGEL, Jean-William, 2009. Plan d'un article : inversez la pyramide ! In : *Courrier des statistiques*. Insee Hors série, décembre 2009, pp. 21-24. [en ligne]. [Consulté le 9 août 2023]. Disponible à l'adresse : *https://www.bnsp.insee.fr/ark:/12148/bc6p06xt18x.pdf*.

- BATEMAN, Scott, MANDRYK, Regan L., GUTWIN, Carl, GENEST Aaron, Mc DINE, David et BROOKS, Christopher, 2010. *Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Charts*. [en ligne]. [Consulté le 9 août 2023]. Disponible à l'adresse : *http://www.stat.columbia.edu/~gelman/communication/Bateman2010.pdf*.

- BERTIN, Jacques, 1967. *Sémiologie graphique : les diagrammes, les réseaux, les cartes*. Gauthier-Villars.

- BERTIN, Jacques, *1977. La graphique et le traitement graphique de l'information*. Flammarion.

- BRYSBAERT, Marc, 2019. *How many words do we read per minute? A review and meta-analysis of reading rate*. Journal of Memory and Language, Volume 109, Décembre 2019.

- CLEVELAND, William S. and McGILL, Robert, 1984. *Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods*. Journal of the American Statistical Association, Vol. 79, No. 387. Septembre 1984, pp. 531-554.

- CLEVELAND, *William S. How William Cleveland Turned Data Visualization Into a Science*. [en ligne]. [Consulté le 9 août 2023]. Disponible à l'adresse : *https://priceonomics.com/how-william-cleveland-turned-data-visualization/*.

- DARRIAU, Valérie, 2020. Les données carroyées, des outils et méthodes innovants. Pour percevoir la réalité des territoires. In : *Courrier des statistiques*. [en ligne]. 31 décembre 2020. N° N5, pp. 53-73. [Consulté le 9 août 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/5008701?sommaire=5008710*.

- DESTANDAU, Sophie, LADIRAY Dominique, LE GUEN Monique, 1999. In : *Courrier des statistiques*. [en ligne]. Juin 1999. Insee. N°90. [Consulté le 9 août 2023]. Disponible à l'adresse : *https://www.bnsp.insee.fr/ark:/12148/bc6p06xt287/f1.pdf*.

- 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 1er décembre 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635827?sommaire=7635842*.

- DORTIER, Jean-François, 2012. *Une histoire des sciences humaines*. pp.150-153. ISBN 978 -2361061678

- HUFF, Darrel, 1954. *How to Lie With Statistics*. Norton, New York, ISBN 0-393-31072-8.

- KIRK, Andy, 2016. *Data visualisation: A handbook for data driven design*. Londres, Royaume-Uni : Sage Publications.

- KORZYBSKI, Alfred, 1933. *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics*. p. 58.
- LE GRAND, Hervé, 2022. Le recensement agricole de 2020, cinq innovations qui feront date. In : *Courrier des statistiques*. [en ligne]. 20 janvier 2022. Insee. N° N7, pp. 48-67. [Consulté le 1er décembre 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6035938?sommaire=6035950*.
- NEURATH, Otto, 1939. *Modern Man in the Making*. Lars Muller Publishers. ISBN 978-3037786765.
- NIGHTINGALE, Florence, 1858. "*Diagram of the causes of mortality in the army in the East*" Notes on Matters Affecting the Health, Efficiency, and Hospital Administration of the British Army.
- NORMAN, Donald A., 2012. *Design émotionnel : pourquoi aimons-nous (ou détestons-nous) les objets qui nous entourent ?* Bruxelles, Belgique : De Boeck.
- PALSKY, Gilles, 1996. *Des chiffres et des cartes. Naissance et développement de la cartographie quantitative au XIXe siècle*. Collection : Mémoires de la section de géographie physique et humaine - n° 19.
- PLAYFAIR, William, 1805. *A Statistical Account of the United States of America by D. F. Donnant*. London: J. Whiting. William Playfair
- RENDGEN, Sandra, 2020. *Le Système Minard - Anthologie des représentations statistiques de Charles-Joseph Minard* - Collection de l'École nationale des ponts et chaussées. 20 novembre 2020. Éditions B42. ISBN 978-2-490077-45-8.
- RODIGHIERO Dario, 2021. *Mapping Affinities: Democratizing Data Visualizations*. Métis Presses.
- STANDING, Lionel, CONEZIO, Jerry et HABER, Ralph, 2013. *Perception and memory for pictures: Single-trial learning of 2500 visual stimuli*. [en ligne]. [Consulté le 9 août 2023]. Disponible à l'adresse : *https://link.springer.com/article/10.3758/BF03337426*.
- TUFTE, Edward, 2001. *The Visual Display of Quantitative Information*. Graphics Press.

L'ouverture des données au ministère des Armées

Pierre Greffet*

Le ministère des Armées, dont la mission prioritaire est d'assurer la protection du territoire national, de la population et des intérêts français partout dans le monde, produit des données. Certaines peuvent être placées en source ouverte, tandis que d'autres sont couvertes par le secret de la défense nationale et à ce titre inaccessibles au grand public. Dans ce monde binaire, il existe toutefois des cas intermédiaires où la donnée présente une certaine sensibilité mais aussi un intérêt pour les travaux de recherche. Le service statistique de ce ministère, S2E[1], qui bénéficie des mêmes prérogatives que les quinze autres services statistiques ministériels (SSM), est au cœur de ce qui semble a priori inconciliable : préserver la sécurité des données tout en en favorisant l'ouverture. En complément de l'application stricte du secret statistique du fait de leur sensibilité, certaines données peuvent nécessiter la mise en place de mécanismes supplémentaires pour en assurer la diffusion ou l'accès, générant une impression de rareté voire d'absence. La volonté de ne pas mettre en libre circulation des informations sensibles sur les armements ou l'industrie de défense, voire les données opérationnelles des armées, est la première explication à cette impression de rareté de la donnée de défense. La seconde concerne plus directement l'environnement du statisticien ; l'absence de nomenclatures propres au domaine de la défense rend nécessaire des investissements statistiques complémentaires. Toutes les données produites par le SSM défense ne sont pas dans ce cas de figure et certaines ne posent aucun problème d'ouverture. À travers quelques exemples, les contraintes s'imposant à la diffusion de données statistiques dans un contexte de demande croissante d'ouverture des données sont exposées dans cet article ainsi que les mécanismes originaux proposés pour s'en affranchir.

The Ministry of the Armed Forces, whose priority mission is to ensure the protection of France's territory, population and interests throughout the world, produces data. Some data can be placed in open source, while others are protected by national defence secrecy and are therefore off-limits to the general public. In this binary world, however, there are intermediate cases where the data is both sensitive and of interest for research purposes. The ministry's statistical service, S2E[1], which has the same prerogatives as the fifteen other ministerial statistical services (SSM), is at the heart of what seems at first sight to be an irreconcilable conflict: preserving data security while encouraging openness. In addition to the strict application of statistical confidentiality due to their sensitivity, some data may require additional mechanisms to be implemented to ensure dissemination or access, generating an impression of scarcity or even absence. The first reason for this impression of defence data scarcity is the desire not to allow sensitive information on armaments or the defence industry, or even battlefield data, to be disseminated freely. The second relates more directly to the statistician's environment; the lack of defence-specific nomenclatures requires additional statistical investments. Not all the data produced by the Defence SSM falls into this category, and some do not present any problems of openness. Using a number of examples, this article describes the constraints imposed on the dissemination of statistical data in a context of increasing demand for open data, as well as the innovative mechanisms proposed to overcome them.

* Sous-Directeur, chargé du Service Statistique de la Défense
pierre.greffet@intradef.gouv.fr

1 La Sous-direction Statistiques et Études économiques (S2E) autrefois appelée Observatoire Économique de la Défense (OED) est le service statistique ministériel (SSM) du ministère des Armées.

Durant la première partie de la Guerre froide, immédiatement après la mort de Staline, les deux puissances situées de part et d'autre du rideau de fer se sont impliquées dans un partage de données massives. À l'occasion de l'année géophysique internationale[2] (1957-1958), les États-Unis et l'URSS ont démontré qu'il était possible de partager des données stratégiques[3] malgré un contexte politique qui ne s'y prêtait guère et ceci bien avant l'avènement des technologies électroniques utilisées dans notre quotidien (*Aronova, 2017*). À cette époque, les principes fondateurs de la gouvernance de la donnée telle que pratiquée encore actuellement, prennent naissance : des données en source ouverte stockées de façon centralisée dans des centres de données et rendues largement accessibles aux chercheurs. À partir d'un exemple de la statistique publique, on démontre que l'ouverture de certaines données dans un contexte qui ne s'y prête *a priori* pas, est possible et présente même un intérêt certain pour leur producteur.

▶ Des données du ministère des Armées en source ouverte

Comme dans tous les ministères, des données sur le personnel (composé de 204 144 militaires et 61 908 civils[4] en 2022) sont largement accessibles au public à travers le rapport social unique qui détaille les effectifs (***figure 1***) voire les rémunérations. Certaines données caractérisant les forces armées sont aussi accessibles en source ouverte (*open data*) à travers le portail *www.data.gouv.fr*. À ce jour, environ 200 jeux de données portent essentiellement sur les ressources humaines mais pas seulement[5].

Les données concernant les personnels militaires ont toutefois une spécificité, une réglementation, qui impose le respect du plus strict anonymat dans certaines situations (***Encadré 1***). Il est impossible avec ces données de pouvoir localiser des membres des forces spéciales, par exemple. Le sujet de la sensibilité des informations individuelles des

▶ Encadré 1. Le respect de l'anonymat des personnels civils et militaires du ministère de la Défense

Les personnels civils et militaires du ministère des Armées* sont couverts par une réglementation spécifique (arrêté du 7 avril 2011 actualisé le 11 mai 2020, relatif au respect de l'anonymat des militaires et des personnels civils du ministère de la Défense et l'article 39 sexies de la loi du 29 juillet 1881 sur la liberté de la presse**). L'arrêté établit une liste détaillée de 78 services (Direction générale de la Sécurité extérieure (DGSE), Direction du renseignement et de la sécurité de la Défense (DRSD), Direction du Renseignement militaire (DRM), etc.) et d'« unités » (régiments, escadrons de chasse, flottilles, etc.) dispersées sur 226 entités géographiques. Pour respecter le plus strict anonymat, il est impossible de diffuser des détails sur les individus composant ces populations : adresse du domicile, caractéristiques de l'établissement employeur ainsi que les caractéristiques d'état civil.

* *Le ministère de la Défense est l'ancienne dénomination du ministère des Armées pour la période 1974-2017. Les attributions de ce ministère n'ont pas changé en 2017 avec l'adoption de l'appellation actuelle « ministère des Armées ».*

** *Voir les fondements juridiques en fin d'article.*

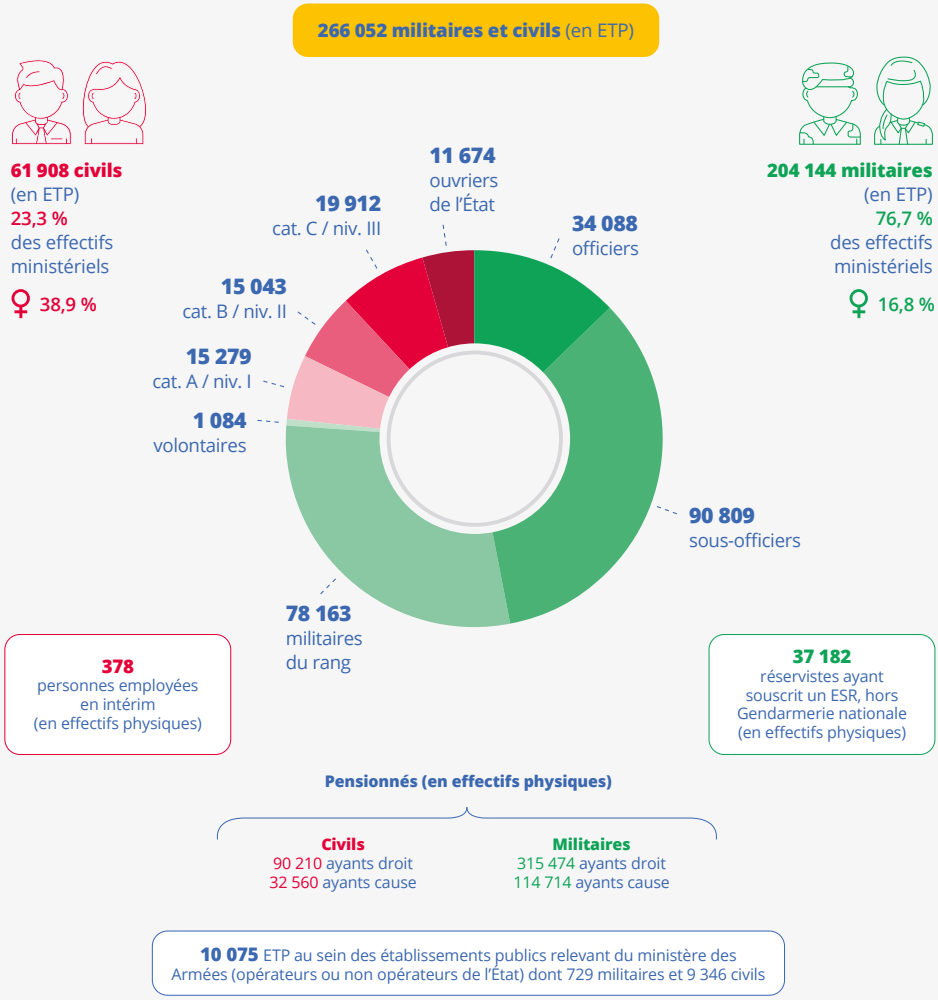
2 L'année géophysique internationale est la période s'étendant du 1er juillet 1957 au 31 décembre 1958, qui a coïncidé avec une activité solaire maximale, et au cours de laquelle plusieurs dizaines de pays ont déployé conjointement un effort particulier dans quatorze disciplines des sciences de la Terre.

3 Il s'agissait de données portant sur l'environnement géophysique : météorologie, géomagnétisme, glaciologie, gravité, radiation nucléaire, océanographie, sismologie, etc.

4 *Rapport social unique 2022* du ministère des Armées.

5 Liste des sites mémoriels, part des véhicules à faible émission dans le parc automobile, données du baromètre de la loi de programmation militaire (LPM) 2019-2025.

Figure 1 - Effectifs des personnels du ministère des Armées en 2022



ESR : Engagement à servir dans la réserve
ETP : Équivalent temps plein

Source : Rapport social unique 2022 – ministère des Armées

personnels militaires est à l'origine de la création, en 1978, du Bureau central de Statistique placé auprès du Secrétaire général pour l'Administration (*De Lapparent, 1980*), première instance de l'actuel SSM défense.

Par ailleurs, les données budgétaires de ce ministère sont aussi accessibles au grand public à travers le portail *www.budget.gouv.fr* ou directement depuis *www.defense.gouv.fr*.

On y apprend que le budget des armées est le troisième poste de dépense du budget général de l'État. Ce budget sert à doter les armées des équipements nécessaires pour accomplir leur mission. Ainsi, le ministère des Armées est aussi un acteur économique public de premier plan qui se caractérise par le montant très important des investissements qu'il réalise chaque année que ce soit au profit des grands groupes industriels mais aussi des PME et TPE : 16,2 Md€[6] de crédits d'investissement inscrits au projet de loi de finances (PLF) 2024[7], soit 75 % des investissements de l'État.

> ***Les données caractérisant le domaine de la défense s'avèrent être rares et cela en constitue une spécificité.***

Au-delà de ces quelques exemples, par rapport à d'autres périmètres ministériels tels que la santé, l'emploi, le commerce extérieur[8], les données caractérisant le domaine de la défense s'avèrent être rares et cela en constitue une spécificité. La statistique de ce domaine ne fait pas exception.

Les données statistiques sur l'économie de défense sont rares

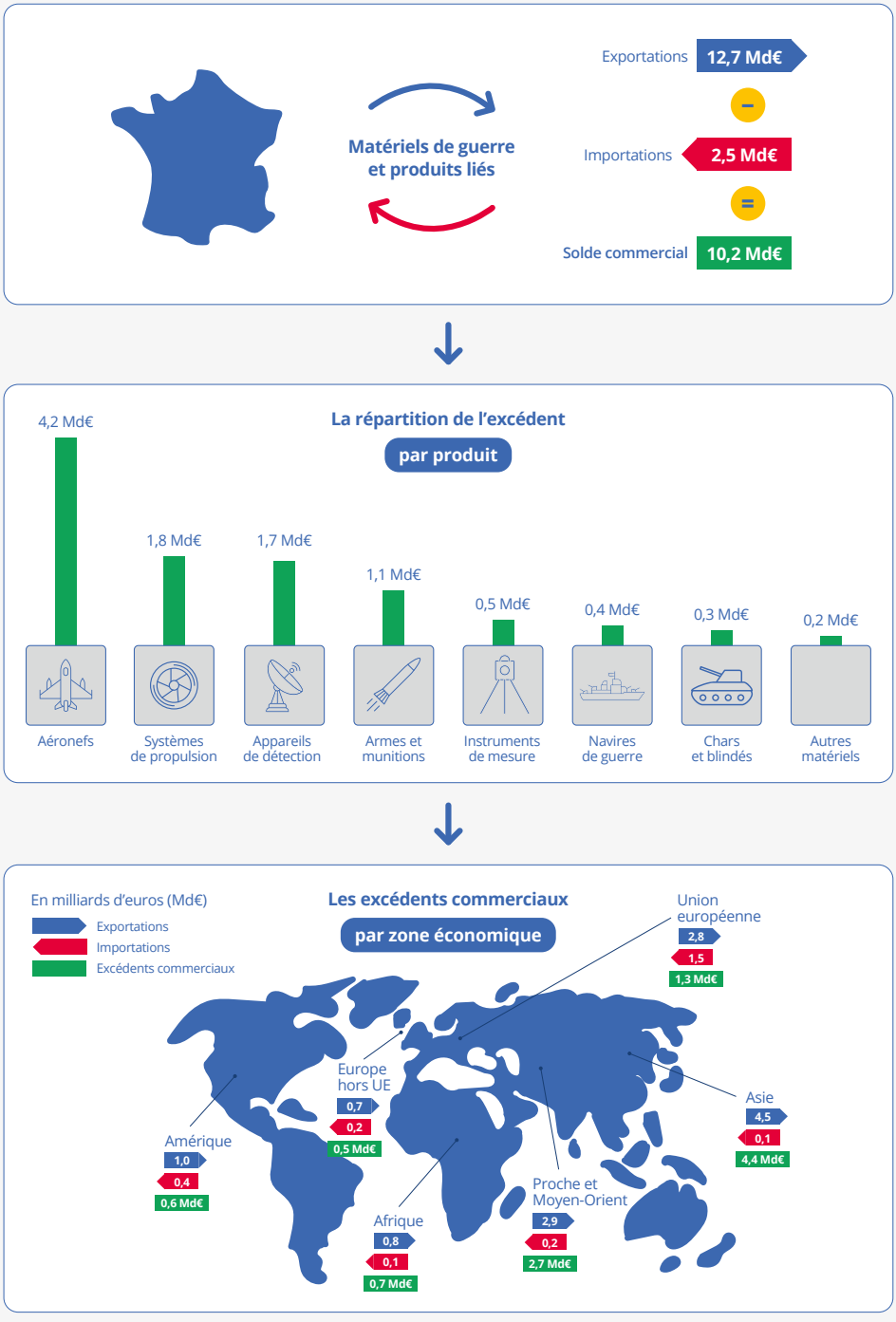
Les données statistiques caractérisant le tissu économique de la défense, désigné sous le vocable de « base industrielle et technologique de défense (BITD) », ainsi que son activité d'exportation, ne sont pas accessibles en source ouverte mais sous forme d'indicateurs figurant dans des notes ou rapports publics.

Dans le domaine des exportations, les seuls indicateurs accessibles au public figurent dans le rapport annuel au Parlement[9] ainsi que dans la note annuelle EcoDef[10] du SSM du ministère des Armées[11] sur les données du commerce extérieur pour 2021 (*Wyckaert, 2023*) (Direction générale des douanes et droits indirects - DGDDI) (***figure 2***).

Le SIPRI (Stockholm International Peace Research Institute) diffuse également des données sur les exportations d'armement (*Stockolm International Peace Research Institute (SIPRI), 2023*) *via* son site internet et selon une méthode originale[12] dans laquelle la valorisation des équipements livrés se fait selon une unité non monétaire, le TIV (*trend-indicator value*) permettant de suivre des tendances historiques mais sans permettre de rapprochement avec des grandeurs macro-économiques telles que le PIB ou la dépense d'armement. Les sources utilisées sont publiques (Defense News, Jane's Defence Weekly, journaux, rapports officiels et le registre des Nations Unies) et les données sont collectées *via* des techniques de « *web scraping* ».

6 *https://www.budget.gouv.fr/documentation/file-download/9508*.
7 « *La mission Défense reste toutefois la première source des dépenses d'investissement du budget de l'État dont elle prend en charge 78 % des crédits de titre 5 (contre 79 % en 2020) alors qu'elle représente (hors CAS Pensions) 9,6 % de l'ensemble de ces dépenses, contre 11,2 % en 2020.* », Cour des Comptes, Note d'analyse de l'exécution budgétaire 2021, Mission Défense.
8 *https://www.health-data-hub.fr/* pour la santé, *https://dares.travail-emploi.gouv.fr/donnees* pour le travail, l'emploi et la formation professionnelle et *https://lekiosque.finances.gouv.fr/site_fr/telechargement/telechargement_SGBD.asp* pour le commerce extérieur.
9 *https://www.defense.gouv.fr/rapport-au-parlement-2022-exportations-darmement-france*.
10 *https://www.defense.gouv.fr/ssm/actualites/ecodef-statistiques-ndeg223-2021-lexcedent-commercial-lie-aux-materiels-guerre-est-au-plus-haut-10*.
11 Le service statistique du ministère des Armées est la sous-direction des Statistiques et Études économiques depuis novembre 2022 (Décret n° 2022-1414 du 8 novembre 2022 modifiant le décret n° 2009-1179 du 5 octobre 2009 fixant les attributions et l'organisation du Secrétariat général pour l'administration du ministère de la Défense).
12 *https://www.sipri.org/databases/armstransfers/sources-and-methods*.

▶ Figure 2 - Exportations de matériels de guerre et produits liés en 2021



Sources : Direction générale des douanes et droits indirects (DGDDI), retraitements Observatoire économique de la défense (OED)

Pour ce qui est de la caractérisation économique de l'activité de l'industrie de défense (BITD), celle-ci est réalisée par S2E au moyen d'une enquête de la statistique publique effectuée deux fois, en 2018 et en 2023. Les résultats du millésime 2018 figurent dans la publication Ecodef n°133[13] et ne comportent pas de données statistiques détaillées sur la répartition par code de la nomenclature d'activités française (NAF) ou par région mais seulement des données statistiques agrégées.

Des nomenclatures statistiques inadaptées à l'activité de défense

> ***La rareté des données statistiques en source ouverte concernant l'activité économique dans le domaine de l'armement est pour partie due à une insuffisante description de l'activité de défense dans les nomenclatures statistiques.***

La rareté des données statistiques en source ouverte concernant l'activité économique dans le domaine de l'armement est pour partie due à une insuffisante description de l'activité de défense dans les nomenclatures statistiques[14] (*Camus, 2022*).

Pour la NAF, par exemple, la sous-classe 30.30Z « Construction aéronautique et spatiale » ne fait pas la différence entre les aéronefs civils et militaires.

Concernant la nomenclature des douanes SH[15], ce sont seulement 7 positions[16] qui peuvent caractériser les matériels de guerre voire les biens à double usage (civil et militaire), et il n'y a pas de distinction claire entre un équipement civil et son équivalent militaire.

Un autre enjeu pour le statisticien est de disposer d'un cadre de référence international partagé permettant d'assurer la comparabilité spatiale des données.

L'absence de règlement européen sur la statistique dans le domaine des activités de défense ne permet pas, en outre, de définir un cadre commun de collecte de données dans ce domaine, et donc d'assurer une comparabilité des données entre les différents États membres. Toutefois, le règlement (UE) 2021/690 du Parlement européen et du Conseil du 28 avril 2021 prévoit dans son annexe II « *la fourniture de statistiques à l'appui de la politique européenne de défense, sous réserve d'études de faisabilité en tenant compte de la sensibilité des données statistiques* »[17].

13 *EcoDef n°133*, « *Près de 30 milliards de chiffre d'affaires militaire pour les entreprises industrielles de la BITD en 2017* », septembre 2019.
14 Sur la nécessité pour le statisticien de disposer de nomenclatures adaptées, l'article « *Le défi de l'élaboration d'une nomenclature statistique des infractions* » du Courrier des statistiques numéro N7 constitue une parfaite illustration.
15 Nomenclature du « système harmonisé » (SH) établie sous la responsabilité de l'Organisation mondiale des douanes.
16 Positions 84, 85, 87, 88, 89, 90, 93.
17 Voir les fondements juridiques en fin d'article.

La production statistique face au secret de la défense nationale

Certaines données produites par le ministère des Armées et dont S2E a besoin pour sa production statistique courante bénéficient d'accès très encadrés, particulièrement dans les cas où s'applique le secret de la défense nationale[18].

Ainsi, pour pouvoir travailler sur ces données, il est nécessaire de disposer d'une habilitation :

« *Conformément aux articles 413-10 et suivants du Code pénal, l'accès par des personnes non qualifiées à des informations ou supports protégés par le secret de la défense nationale est prohibé.*

Pour qu'une personne physique puisse être considérée comme qualifiée au sens du Code pénal, elle doit répondre à deux exigences cumulatives :

- *avoir été dûment habilitée au niveau de classification requis, à l'issue d'une enquête administrative destinée à évaluer les vulnérabilités qu'elle est susceptible de présenter pour le secret de la défense nationale (cf. 3.3) ou être habilitée ès qualités de par la loi ou son statut constitutionnel (cf. 3.1.4) ;*
- *justifier du besoin d'en connaître.* »

Tous les collaborateurs du service statistique ministériel disposent d'une telle habilitation mais cela n'est parfois pas suffisant pour avoir accès aux différentes sources nécessaires pour la production de la statistique. La réglementation prévoit en outre qu'il faut « *justifier du besoin d'en connaître* ». Aussi, afin de faciliter l'accès de S2E aux différentes sources de données du ministère des Armées et prendre en charge de façon générale les sujets relatifs à la confidentialité des données, une organisation originale a été mise en place : le Comité ministériel pour l'information statistique (CoMIS), instance de concertation entre les utilisateurs internes de la statistique au ministère des Armées et le producteur, S2E. À ce titre, le CoMIS (***Encadré 2***) joue, pour le ministère des Armées, un rôle identique à celui du Cnis[19] et reprend le modèle du Conseil de la statistique du ministère de la Justice[20].

> ***Le CoMIS joue, pour le ministère des Armées, un rôle identique à celui du Cnis.***

Dans ses travaux courants relevant de la statistique publique, le SSM se heurte donc à deux difficultés : la sensibilité particulière relative à l'accès aux données du ministère des Armées et l'absence de nomenclature statistique dédiée au sujet de l'économie de défense.

Pour pallier cette deuxième difficulté d'ordre technique, S2E a mis en place un dispositif statistique spécifique : l'enquête sur les entreprises de défense de l'industrie et des services (EDIS).

18 Arrêté du 13 novembre 2020 portant approbation de l'instruction générale interministérielle n°1300 sur la protection du secret de la défense nationale (*http://www.sgdsn.gouv.fr/files/files/Nos_missions/igi-1300-20210809.pdf*).
19 Le Conseil national de l'information statistique (Cnis) assure la concertation entre les producteurs et les utilisateurs de la statistique publique.
20 Arrêté du 11 mars 1994 portant création d'un Conseil de la statistique et des études et d'un Comité de programmation statistique et des études du ministère de la Justice (*https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000729208*).

▶ Encadré 2. Le CoMIS

Le Comité ministériel pour l'information statistique (CoMIS) est l'instance de coordination des travaux de production et de diffusion de l'information statistique du ministère des Armées.

En outre, ce comité a pour fonction d'analyser les contraintes éventuelles dans l'établissement puis la publication des informations statistiques, du fait des missions de sécurité et de défense du ministère.

Les éventuelles exceptions partielles ou totales au principe d'ouverture des données sont appréciées par le CoMIS dans le cadre des textes applicables à la lumière d'une analyse des risques*.

En cela, le CoMIS assure les fonctions du Cnis et du Comité du secret pour le ministère des Armées.

Le CoMIS fait suite au Comité statistique de la Défense créé en 1978, organisme qui réunissait sous la présidence du Secrétaire général pour l'Administration et la vice-présidence du Délégué général pour l'Armement, chacun des chefs d'état-major, le directeur de la Gendarmerie et de la Justice militaire, le chef du Contrôle général des Armées et le chef du Service d'Information et de Relations publiques des Armées.

* *Instruction N° 2804/ARM/CAB du 25 avril 2022 relative à l'information statistique au ministère des Armées.*

▶ La première enquête de la statistique publique sur l'industrie de défense

Lorsqu'on mentionne l'industrie de défense, plusieurs définitions peuvent être données en fonction des liens que les fournisseurs entretiennent avec le ministère des Armées. Une acception large pourrait être de considérer l'ensemble de ces entreprises tel qu'elles sont enregistrées dans Chorus, l'outil de pilotage des dépenses de l'État. L'inconvénient est l'absence de connaissance des sous-traitants et co-traitants intervenant dans la chaîne de valeur.

Le SSM a choisi une approche différente pour caractériser de façon statistique la base industrielle et technologique de défense (BITD) (***figure 3***) : toute entreprise avec plus de 1 % de son chiffre d'affaires dans le domaine de la défense est considérée comme relevant de ce périmètre.

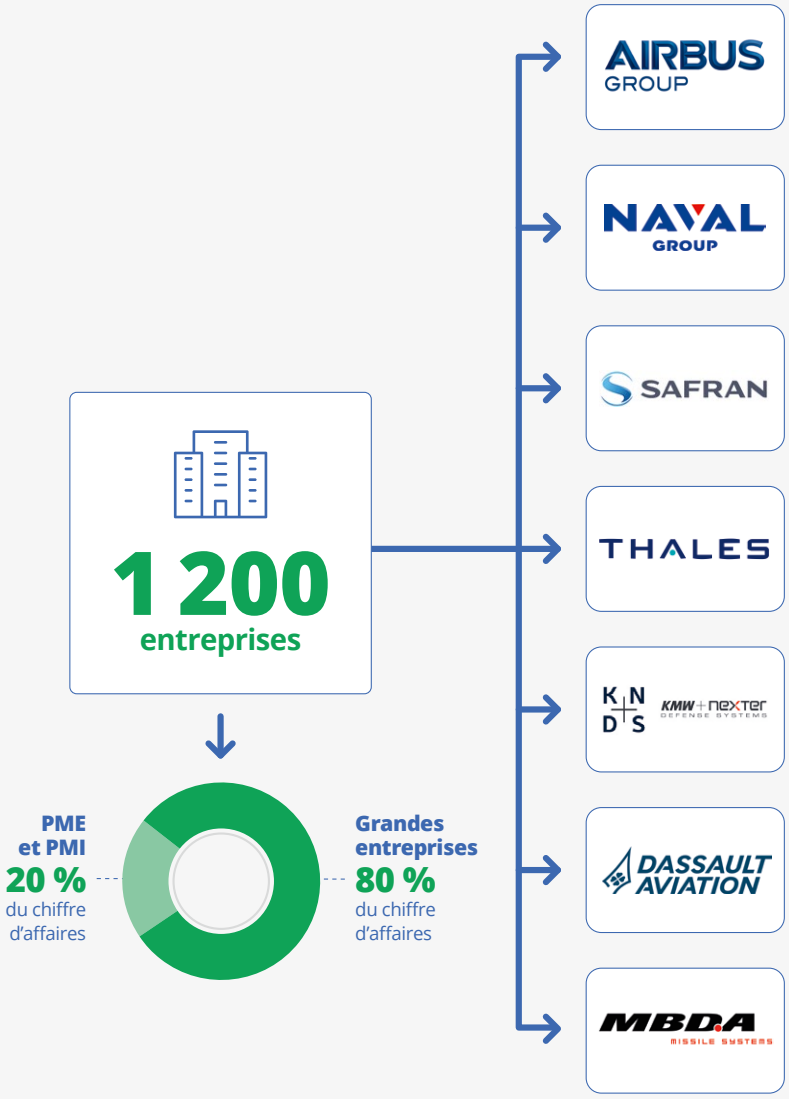
Pour constituer la base d'échantillonnage de cette enquête en l'absence de nomenclature statistique dédiée, S2E a eu recours à une approche empirique.

Jusqu'en 2018, le SSM tenait à jour le répertoire Sandie (Statistiques Annuelles sur la Défense, son Industrie et ses Entreprises) des entreprises liées à la défense sur le territoire français. Il s'agit des entreprises qui fournissent directement ou indirectement des biens et des services utilisés par la communauté de la défense, à savoir les ministères de la Défense (français et étrangers) et les entreprises de la défense elles-mêmes (par les relations de sous-traitance).

Le répertoire Sandie servait jusqu'en 2018 à assurer le suivi des entreprises de défense par le SSM[21]. Il était alimenté à partir de données administratives (Chorus), de données de la statistique publique provenant de l'Insee et de données directement transmises par les grands maîtres d'œuvre industriels. Cette base était confrontée, comme cela peut être le cas pour d'autres répertoires utilisés à des fins statistiques (*Rivière, 2022*), à des difficultés d'actualisation[22].

21 *https://www.irsem.fr/data/files/irsem/documents/document/file/676/EcoDef_55.pdf.*
22 En l'espèce, les fermetures d'entreprises moins bien prises en compte que les créations engendraient des biais dans la qualité des données.

Figure 3 - Les entreprises de la base industrielle et technologique de défense française (BITD) en 2018



PME : Petites et moyennes entreprises.
ETI : Entreprises de taille intermédiaire.

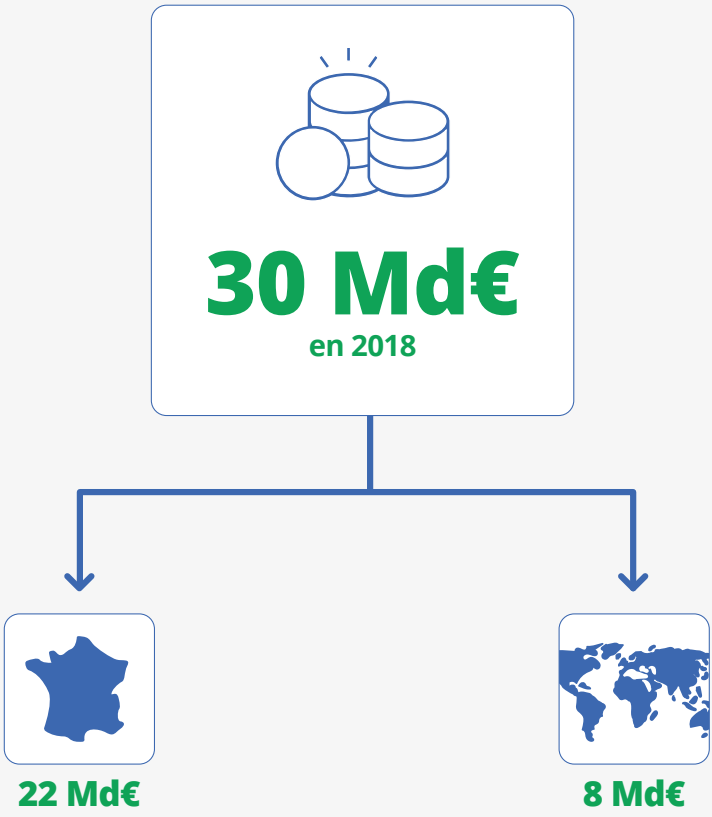
Source : Observatoire économique de la défense

En 2014, le répertoire Sandie comportait 1882 unités légales du secteur marchand. Sa mise à jour a été suspendue en 2018. À la suite de quoi, le SSM a lancé une enquête de la statistique publique, l'enquête sur les entreprises des industries de défense (EID 2018).

À partir de Sandie, une liste de codes NAF a été créée, ceux des entreprises ayant une forte probabilité de se rapporter au secteur de la défense.

La collecte de cette enquête labellisée[23] a débuté le 3 septembre 2018. Le taux de réponse final s'est établi à 85 %, un taux élevé probablement en raison des liens forts entre les entreprises de la défense et le ministère des Armées. La collecte était de type auto-administré avec un questionnaire papier.

Figure 4 - Le chiffre d'affaires de la base industrielle et technologique de défense française (BITD) en 2018



Source : Observatoire économique de la défense

23 L'enquête EID a recueilli un avis favorable du Comité du label du 14 mars 2018.

> ***L'EID a permis de publier pour la première fois une estimation du chiffre d'affaires militaire dans l'industrie de la défense : 30 milliards d'euros en 2018.***

L'EID a permis de publier pour la première fois une estimation du chiffre d'affaires militaire dans l'industrie de la défense : 30 milliards d'euros en 2018 (***figure 4***). Les résultats ont été diffusés sous la forme d'une publication EcoDef mise en ligne sur le site du ministère des Armées en janvier 2019[24]. Quatre autres publications[25] ont par la suite été réalisées.

Cette enquête totalement inédite a été largement valorisée à travers des publications EcoDef ; elle a fait depuis l'objet d'une actualisation et d'une extension de son périmètre pour assurer une meilleure couverture au niveau des entreprises liées à la défense.

▶ Une enquête statistique rénovée

Une nouvelle enquête a été lancée en 2022 : l'enquête sur les entreprises de défense de l'industrie et des services (EDIS 2023) (***Encadré 3***).

Par rapport à la précédente, ses objectifs sont :

- de délimiter précisément le périmètre des industries manufacturières et tertiaires de défense, et de mesurer leur poids dans l'économie française ;
- de collecter l'information statistique nécessaire à la description de son fonctionnement, en particulier son comportement en 2020, en période de crise sanitaire ;
- d'actualiser les données produites.

▶ Encadré 3. Les caractéristiques techniques de l'enquête sur les entreprises de défense de l'industrie et des services : EDIS 2023

- avis d'opportunité obtenu le 07/10/2021 ;
- label le 08/12/2022 (commission « Entreprises ») ;
- unités statistiques : unités légales autres que microentreprises (29 361 unités dans la base de sondage) ;
- champ géographique : France hors Mayotte ;
- variables : la part du chiffre d'affaires militaire des sociétés, part consacrée à la R&D, répartition du chiffre d'affaires par fonction, répartition du chiffre d'affaires par produits militaires et part de l'emploi affectée à la production de biens et services militaires ;
- plan de sondage : 359 strates et 12 100 unités légales ;
- tirage des unités réalisé par la division Sondage du Département des Méthodes statistiques (DMS) de la Direction générale de l'Insee ;
- traitements post collecte : reprise de ceux de la précédente enquête ;
- opérations sous-traitées à un prestataire *via* un marché (avis du Comité du secret du 06/12/2022) : envoi des questionnaires papiers aux entreprises échantillonnées, relances par téléphone, réception des questionnaires, numérisation et saisie ;
- sécurisation : échanges chiffrés avec le prestataire, respect du secret de la défense nationale, destruction à l'issue de l'opération de tous les documents et fichiers relatifs à la collecte.

24 *https://www.defense.gouv.fr/sites/default/files/ssm/EcoDef%20133.pdf.*
25 « Existe-t-il un antagonisme entre défense et environnement ? », EcoDef n°135, septembre 2019, « Dépendance stratégique aux matériaux critiques de la BITD française », EcoDef n°143, janvier 2020, « Les déterminants économiques des exportations de matériels militaires des entreprises industrielles de la BITD française », EcoDef n°147, février 2020, « Le rôle contracyclique joué par les activités militaires dans la crise économique », EcoDef n°196, octobre 2021.

Cette enquête est intégrée au plan d'action pluriannuel du système statistique européen 2021-2027 (MAP pour Multi-annual Action Plan). Elle permet de produire de nouvelles statistiques dans le domaine de la défense, où elles sont encore peu disponibles.

La cible de l'enquête est l'ensemble des entreprises (unités légales marchandes), hors microentreprises et hors entreprises individuelles, de France entière (y compris départements et régions d'outre-mer mais hors collectivités d'outre-mer), quelle que soit leur taille ou leur localisation sur le territoire.

Sont interrogées les entreprises présentes dans le répertoire des entreprises fournisseurs de la défense (REFD[26]), celles qui exportent des matériels de guerre et celles identifiées lors de la précédente enquête. Ce sont environ 12 100 unités légales marchandes qui sont interrogées.

Un comité de pilotage interne au ministère de la Défense a été constitué pour l'élaboration du questionnaire et le suivi du déroulé du projet. S2E assure la maîtrise d'ouvrage de l'ensemble de la procédure. La Direction générale de l'armement (DGA) et les groupements professionnels[27] y ont été associés.

La collecte s'est déroulée de mars à juillet 2023. L'enquête donnera lieu à des premiers résultats publiés en ligne dans la collection EcoDef Statistiques de S2E dès début 2024.

Afin d'étendre la connaissance du champ de l'économie de la défense et en raison de l'impossibilité de multiplier les enquêtes auprès des entreprises, un rapprochement avec le monde de la recherche académique dans ce domaine s'est avéré nécessaire. Ne pouvant héberger des chercheurs en permanence, il a été décidé de construire des partenariats extérieurs avec des organismes de recherche dans le domaine de l'économie de défense.

▶ Le partenariat avec la recherche publique en économie de défense se développe

S2E, consciente des difficultés de se lancer seule dans l'aventure de la caractérisation de l'économie de défense, a toujours veillé à entretenir des liens privilégiés avec la recherche académique dans ce domaine. La connaissance des entreprises de défense est l'une des cinq thématiques principales de la Chaire Économie de défense (***Encadré 4***) de l'Institut des hautes études de défense nationale (IHEDN). Compte tenu de leur richesse mais aussi de leur rareté, les données sur l'économie de la défense sont très demandées ; cependant, comme toutes données statistiques, leur utilisation doit être accompagnée notamment pour assurer qu'elles soient correctement interprétées. Dans ce but, S2E est étroitement associée aux travaux de la Chaire Économie de défense de l'IHEDN en participant à son Comité de pilotage ainsi qu'à son Conseil scientifique[28]. Les travaux réalisés conjointement se concrétisent par des publications communes avec des chercheurs[29] et des participations à des manifestations publiques[30].

26 Le REFD est produit par le SSM défense à partir de données issues du système Chorus.
27 Groupement des Industries Françaises Aéronautiques et Spatiales (GIFAS), Groupement des industries françaises de défense et de sécurité terrestres et aéroterrestres (GICAT), Groupement des Industries de Construction et Activités Navales (GICAN).
28 *https://ecodef-ihedn.fr/gouvernance/#:~:text=Le%20Conseil%20scientifique%20de%20la,et%20du%20Fonds%20de%20dotation.*
29 EcoDef n° 70, EcoDef n° 76, EcoDef n° 94, EcoDef n° 106, EcoDef n° 108-109, EcoDef n° 135, EcoDef n° 178.
30 Journées de l'innovation, colloques organisés par le Réseau de Recherche sur l'Innovation (RRI).

▶ Encadré 4. La Chaire Économie de défense de l'IHEDN

La Chaire Économie de défense de l'IHEDN (Institut des hautes études de défense nationale) est le fruit d'une initiative conjointe « État-Industrie » de soutien à la recherche académique en économie de défense. Depuis sa création en janvier 2014, la Chaire produit des analyses, utilisées comme références par les décideurs publics ou privés (ministère des Armées, ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, Parlement, entreprises, etc.).

Localisée à l'École militaire, dans le 7e arrondissement de Paris, la Chaire est composée d'une équipe de recherche, d'un Conseil scientifique et d'un Comité de pilotage.

Le Comité de pilotage se compose des représentants des mécènes (Airbus, Arquus, MBDA, Naval Group, Nexter, Safran et Thales), de l'IHEDN et des partenaires étatiques (Direction générale de l'armement (DGA), Direction générale des relations internationales et de la stratégie (DGRIS) et Secrétariat général pour l'administration (SGA)). Il fixe les grands objectifs de la Chaire et valide les activités.

Les principales thématiques de recherche sont :

- les impacts économiques et sociaux des efforts de défense ;
- les relations entre acteurs : États et industries ;
- l'économie de défense dans le contexte international ;
- les organisations industrielles ;
- les relations entre défense, recherche et enseignement supérieur ;
- les retombées économiques de l'exportation de défense ;
- les bénéfices économiques des nouvelles coopérations européennes.

L'enquête sur les entreprises des industries de défense (EID 2018) et l'enquête sur les entreprises de défense de l'industrie et des services (EDIS 2023) s'inscrivent dans cette thématique de l'économie de défense en identifiant les entreprises concernées ainsi que leurs caractéristiques détaillées.

L'écosystème de la recherche en économie de défense étant restreint et les travaux du SSM défense étant méconnus, S2E pourrait s'engager à aller au-devant des équipes de recherche dans différentes universités, laboratoires et centres de recherche en économie afin de leur présenter son activité en matière de production de données dans ce domaine. Cette démarche devrait conduire à la mise en place de partenariats pour des travaux d'étude sur des domaines intéressant le ministère des Armées. Les bénéfices potentiels pour les chercheurs seraient de disposer de données nécessaires à la réalisation de leurs travaux dans le domaine de l'économie de défense et pour le SSM défense, et par extension le ministère des Armées, de pouvoir bénéficier d'éclairages externes complémentaires sur les grands enjeux actuels et à venir (économie de guerre, base industrielle et technologique de défense (BITD) européenne, lien armée-nation, etc.).

▶ La promotion de l'économie de défense se renforce

Dans le cadre de son activité de promotion de l'économie de défense auprès des chercheurs (***figure 5***), S2E organise chaque année la remise d'un prix qui récompense des travaux académiques dans ce domaine. Il est décerné par un jury présidé par la directrice des affaires financières du ministère des Armées et composé d'universitaires, de la DGA[31], DGRIS[32] et de l'EMA[33]. L'arrêté du 9 juin 2015[34] en décrit le règlement.

31 Direction générale de l'armement.
32 Direction générale des relations internationales et de la stratégie.
33 État-major des armées.
34 *https://www.defense.gouv.fr/ssm/prix-deconomie-defense*.

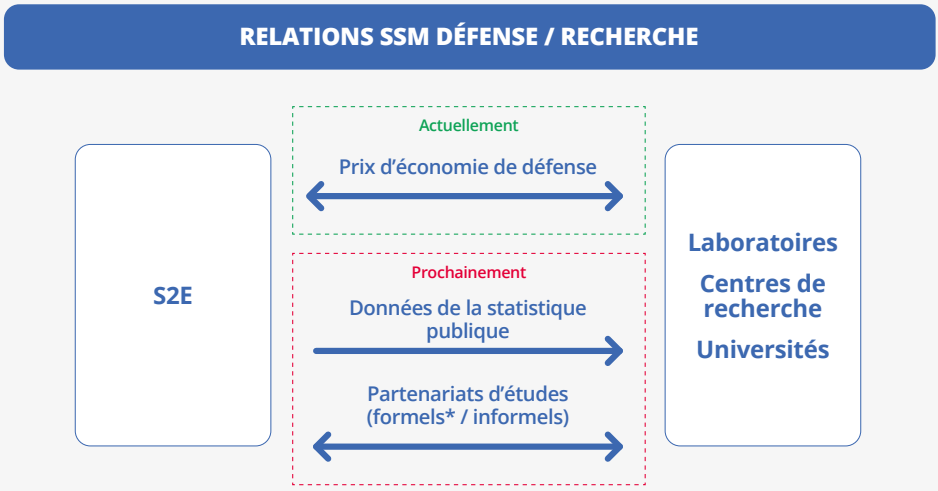
> ***S2E organise chaque année la remise d'un prix qui récompense des travaux académiques dans ce domaine.***

Les sujets de thèse ou de mémoire récompensés jusqu'à présent sont cités dans la rubrique « Bibliographie » en fin d'article (*Droff, 2014 ; Mie, 2016 ; Pietri, 2016 ; Kundu, 2017 ; Meunier, 2017 ; Fauconnet, 2019 ; Fawaz, 2021*).

Une action complémentaire d'information et de sensibilisation est conduite par S2E au profit des hautes autorités du ministère en organisant tous les trimestres des petits déjeuners de l'économie faisant intervenir des grands témoins issus d'horizons divers.

Depuis 2021, sept petits déjeuners thématiques sur des sujets économiques d'intérêt ont été organisés[35].

Figure 5 - Les relations du SSM défense et de la recherche dans le domaine de l'économie de défense



** Conventions*

S2E : Sous-direction des Statistiques et Études économiques, autrefois appelée Observatoire économique de la défense (OED). C'est le service statistique ministériel (SSM) du ministère des Armées.

35 Y ont participé Patrick Artus (Natixis), Xavier Ragot (OFCE), Agnès Benassy-Quéré (DG Trésor), Jean-Luc Tavernier (Insee), Jean-Marc Daniel (professeur émérite à ESCP Business School), Julien Malizard (IHEDN), Maya Atig (Fédération des banques françaises).

Une sensibilité des données qui peut dépasser le secret statistique

Les chercheurs, pour leurs travaux académiques, ne peuvent se contenter de statistiques publiées jusqu'à présent sur le site internet du SSM défense dans ses collections EcoDef. Ils ont besoin d'accéder aux données statistiques détaillées produites par le SSM.

Ces données étant issues d'enquêtes de la statistique publique, celles-ci sont enregistrées auprès du Comité du secret (*Redor, 2023*)[36]. Deux sources y sont référencées : l'enquête sur les entreprises des industries de défense (EID 2018) et l'enquête sur la fréquentation des lieux de mémoire (EFLM) (***Encadré 5***). Au regard des critères de sensibilité, les données statistiques de ces deux enquêtes ne se situent pas sur un même plan. Celles portant sur la fréquentation des lieux de mémoire ne relèvent pas du secret de la défense nationale ni même du secret statistique. Cette enquête est la réponse du SSM défense au besoin d'un ministère très attaché aux symboles et à la mémoire[37].

Encadré 5. L'enquête sur la fréquentation des lieux de mémoire

Cette enquête annuelle interroge l'ensemble des lieux de mémoire (musées, mémoriaux, centres d'interprétation, nécropoles, etc.) des conflits contemporains (guerre de 1870, Première et Seconde Guerres mondiales, conflits postérieurs à 1945), situés en France métropolitaine*, afin de recueillir les données détaillées de fréquentation.

La collecte porte sur la fréquentation de l'année N, est réalisée par un prestataire et se déroule de janvier à mi-mai de l'année N+1. Les résultats sont valorisés dans un EcoDef** diffusé en octobre de l'année N+1 (*Prénée, 2023*).

L'enquête a reçu un avis d'opportunité favorable de la part de la commission « Entreprises et stratégie de marché » du Cnis lors de sa réunion du 29 septembre 2017.

Les questions portent sur le statut juridique du site, le type d'événement qui y est organisé, les conflits concernés, la fréquentation en nombre d'entrées, type de groupes, nationalités étrangères les plus représentées ainsi que les outils numériques mis à la disposition des visiteurs (bornes interactives, applications mobiles, casques de réalité virtuelle, réseaux sociaux, etc.). La liste des sites évoluant régulièrement***, chaque questionnaire de l'année porte sur les deux exercices des années N-1 et N-2 afin de garantir une comparabilité sur deux années.

Le taux de réponse est d'environ 79 %.

* *413 sites ont été interrogés en 2023.*

** *Ecodef n°219 « La fréquentation des lieux de mémoire des conflits contemporains en 2021 », février 2023.*

*** *Le répertoire compte actuellement 423 sites.*

Les données individuelles de l'enquête sur les entreprises de défense de l'industrie et des services EDIS 2023 sont d'un niveau de sensibilité bien plus élevé au regard de la doctrine de préservation de la souveraineté de la BITD française. Ces données sont également couvertes par le secret statistique qui garantit le respect du secret commercial et des affaires.

Comme pour tout ministère, le besoin d'accès aux données se manifeste de la part d'un public, notamment de chercheurs, pour qui cette ressource représente un enjeu scientifique majeur. La rareté des sources accessibles ainsi que la sensibilité à leur ouverture sont également des caractéristiques communes à toutes les administrations.

36 *https://cdap.casd.eu/referentiel.*

37 Les crédits alloués à la mission « Anciens combattants, mémoire et liens avec la Nation » sont de 1,8 Md€ en LFI 2023 et PLF 2024 dont des crédits pour la politique mémorielle de 19,8 M€ pour le 80e anniversaire des débarquements.

> ***Le besoin d'accès aux données se manifeste de la part d'un public, notamment de chercheurs, pour qui cette ressource représente un enjeu scientifique majeur.***

Dans le cas général, la réponse donnée à cette demande légitime d'accès à des données d'enquêtes couvertes par le secret statistique repose sur une procédure formelle d'autorisation qui inclut l'accord de l'autorité dont émanent les données (généralement l'Insee ou un service statistique ministériel), l'avis du Comité du secret statistique[38], puis une décision de l'administration des Archives, puisque les enquêtes statistiques sont considérées comme des archives publiques[39]. *In fine* et après accord de ces instances, l'accès aux données anonymisées se fait *via* le réseau Quetelet-Progedo Diffusion pour des fichiers de production et de recherche ou *via* le Centre d'accès sécurisé aux données (CASD)[40] pour les données les plus détaillées (micro-données) (*Bozio et alii, 2017*). Ce mode de fonctionnement dit « classique » serait *a priori* insuffisant pour répondre aux exigences du ministère des Armées et à sa sensibilité toute particulière vis-à-vis des données.

Pour la mise à disposition des micro-données de l'enquête EID 2017-2018, il a été décidé d'une double autorisation[41] : une autorisation préalable des autorités du ministère des Armées compétentes en la matière (SGA/DGA) puis un avis favorable du Comité du secret statistique.

Pour la mise à disposition des micro-données de l'enquête EDIS et du fait de la mise en place en 2022 du CoMIS, des adaptations à la procédure d'accès aux données devraient être mises en place selon les hypothèses de travail développées ci-après.

La procédure de demande d'accès, comme pour toute demande d'accès à des données d'enquêtes de la statistique publique se ferait dans un premier temps *via* le portail CDAP[42] (*confidential data access portal*). Dans un second temps (encore en cours d'instruction), afin de tenir compte de la double autorisation, l'avis du CoMIS, instance compétente en matière de diffusion de données[43], serait sollicité après une phase d'instruction de la demande par S2E à la lumière d'une analyse des risques. L'avis du CoMIS serait alors transmis au Comité du secret.

Cette procédure (***figure 6***) serait identique à celle mise en place pour toute demande d'accès à des données couvertes par le secret statistique *via* le portail CDAP du Comité du secret, excepté l'ajout de l'étape d'instruction par le CoMIS compte tenu de la sensibilité des données. Le CoMIS donnerait ou non l'accès (accréditation) à la *Data Room*[44]. La saisine du CoMIS serait alors systématique pour toute demande d'accès à la *Data Room*.

38 L'accord de la CNIL doit être sollicité pour tout accès à des données permettant l'identification des personnes physiques.

39 *https://www.comite-du-secret.fr/wp-content/uploads/2017/12/Acc%C3%A8s-%C3%A0-des-donnes-confidentielles-J-P-Le-Gl%C3%A9au.pdf*.

40 Lors des demandes d'accès aux données détenues par le Service Statistique Public *via* le Comité du secret, il est d'usage de les mettre à disposition sur le CASD.

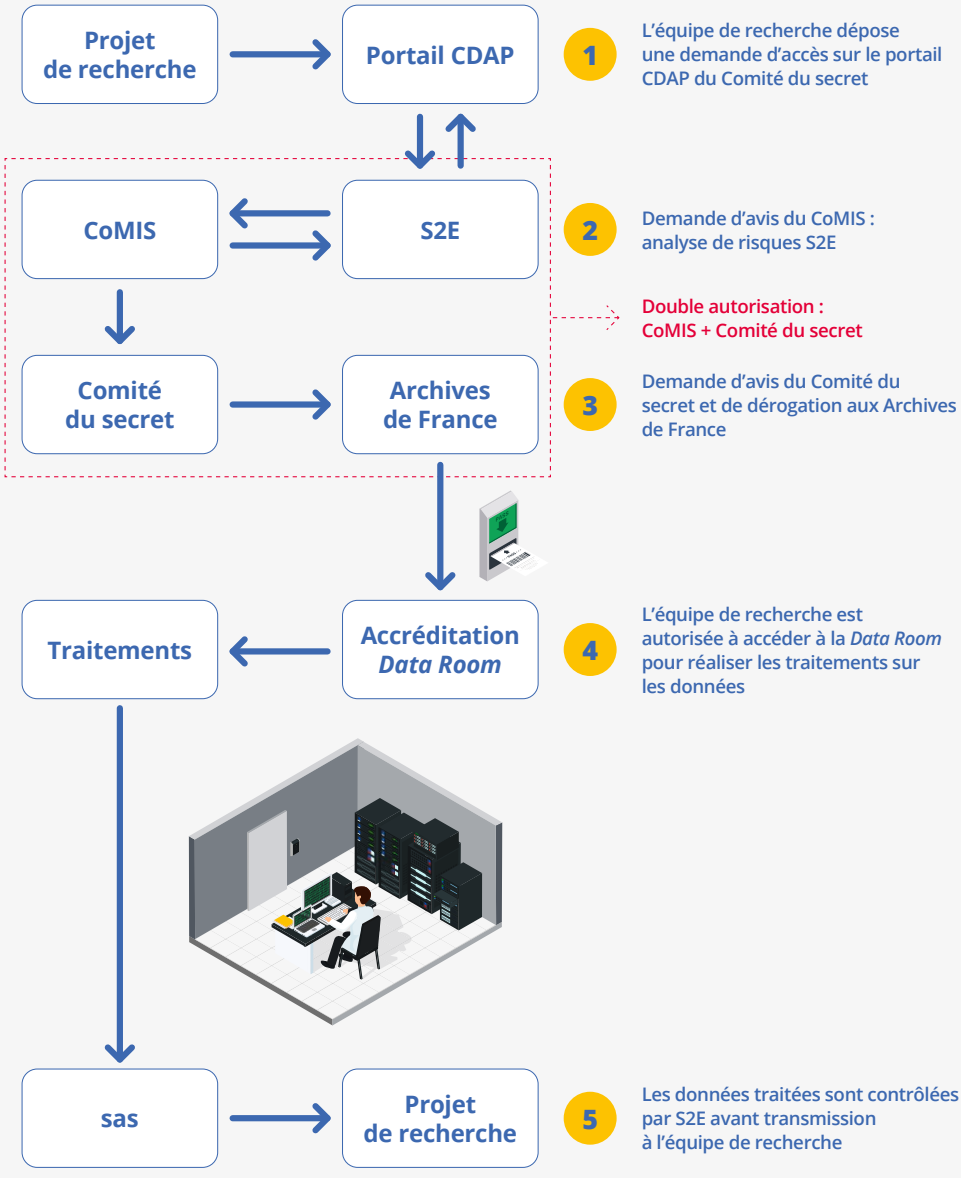
41 *https://www.insee.fr/fr/metadonnees/source/operation/s2056/acces-micro-donnees*.

42 *https://cdap.casd.eu*.

43 « *Le principe d'ouverture des données et leur valorisation peut donc comporter des exceptions, partielles ou totales, dont la nature et la portée seront appréciées par le comité ministériel, dans le cadre des textes applicables, à la lumière d'une analyse des risques.* » (Instruction ministérielle N° 2804/ARM/CAB du 25 avril 2022 relative à l'information statistique au ministère des Armées).

44 Au moment de la rédaction de cet article, le projet de *Data Room* ainsi que la procédure d'accès aux données sont des hypothèses de travail.

▶ Figure 6 - Accès aux données de la *Data Room* pour les chercheurs



S2E : Sous-direction des Statistiques et Études économiques, autrefois appelée Observatoire économique de la défense (OED). C'est le service statistique ministériel (SSM) du ministère des Armées.
CoMIS : Comité ministériel pour l'information statistique. C'est l'instance de coordination des travaux de production et de diffusion de l'information statistique du ministère des Armées.
CDAP : *confidential data access portal*.

La sécurité des données étant de la responsabilité du producteur, à savoir le SSM défense qui relève du ministère des Armées, le principe de souveraineté s'appliquerait de plein droit aux données statistiques et il n'est donc pas envisageable de confier l'hébergement de ses données à un tiers.

La solution couramment privilégiée par les acteurs de la statistique publique pour mettre à disposition leurs données passe par le recours à un tiers, un centre sécurisé d'accès distant aux données (le CASD par exemple) qui permet aux chercheurs habilités d'accéder aux seules données dont ils ont besoin pour leur projet de recherche et de respecter les règles des différents secrets s'y appliquant.

Ce mode de mise à disposition des données n'est pas exclusif, d'autres solutions ont été mises en œuvre par le passé par les producteurs de données eux-mêmes telles que l'*Open Data Room* (ODR) de la Banque de France.

Une solution d'accès à distance aux données présente plusieurs vulnérabilités par rapport à une solution sur site dans la mesure où l'on ne peut garantir à 100 % l'identité de l'utilisateur et l'usage qui est fait des données, puisque, par définition, l'utilisateur n'est pas sur place.

En outre, dans le contexte actuel où la menace d'attaques cyber est particulièrement présente, un accès à distance à des données représente une forte vulnérabilité dans les systèmes d'information et là encore il est impossible de garantir leur inviolabilité.

Implicitement il faut gérer le paradoxe de l'obligation d'ouverture des données d'enquêtes labellisées au Cnis et de la contrainte forte de sécurité liée à la nature même des données concernées.

▶ La « *Data Room* », une solution d'ouverture des données originale

Pour remédier à cela, la solution de mise à disposition des données qui pourrait être envisagée repose sur un principe simple : la mise à disposition des données ne pouvant être confiée à un tiers, l'accès aux données se ferait exclusivement à partir des locaux du ministère des Armées.

Ainsi, l'identité des chercheurs serait vérifiée par les instances du ministère des Armées selon des procédures robustes et maîtrisées et le poste de travail, sur lequel ils effectueraient leurs travaux, serait totalement isolé physiquement de toute connexion à un réseau (principe du « coffre ») ; le risque d'attaque cyber serait donc plus faible que celui d'une solution d'accès à distance.

La solution proposée, la « *Data Room* », consisterait à mettre à disposition des chercheurs, dans un local dédié, un équipement accessible uniquement sur place et doté des ressources nécessaires à leurs travaux de recherches après habilitation de leur projet par le Comité du secret et le CoMIS.

Les résultats des travaux menés sur place par le chercheur pourraient lui être remis après passage dans un système de sas dans lequel des contrôles manuels sur le respect des différents secrets[45] seraient pratiqués systématiquement. La règle appliquée concernant les données

45 Les règles d'application du secret statistique sont décrites dans le guide accessible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/1300624/guide_secret_avril_2023.pdf.*

relatives aux entreprises est celle appliquée par l'Insee. Le respect du secret statistique serait donc contrôlé systématiquement de façon manuelle par les agents compétents du SSM défense selon les mêmes procédures que celles utilisées au CASD (*Gadouche, 2019*) comme il peut l'être pour la production des publications EcoDef.

En outre, le fait d'accueillir les chercheurs dans les locaux du ministère des Armées permettrait à S2E d'être en contact direct avec le monde académique, d'être à l'écoute des attentes des chercheurs, de se tenir mutuellement informés et de leur apporter l'aide dont ils auraient besoin pour mener à bien leurs travaux. La communauté de recherche intéressée par ces sujets étant relativement réduite, il est important pour le ministère des Armées de soutenir ces travaux qui mettent en lumière la place de la défense dans l'économie.

À terme, une solution reprenant les grands principes du SSP *Cloud* (*Comte et alii, 2022*) avec l'interface Onyxia serait pertinente pour compléter l'offre du ministère des Armées à destination des chercheurs : une offre de « *cloud computing* » privée, dédiée, sécurisée et certifiée par les instances techniques du ministère des Armées. Comme pour la solution technique adoptée par le SSPLab, il s'agirait de mettre l'utilisateur au centre des traitements et des données. L'utilisateur serait en capacité de construire l'environnement de travail adapté à son besoin de traitement de données à partir des « briques logicielles » qui lui seraient proposées[46]. Des technologies de « *cloud* » et de conteneurisation seraient mises en œuvre. Enfin, pour satisfaire aux exigences de la recherche publique en matière de « scientificité » et en particulier la reproductibilité des résultats, la solution de conteneurisation de l'environnement offrirait toute la souplesse nécessaire.

En conclusion, produire des statistiques publiques au sein du ministère des Armées ne constitue nullement un obstacle à leur ouverture, même à un niveau fin (micro-données). Des procédures spécifiques tenant compte de la nature même des données et du public souhaitant y avoir accès, sont envisageables. L'hypothèse de travail développée, devrait répondre à une contrainte de sécurisation imposant le recours à une solution « sur site » en lieu et place d'un accès distant offert par le CASD. Une telle solution, outre les garanties qu'elle offrirait en matière de souveraineté, permettrait de faciliter les échanges entre la communauté de recherche académique sur l'économie de défense et les statisticiens et économistes du SSM défense. Internaliser la solution d'accès permettrait de la faire évoluer au gré des besoins facilement. Le nombre d'accès à gérer serait relativement faible, compte tenu de la taille encore modeste de la communauté de recherche dans ce domaine. La démarche entreprise par le SSM défense consiste à démontrer la pertinence de l'ouverture des données ; d'autres directions du ministère des Armées pourraient réfléchir à l'accessibilité de leurs données. Ceci pourrait entraîner la mobilisation d'une plus large communauté de chercheurs susceptibles de répondre aux besoins spécifiques du ministère des Armées.

> ***Produire des statistiques publiques au sein du ministère des Armées ne constitue nullement un obstacle à leur ouverture, même à un niveau fin (micro-données).***

46 Ressources informatiques (mémoire, disque, CPU), logiciels (R, Python), bibliothèques indispensables aux traitements (packages R et Python).

▶ Fondements juridiques

- Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données). In : *Journal officiel de l'Union européenne*. [en ligne]. Mis à jour le 4 mai 2016. [Consulté le 25 octobre 2023]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX%3A32016R0679*.

- Règlement (UE) 2021/690 du Parlement européen et du Conseil du 28 avril 2021 établissant un programme en faveur du marché intérieur, de la compétitivité des entreprises, dont les petites et moyennes entreprises, du secteur des végétaux, des animaux, des denrées alimentaires et des aliments pour animaux et des statistiques européennes (programme pour le marché unique), et abrogeant les règlements (UE) no 99/2013, (UE) no 1287/2013, (UE) no 254/2014 et (UE) no 652/2014 (Texte présentant de l'intérêt pour l'EEE). In : *Journal officiel de l'Union européenne*. [en ligne]. Mis à jour le 3 mai 2021. [Consulté le 25 octobre 2023]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX%3A32021R0690*.

- Article 39 sexies de la loi du 29 juillet 1881 sur la liberté de la presse. In : *site de Légifrance*. [en ligne]. Mis à jour le 24 décembre 2021. [Consulté le 25 octobre 2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/article_lc/LEGIARTI000044568148*.

- Loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques. In : *site de Légifrance*. [en ligne]. Mis à jour le 25 mars 2019. [Consulté le 25 octobre 2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000888573*.

- Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. In : *site de Légifrance*. [en ligne]. Mis à jour le 26 janvier 2022. [Consulté le 25 octobre 2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000886460*.

- Arrêté du 7 avril 2011 actualisé au 7 mai 2020 relatif au respect de l'anonymat des militaires et des personnels civils du ministère de la Défense. In : *site de Légifrance*. [en ligne]. Mis à jour le 11 mai 2020. [Consulté le 25 octobre 2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000023865735*.

- Instruction n°2804/ARM/CAB relative à l'information statistique au ministère des Armées du 25 avril 2022. [en ligne]. [Consulté le 25 octobre 2023]. Disponible à l'adresse : *https://www.defense.gouv.fr/sites/default/files/sga/16%20INSTRUCTION%20N%C2%B0%202804ARMCAB.pdf*.

▶ Bibliographie

- ARONOVA, Elena, 2017. Geophysical Datascapes of the Cold War: Politics and Practices of the World Data Centers in the 1950s and 1960s. In : *Data Histories*. Septembre 2017. Osiris, Volume 32, N° 1, pp. 307-327.

- BOZIO, Antoine, BREUIL, Pascale, GEOFFARD, Pierre-Yves, MALVERTI, Clément et PERRIERE, Manon, 2017. L'accès des chercheurs aux données administratives − État des lieux et propositions d'actions. In : *Rapport du groupe de travail du Cnis*. [en ligne]. Mars 2017. N° 147. [Consulté le 25 octobre 2023]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2019/07/rapportcnis147completweb.pdf*.

- CAMUS, Benjamin, 2022. Le défi de l'élaboration d'une nomenclature statistique des infractions. In : *Courrier des statistiques*. [en ligne]. 20 janvier 2022. Insee. N° N7, pp. 146-161. [Consulté le 25 octobre 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6035948?sommaire=6035950*.

- COMTE, Frédéric, DEGORRE, Arnaud et LESUR, Romain, 2022. Le SSPCloud : une fabrique créative pour accompagner les expérimentations des statisticiens publics. In : *Courrier des statistiques*. [en ligne]. 20 janvier 2022. Insee. N° N7, pp. 68-87. [Consulté le 25 octobre 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6035940?sommaire=6035950*.

- DE LAPPARENT, Jean, 1980. Le Bureau central de la statistique du ministère de la Défense. In : *Courrier des statistiques*. [en ligne]. Avril 1980. N°14, pp.9-10. [Consulté le 25 octobre 2023]. Disponible à l'adresse : *https://www.bnsp.insee.fr/ark:/12148/bc6p06z974h*.

- DROFF, Josselin, 2014. *Le facteur spatial en économie de la défense : application à l'organisation du Maintien en Condition Opérationnelle (MCO) des matériels de défense*. Thèse de doctorat en Économies et finances. Université de Bretagne occidentale. [en ligne]. 1er décembre 2014. [Consulté le 25 octobre 2023]. Disponible à l'adresse : *https://theses.hal.science/tel-00942906*.

- FAUCONNET, Cécile, 2019. *La structuration des bases de connaissance des entreprises de défense*. Thèse de doctorat en Sciences économiques. Université Paris 1 Panthéon-Sorbonne. [en ligne]. 16 octobre 2019. [Consulté le 25 octobre 2023]. Disponible à l'adresse : *https://www.theses.fr/2019PA01E040*.

- FAWAZ, Mahmad. M., 2021. *La dynamique des conflits armés. Contribution à une analyse interdisciplinaire : l'apport de l'économie et du droit*. Thèse de doctorat en Sciences économiques. Université de Bordeaux. [en ligne]. 3 décembre 2021. [Consulté le 25 octobre 2023]. Disponible à l'adresse : *https://www.theses.fr/2021BORD0316*.

- GADOUCHE, Kamel, 2019. Le Centre d'accès sécurisé aux données (CASD), un service pour la *data science* et la recherche scientifique. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. Insee. N° N3, pp. 76-92. [Consulté le 25 octobre 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4254227?sommaire=4254170*.

- KUNDU, Oishee, 2017. *Risks in Defence Procurement: India in the 21st Century*. Mémoire de master. Université de Manchester. [en ligne]. 24 juillet 2019. [Consulté le 25 octobre 2023]. Disponible à l'adresse : *https://research.manchester.ac.uk/files/117821253/manuscript.pdf*.

- MEUNIER, François-Xavier, 2017. *Innovation technologique duale : une analyse en termes d'influence et de cohérence*. Thèse de doctorat en Économie. Université Paris 1 Panthéon-Sorbonne. [en ligne]. 15 septembre 2017. [Consulté le 25 octobre 2023]. Disponible à l'adresse : *https://www.theses.fr/2017PA01E047*.

- MIE, Flavien, 2016. *Un marché de l'observation de la Terre depuis l'espace en mutation*. Mémoire de l'Université Paris II Panthéon-Assas.

- PIETRI, Antoine, 2016. *L'analyse économique des conflits à la lumière de la « Contest Theory »*. Thèse de doctorat en Sciences économiques. Université Paris 1 Panthéon-Sorbonne. [en ligne]. 8 décembre 2016. [Consulté le 25 octobre 2023]. Disponible à l'adresse : *https://www.theses.fr/2016PA01E052*.

- PRENÉ, Léa, 2023. La fréquentation des lieux de mémoire des conflits contemporains en 2021. In : *EcoDef Statistiques*. [en ligne]. Février 2023. Observatoire Économique de la Défense. N°219. [Consulté le 25 octobre 2023]. Disponible à l'adresse : *https://www.defense.gouv.fr/sites/default/files/ssm/Ecodef%20219%20fevrier%202023.pdf*.

- REDOR, Patrick, 2023. Confidentialité des données statistiques : un enjeu majeur pour le service statistique public. In : *Courrier des statistiques*. [en ligne]. 30 juin 2023. Insee. N° N9, pp. 46-63. [Consulté le 25 octobre 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/7635823?sommaire=7635842*.

- RIVIÈRE, Pascal, 2022. Qu'est-ce qu'un répertoire ? De multiples exigences pour un système complexe. In : *Courrier des statistiques*. [en ligne]. 29 novembre 2022. Insee. N° N8, pp. 52-71. [Consulté le 23 octobre 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6665186?sommaire=6665196*.

- Secrétariat général pour l'administration, DRH du ministère de la Défense, 2023. *Rapport Social Unique 2022 du ministère des Armées*. [en ligne]. [Consulté le 25 octobre 2023]. Disponible à l'adresse : *https://www.defense.gouv.fr/sites/default/files/sga/rapport%20social%20unique%20du%20minist%C3%A8re%20des%20Arm%C3%A9es_RSU_2022.pdf*.

- STOCKHOLM INTERNATIONAL PEACE RESEARCH INSTITUTE (SIPRI), 2023. *Armaments, Disarmament and International Security*. In : SIPRI Yearbook 2023. [en ligne]. [Consulté le 25 octobre 2023]. Disponible à l'adresse : *https://www.sipri.org/sites/default/files/2023-09/yb23_summary_fr.pdf*.

- WYCKAERT, Matthieu, 2023. En 2021, l'excédent commercial lié aux matériels de guerre est au plus haut depuis 10 ans. In : *EcoDef Statistiques*. [en ligne]. Mars 2023. Observatoire Économique de la Défense. N° 223. [Consulté le 25 octobre 2023]. Disponible à l'adresse : *https://www.defense.gouv.fr/ssm/ecodef-statistiques/ecodef-statistiques-ndeg223-2021-lexcedent-commercial-lie-aux-materiels-guerre-est-au-plus*.

Quantifier la pratique sportive : une approche sociologique et sanitaire

Augustin Vicard *

De quels outils dispose la statistique publique pour quantifier la pratique sportive ? L'Institut national de la jeunesse et de l'éducation populaire (Injep) pilote le service statistique ministériel chargé du sport. Il a développé un appareil de mesure permettant de quantifier la pratique sportive, avec notamment l'enquête nationale sur les pratiques physiques et sportives (ENPPS) et le recensement des licences et clubs sportifs. Les statistiques produites sont largement mobilisées par les pouvoirs publics, pour suivre les résultats des politiques menées pour développer le « sport pour tous », tout comme par les acteurs économiques de la filière sport. Ce système d'observation s'appuie sur une approche sociologique et économique, visant à mieux comprendre la place du sport dans les loisirs. En cela, elle diffère mais complète une vision purement sanitaire, s'intéressant à la pratique sportive en tant qu'activité physique participant à la lutte contre la sédentarité. Les outils statistiques de l'Injep apparaissent ainsi complémentaires aux enquêtes sur l'activité physique conduites, par exemple, par Santé publique France ou la Direction de la recherche, des études, de l'évaluation et des statistiques (Drees). Ces deux approches nécessitent des outils d'observation différents, et conduisent à structurer différemment les enquêtes (période de référence, définition du sport, etc.). Plusieurs pistes sont évoquées pour améliorer le système d'observation des pratiques sportives ; par exemple, analyser le potentiel des données issues des nouvelles technologies, comme les applications dédiées au sport, mais aussi outiller les acteurs publics locaux, dont le rôle est décisif en matière de politique sportive, à travers une déclinaison territoriale des indicateurs statistiques.

What methods are used by official statistics to quantify the practice of sport? The Institut national de la jeunesse et de l'éducation populaire ***(Injep) runs the ministerial statistical service responsible for sport. It has developed a set of indicators to quantify sporting activity, including the national survey of physical and sporting activity (ENPPS) and the census of sports licences and clubs. The statistics produced are widely used by public authorities to monitor the results of policies designed to develop 'sport for all', as well as by economic players in the sports sector. This observation system is based on a sociological and economic approach, aimed at better understanding the place of sport in leisure time. In this respect, it differs from yet completes a purely health-based vision, focusing on sport as a physical activity that contributes to the fight against a sedentary lifestyle. The Injep's statistical tools are thus complementary to the surveys on physical activity conducted, for example, by Santé publique France or the Directorate of Research, Studies, Evaluation and Statistics (Drees). These two approaches require different observation tools, and lead to structure the surveys differently (reference period, definition of sport, etc.). A number of possibilities have been put forward to improve the system of observing sporting practices; for example, analysing the potential of data obtained by new technologies, such as applications dedicated to sport, but also helping local public players, whose role is crucial in terms of sports policy, by adapting statistical indicators to the local level.***

* Directeur, Institut national de la jeunesse et de l'éducation populaire, *augustin.vicard@jeunesse-sports.gouv.fr*

Le temps consacré aux loisirs excède désormais, en moyenne, celui dévolu au travail ou à la formation, selon les enquêtes Emploi du temps de l'Insee. La question de l'occupation du temps de loisir devient dès lors centrale pour mieux comprendre la société contemporaine.

Dans le temps dédié aux loisirs, le sport occupe une place relativement faible au regard d'autres pratiques, comme regarder la télévision ou d'autres supports numériques. Pour autant, il s'agit d'un loisir valorisé socialement, souvent associé à des valeurs positives, comme en témoigne la place occupée par le suivi des compétitions sportives (50 % des Français suivent régulièrement l'actualité sportive, selon la dernière enquête Pratiques culturelles, menée en 2018, par le département des études, de la prospective et des statistiques (DEPS), service statistique ministériel de la culture).

Des politiques publiques pour promouvoir le « sport pour tous »

> ***L'objectif gouvernemental est d'augmenter le nombre de sportifs réguliers de 3 millions à l'horizon 2024.***

Les politiques publiques pour promouvoir la pratique sportive se développent également en direction des enfants : que ce soit dans le cadre scolaire, avec l'éducation physique et sportive (30 minutes de sport par jour à l'école primaire par exemple), ou dans le cadre péri- ou extrascolaire, *via* par exemple le Pass'Sport, un chèque de 50 € institué en 2021, permettant, sous conditions de ressources, de réduire les frais d'inscription à une activité sportive pour les enfants ou les étudiants boursiers.

Les pouvoirs publics proposent également plusieurs dispositifs à destination des adultes et des personnes âgées, dans une perspective de « sport pour tous » (*Hurtis et Sauvageot, 2018*) : par exemple, le développement du sport-santé a permis de labelliser plus de 550 « maisons Sport-Santé » depuis 2019. L'objectif gouvernemental est d'augmenter le nombre de sportifs réguliers de 3 millions à l'horizon 2024 et de développer le parasport pour les personnes en situation de handicap (*Carlac'h et Le Fur, 2023*), ou encore la pratique sportive des étudiants (*Muller et Lombardo, 2019*).

Deux dimensions d'analyse complémentaires : le sport comme activité physique et comme fait social

Pourquoi les pouvoirs publics veulent-ils promouvoir la pratique sportive ? Tout d'abord, il s'agit d'un loisir actif et souvent collectif, par opposition à des loisirs jugés passifs et solitaires, comme la télévision ou les jeux vidéo. La préoccupation de privilégier les loisirs actifs est au cœur des politiques de santé publique de lutte contre la sédentarité, à l'heure où est évoquée une épidémie de surpoids et d'obésité dans les pays occidentaux.

Cependant, même si cette motivation de santé publique est prégnante dans le soutien des pouvoirs publics à la pratique sportive, elle n'apparaît pas comme le seul déterminant pour les pratiquants eux-mêmes. S'ils mettent en avant l'amélioration de leur santé comme premier motif de pratique, nombreux sont ceux qui insistent également sur le plaisir et la

> ***Le statisticien se doit dès lors de ne pas interroger le phénomène en chaussant les seules lunettes « sanitaires ».***

sociabilité (*Muller et Lombardo, 2023* ; *Croutte et alii, 2019*), certains travaux testant même l'hypothèse d'une contribution significative de la pratique sportive au bien-être des sportifs (*Ruseski et alii, 2014*).

Le statisticien se doit dès lors de ne pas interroger le phénomène en chaussant les seules lunettes « sanitaires » ; il doit également s'interroger sur les manifestations sociologiques et économiques du phénomène sportif, et notamment sur la place prise par le sport dans la vie de nos concitoyens en fonction de leur âge ou de leur catégorie sociale. Comme pour d'autres loisirs, le processus de « distinction » joue ici un rôle essentiel : les catégories favorisées ne pratiquent pas les mêmes sports que les catégories défavorisées, et elles n'ont pas non plus les mêmes conditions de pratique. Les goûts sportifs se construisent en fonction de sa classe sociale et de son genre (*Guérandel et Mardon, 2022*).

▶ L'approche sanitaire de l'activité physique et sportive

Analyser le sport comme activité physique ou comme fait social ne nécessite pas les mêmes outils de suivi statistique et n'aboutit pas aux mêmes conclusions. D'où un paradoxe apparent : la pratique sportive s'est répandue au sein de la société, au point qu'on estime que deux tiers des Français pratiquent régulièrement une activité physique et sportive[1] ; mais, dans le même temps, selon les indicateurs mobilisés en santé publique, les Français sont trop sédentaires. Comment l'expliquer ?

Dans les enquêtes de santé publique, l'activité physique est définie comme « tout mouvement corporel produit par la contraction des muscles squelettiques entraînant une augmentation de la dépense énergétique par rapport à la dépense énergétique de repos » (*Inserm, 2008*). Dès lors, les pratiques sportives ne sont qu'une situation parmi de nombreuses autres qui participent à la dépense énergétique : le travail, les transports, les activités domestiques et les autres loisirs y jouent également un rôle primordial. Selon l'enquête INCA3 menée en 2014-2015[2], les activités hors sport (activités domestiques, de loisirs, transport et travail) représentent 62 % de la sollicitation cardiorespiratoire moyenne par semaine, contre 38 % pour le sport. Ainsi, pour expliquer les problèmes de sédentarité et de manque d'activité physique, de nombreux facteurs jouent un rôle au moins aussi important que l'évolution des pratiques sportives : l'évolution des métiers moins contraignants physiquement, et des pratiques de loisirs non sportives notamment avec la multiplication du temps passé devant les écrans.

1 Au sens qui sera précisé ci-dessous, c'est-à-dire qu'ils ont réalisé au moins 52 « séances » d'activités physiques et sportives au cours de l'année, cf. Didier M., Lefèvre B et Raffin V. « Deux tiers des 15 ans ou plus ont une activité physique ou sportive régulière en 2020 » in *France, Portrait social*, Insee, 2022.

2 Le constat dépend des indicateurs mobilisés. L'inactivité physique est définie par référence à un seuil recommandé (par exemple 30 minutes d'activité physique d'intensité modérée au minimum cinq fois par semaine). Selon l'enquête INCA3 (consommations et habitudes alimentaires) menée en 2014-2015, lorsqu'on considère les seuils de durée de sollicitation cardiorespiratoire (150 min/sem) et de travail musculaire en résistance (TMR) (40 min/sem), la moitié des participants n'atteint pas tous les seuils simultanément (49 %). Les hommes sont plus nombreux que les femmes à atteindre les deux seuils simultanément (63 % contre 34 %) ; 18 % des femmes n'atteignent aucun des seuils contre 6 % des hommes (cf. « Évaluation des risques liés aux niveaux d'activité physique et de sédentarité des adultes de 18 à 64 ans, hors femmes enceintes et ménopausées », Avis de l'Anses, 2022).

Analyser le sport en chaussant ses lunettes de spécialiste de la santé publique oblige donc à l'observer comme une activité physique parmi d'autres, en mesurant son caractère plus ou moins intensif[3]. Un calendrier d'activité très détaillé, sur une journée ou une semaine est ainsi réalisé. À l'inverse, dans une perspective sociologique de suivi de la pratique sportive, intéressant les acteurs du sport, la vision est élargie, pour embrasser l'ensemble des activités physiques et sportives au cours de l'année (ce que l'on qualifie dans la littérature de « portefeuilles de pratiques » (*Michot, 2021*)), en les mettant en (cor)relation avec des indicateurs sociodémographiques.

▶ Établir des « faits stylisés » pour éclairer les acteurs du mouvement sportif et outiller les décideurs publics

Il semble difficile de réunir au sein d'une même source statistique l'ensemble des éléments permettant de décrire à la fois la pratique sportive comme fait social et comme indicateur de santé publique. Ainsi, en France comme à l'étranger, se sont développées deux catégories d'enquêtes ou de dispositifs de suivi statistique malgré la très grande diversité et hétérogénéité des systèmes d'observation chez nos voisins européens et dans le monde ***(encadré 1)***.

> ***Les réponses constituent des leviers de politique publique, pour augmenter la pratique ou mieux réguler le secteur sportif.***

L'enjeu est de disposer d'une vision claire de l'évolution de la pratique sportive, au niveau général mais aussi pour les principaux sports pratiqués, sous forme de messages généraux (« faits stylisés »). Ceux-ci permettent notamment de répondre aux questions qui intéressent les instances, clubs et fédérations sportives, les acteurs de l'économie du sport (grande distribution spécialisée, branches professionnelles), les pouvoirs publics (ministère, Agence nationale du sport, etc.), et les départements STAPS (sciences et techniques des activités physiques et sportives) des universités.

- Quel est le portrait-type du sportif ou de la sportive ?
- Quelles sont les différences de pratique entre femmes et hommes ?
- Comment la pratique sportive évolue-t-elle au cours de la vie ?
- Quels sports sont les plus pratiqués, et par qui ?
- Quels sont les sports émergents ?
- Quelle est la place des clubs et associations dans la pratique sportive ? Quelle est la place pour la pratique non encadrée et celle dans les structures privées ?
- Comment s'articule la pratique sportive avec le travail, la vie quotidienne et les études ?
- Pour quelles raisons les sportifs pratiquent, et pour quels motifs les non sportifs ne pratiquent pas ?

3 En utilisant par exemple des appareils de mesure embarqués, qui permettent de mesurer l'intensité d'une activité physique, établie en équivalent métabolique ou MET (*Metabolic Equivalent of Task*). La dépense énergétique au repos, assis sur une chaise, est proche d'un MET.

Au-delà de l'intérêt pour l'observateur averti de la vie sociale et du secteur sportif, les réponses constituent des leviers de politique publique, pour augmenter la pratique ou mieux réguler le secteur sportif, dont le poids dans l'économie est loin d'être négligeable[4].

▶ Encadré 1. Un cadre de comparaison internationale à construire

La plupart des pays développés se sont dotés d'outils d'observation de la pratique physique et sportive. De ce point de vue, le Royaume-Uni est de loin le plus avancé. Au travers de sa grande enquête bi-annuelle auprès des adultes (*Active Lives Adults*), menée auprès de 180 000 personnes chaque année, et de son enquête annuelle auprès des enfants et des jeunes (*Active Lives Children and Young People*), Sport England évalue de manière très complète, à une maille géographique fine, l'évolution des pratiques sportives, qu'il s'agisse d'activité physique, de suivi des compétitions sportives ou même de bénévolat au sein des clubs.

Pour autant, contrairement à de nombreux autres domaines d'analyse statistique, aucune nomenclature internationale commune n'a été construite. Sur le plan sanitaire, l'Organisation mondiale de la Santé a certes publié des recommandations mondiales, qui préconisent au moins 150 minutes d'activité physique d'intensité modérée, ou 75 minutes d'intensité soutenue par semaine pour les adultes. Celles-ci font l'objet d'un suivi statistique irrégulier et sans collecte uniformisée, les dernières données disponibles datant de 2016.

La pratique sportive à proprement parler, ne fait pas l'objet d'un suivi, même au niveau européen. En effet, le sport ne constitue pas une compétence de l'Union européenne – Eurostat travaille peu sur la pratique sportive – et il n'existe pas d'organisation mondiale du sport, à l'instar du Bureau international du Travail qui a défini les contours du chômage. En conséquence, plusieurs publications mettent logiquement en avant les difficultés des comparaisons internationales*.

Ainsi, à l'heure actuelle, la seule source régulière permettant des comparaisons entre pays européens est l'Eurobaromètre spécial piloté par la Commission européenne et portant sur le sport et l'activité physique, dont la dernière édition date de mi-2022 (et la précédente de 2017) ; il s'agit d'une enquête probabiliste sur la base de la sélection aléatoire d'adresses, auprès d'échantillons relativement faibles dans chaque pays (autour de 1 000 pour la France par exemple). L'interrogation porte sur une question unique : « À quelle fréquence faites-vous du sport ou de l'exercice physique ?** », en proposant une échelle de réponse en cinq points, de « 5 fois par semaine ou plus » à « Jamais ». Cette enquête repose ainsi largement sur ce que les enquêtés considèrent comme relevant du sport et ne permet pas de détailler le type de sport pratiqué. L'enquête santé européenne, pilotée par Eurostat, comporte également quelques questions sur la pratique d'activités physiques et sportives.

* *Par exemple, un essai de meta-analyse note : "Results should be used with caution, as some studies have a high risk of bias, which may not necessarily be representative of the population for a given region" (cf. Hulteen, R. M., Smith, J. J., Morgan, P. J., Barnett, L. M., Hallal, P. C., Colyvas, K., & Lubans, D. R. (2017). Global participation in sport and leisure-time physical activities: A systematic review and meta-analysis. Preventive medicine, 95).*

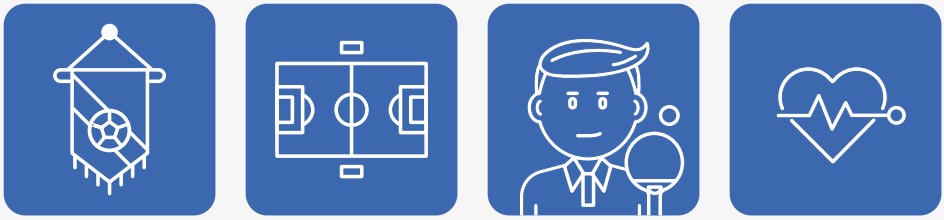
** *En précisant : « Par « faire de l'exercice physique », nous entendons tous les types d'activités physiques que vous pouvez pratiquer dans un contexte sportif ou une infrastructure sportive, comme nager, vous entraîner dans un centre de fitness ou un club sportif, courir dans un parc, etc. ».*

▶ Peut-on se passer d'une enquête ?

Les personnes pratiquant un sport au sein d'un club affilié à une fédération souscrivent une licence sportive, qui leur confère notamment une assurance, et permet le financement des fédérations. À l'heure où l'on enjoint la statistique publique de s'appuyer le plus possible sur les données administratives peu coûteuses et déjà disponibles, et le moins possible

4 En 2020, les dépenses sportives des ménages représentent 0,8 point de PIB, et les dépenses publiques en faveur du sport 0,6 point (*Dietsch, 2022*).

▶ Figure 1 - Les principales sources statistiques sur la pratique sportive



CLUBS SPORTIFS

Recensement des clubs et licences sportifs de l'Injep (annuel)

ÉQUIPEMENTS SPORTIFS

Recensement des équipements sportifs de la Direction des sports, intégré à la Base permanente des équipements de l'Insee (annuel)

PRATIQUE SPORTIVE

Enquête nationale sur les pratiques physiques et sportives (ENPPS, décennal)

Baromètre national des pratiques sportives (bisannuel)

ACTIVITÉS PHYSIQUES

Enquête européenne de santé de la Drees (EHIS, quinquennal)

Baromètre de Santé publique France (irrégulier)

Injep : Institut national de la jeunesse et de l'éducation populaire.
ENPPS : Enquête nationale sur les pratiques physiques et sportives.
EHIS : Enquête européenne de santé (European Health Interview Survey).

sur des données d'enquête réputées onéreuses et chronophages pour les répondants, ne pourrait-on pas suivre la pratique sportive en utilisant le recensement des licences délivrées dans les clubs ? ***(figure 1)***.

Ce recensement est mené auprès des fédérations par l'Injep et s'avère en effet précieux. Tout d'abord, il permet un suivi annuel de la pratique sportive, pour les principaux sports mais aussi pour des activités moins répandues, quand elles sont groupées au sein d'une fédération[5]. Ensuite, ce recensement fournit des informations très précises pour un maillage territorial communal, outillant les acteurs locaux dans le diagnostic des forces et faiblesses de leur territoire en matière sportive. Enfin, il permet également aux fédérations de comparer leurs adhérents, s'agissant de leur féminisation ou encore de la proportion de leurs adhérents vivant en quartier prioritaire de la politique de la ville (QPV) ou en zone de revitalisation rurale. Ces données ont été mobilisées dans le cadre des Conférences régionales du sport, lancées récemment par l'Agence nationale du sport.

5 Pour piquer la curiosité des lectrices et lecteurs, évoquons par exemple le twirling bâton (discipline qui allie la danse, la gymnastique, le théâtre et le maniement du bâton), pratiqué par une personne sur mille environ, ou le pulka (un traîneau utilisé pour la pratique sportive ou le transport), pratiqué par trente fois moins de personnes.

Les limites des données administratives sur les licences sportives

Pour autant, comme toutes données administratives collectées pour d'autres motifs que la connaissance statistique[6], le recensement des licences présente plusieurs fragilités. Tout d'abord, les clubs ne disposent que de peu de variables sociodémographiques sur leurs adhérents, au-delà de leur sexe, de leur âge et de leur lieu de résidence : aussi, le recensement ne permet pas d'interroger les différences de pratique en fonction de la catégorie sociale ou de l'état de santé des personnes.

Ensuite, les personnes licenciées dans deux fédérations différentes sont comptabilisées deux fois, et le nombre de fédérations agréées par le ministère des Sports a augmenté au cours des dernières décennies : il est passé de 111 en 2000 à 119 actuellement, conduisant à surestimer l'essor de la pratique.

Enfin, l'évolution du nombre de licences délivrées est affectée par les changements dans les pratiques « administratives » des fédérations sportives. Celles-ci modifient parfois significativement leur politique vis-à-vis des personnes participant aux activités des clubs affiliés, en requérant par exemple que l'ensemble des participants, y compris dans le cadre d'une pratique non compétitive, prennent une licence. Cela explique des bonds parfois spectaculaires du nombre de licences : la fédération française de canoë-kayak a par exemple quasiment doublé ses licences entre 2016 et 2019 (de 43 000 à 78 000), en créant une licence « pagaie blanche », permettant à de nouveaux adhérents débutants de pratiquer, et en comptabilisant les titres délivrés à l'occasion d'une pratique occasionnelle, par exemple pendant les vacances, comme des licences fédérales.

> ***La pratique sportive au sein d'un club affilié à une fédération sportive est minoritaire.***

Même sans tenir compte de ces quelques limites méthodologiques, qui ne remettent pas en cause la nécessité de recenser les licences, cela reste insuffisant pour suivre la pratique sportive. En effet, s'il permet de donner une image relativement fidèle de la pratique sportive des enfants et adolescents, souvent encadrée, à l'école ou au sein d'un club sportif (*Caille, 2020*), le recensement des licences est fortement trompeur s'agissant des pratiques sportives des jeunes et des adultes. En effet, la pratique sportive au sein d'un club affilié à une fédération sportive est minoritaire. Un seul chiffre suffit à s'en convaincre : en 2020, selon l'enquête nationale sur les pratiques physiques et sportives, seuls trois sportifs réguliers sur dix détiennent une licence d'un club fédéral. Parmi les 15 ans et plus, le palmarès des sports les plus pratiqués est par ailleurs très différent, selon que l'on considère le nombre de licences – football, tennis et équitation se retrouvent sur le podium - ou la pratique sportive régulière – fitness, running, vélo et natation occupent alors les premières places.

6 Plusieurs articles ont traité des enjeux et difficultés liés à l'utilisation de données administratives pour les statistiques publiques (cf. par exemple Hoffman E., "*We must use administrative data for official statistics – but how should we use them?*", Statistical Journal of the United Nations Economic Commission for Europe, 1995, vol. 12, n°1, ou, plus récemment, Bakker B. et Daas P., "*Methodological challenges of register-based research*", Statistica Neerlandica 66.1, 2012).

▶ À la recherche d'une définition de la pratique sportive

Une enquête régulière auprès des ménages est donc nécessaire pour suivre et mesurer la pratique sportive. Une telle enquête suppose cependant de définir la pratique sportive, ce qui, pour certains auteurs, « *semble relever d'un pari intenable, tant les pratiques sont bigarrées et les frontières incertaines* » (*Bromberger, 1995*).

Les débats portent tant sur les activités incluses que sur les conditions de pratique. Chacun s'accorde à considérer comme une séance de sport une activité physique intense, encadrée par des règles, pratiquée régulièrement dans un cadre de loisirs. Mais dès lors que l'on s'éloigne de cette définition principale, des problèmes de frontière entre le sportif et le non sportif apparaissent. Une promenade dominicale en famille constitue-t-elle une activité physique et sportive ? De même, une séance de baignade dans un lac ? Comment classer certaines activités nécessitant une faible dépense énergétique, comme à l'extrême les sports dits cérébraux ou l'*e-sport*, qui empruntent au modèle sportif son caractère compétitif et ses séances d'entraînement[7] ? À l'inverse, les déplacements domicile-travail ou domicile-études à vélo ou en courant, qui occasionnent une dépense énergétique importante mais n'ont à l'évidence pas le caractère d'activité de loisirs, doivent-ils être comptabilisés ?

Ces frontières sont d'autant plus délicates à trancher que les réponses sont situées historiquement et socialement. Tout d'abord, en raison de la relative nouveauté du fait sportif, réservé à une minorité de la population jusqu'au milieu du 20e siècle, et dont le périmètre a fortement évolué depuis lors au cours de sa démocratisation. Les nouvelles activités de loisirs sportifs comme l'accrobranche ou la *via ferrata* sont une nouvelle manifestation de cette porosité, tout comme l'*e-sport* déjà cité, ou encore l'activité physique adaptée dans le cadre du sport-santé[8]. Ensuite, même à un moment précis, la classification d'une activité comme un sport dépend de l'observateur : pour une personne du 4e âge[9], réaliser des exercices simples d'assouplissement ou faire le tour d'un parc en marchant correspondra à une activité sportive, tandis que, pour un triathlète momentanément blessé, une interruption de ses entraînements sera vécue comme un vide sportif, même s'il continue à effectuer chaque matin des exercices de musculation exigeants et difficiles.

> ***Il est ainsi problématique de s'appuyer uniquement sur les activités désignées spontanément comme sportives.***

Il est ainsi problématique de s'appuyer uniquement sur les activités désignées spontanément comme sportives pour établir un taux de sportivité au sein d'une population, comme lorsque l'on mobilise des enquêtes généralistes pour évaluer la pratique sportive à l'aide d'une ou deux questions[10]. La marche constitue ici un exemple éclairant : selon le Baromètre de Santé publique France,

7 Une seule fédération de sport dit cérébral a reçu l'agrément du ministère des Sports : la Fédération française des échecs, en 2000. Les fédérations de bridge (FFB), de poker (FFP), de jeu de dames (FFJD) et de jeux vidéo en réseaux (FFJVR) ont sollicité cet agrément sans succès.

8 Le sport-santé recouvre la pratique d'activités physiques ou sportives qui contribuent à la santé et au bien-être du pratiquant. À ce titre, les activités sport-santé doivent obligatoirement être adaptées au public et encadrées par des éducateurs formés spécifiquement au sport-santé.

9 Le quatrième âge désigne généralement les personnes âgées de 75 ans et plus.

10 Ces dernières s'appuient sur une déclaration spontanée de la fréquence de pratique, sans préciser la nature de celle-ci. Par exemple, l'enquête statistique sur les ressources et conditions de vie (SRCV) ou l'Eurobaromètre.

cette activité est de plus en plus souvent considérée comme sportive. Ainsi, « *en 2017, 60 % des adultes ont déclaré avoir pratiqué un sport au cours des 7 derniers jours, contre 37 % en 2000* » (*Galey et alii, 2020*). Cette hausse spectaculaire s'explique en grande partie par la « *meilleure reconnaissance de la marche de loisirs comme une activité sportive à part entière en 2017* », qui a engendré un doublement du nombre de répondants ayant déclaré la marche au titre de leurs activités sportives.

▶ Le parti pris des enquêtes nationales sur les pratiques physiques et sportives (ENPPS) : établir une cartographie la plus large possible des pratiques sportives

Pour s'abstraire de ces difficultés tout en évitant l'écueil d'une définition figée *a priori*, l'ENPPS repose sur un questionnement en deux temps, spontané puis guidé par la classification des activités en 70 sous-catégories. Cette démarche permet aux enquêtés d'aller au-delà des seules pratiques proposées en ajoutant leurs propres activités physiques et sportives. Au final, près de 530 activités sont recensées, des plus fréquentes aux plus confidentielles.

Dans un premier temps, les personnes déclarant pratiquer une activité physique ou sportive listent l'ensemble des disciplines exercées (déclaration spontanée d'activités). Dans un second temps, pour éviter les oublis, une liste de plus de 70 disciplines est proposée aux enquêtés, qu'ils aient auparavant déclaré faire du sport ou non (relance). Lors des entretiens par téléphone, les disciplines sportives déclarées spontanément sont saisies telles qu'énoncées par le répondant. Des précisions sont éventuellement demandées pour certaines disciplines (marche, vélo, randonnée, natation, etc.). Deux questions permettent ensuite de bien différencier les pratiques de loisirs des pratiques utilitaires, c'est-à-dire des activités dont la finalité est de se déplacer (par exemple déplacement entre le domicile et le travail à vélo ou à pied).

Pour chacune des activités physiques et sportives citées, un ensemble de questions est posé sur les conditions de pratique, à savoir la périodicité, la fréquence, les lieux de pratique, la sociabilité (pratique individuelle ou avec d'autres personnes), l'âge de début, le niveau perçu de pratique, le mode de pratique (fréquentation d'une structure marchande ou non marchande, pratique en autonomie ou encadrée), la détention d'une licence sportive ou la participation à des compétitions ou des manifestations sportives.

▶ Des sports souvent cités spontanément par les répondants, et d'autres beaucoup moins fréquemment

Lors de ce questionnement en deux temps, suivi de demandes de précisions sur les pratiques et les modalités de pratique, les répondants oublient spontanément de nombreuses séances de sport qu'ils ont pourtant réalisées au cours de l'année écoulée, et s'en souviennent quand on leur rafraîchit la mémoire à l'aide d'une liste détaillée d'activités (à l'image de la notoriété dite spontanée, toujours nettement plus faible que la notoriété assistée[11]).

11 Dans les enquêtes de marketing ou sur l'accès aux droits et la connaissance des aides sociales, la notoriété spontanée correspond au pourcentage de personnes capables de citer spontanément le nom d'une marque ou d'un programme de politique publique. La notoriété assistée désigne la proportion de personnes déclarant les connaître une fois qu'ils ont été listés dans l'enquête.

Ainsi, pour les sports les plus répandus, parmi ceux ayant pratiqué au moins une fois dans l'année, une majorité ne le déclare pas spontanément : en 2020, dans l'ENPPS, les taux varient entre 45 % et 55 % pour le football, le basketball, le tennis, l'escalade *indoor*, le vélo de course, le VTT ou la natation. Ils sont de moins d'un tiers pour la plupart des sports pratiqués le plus souvent occasionnellement (badminton, tennis de table) ou lors des vacances (ski alpin, randonnée pédestre, voile, beach-volley, etc.), et à des taux moindres, de l'ordre de 10 %, pour des activités de loisirs à la frontière de la conception habituelle du sport (accrobranche, luge, karting, etc.), dont certaines constituent pourtant un sport fédéral (pétanque, patin à glace, canoë-kayak). Seules les activités immédiatement identifiées comme sportives et le plus souvent pratiquées très régulièrement, comme la course à pied ou la gymnastique de forme ou d'entretien, sont déclarées spontanément chez plus des deux tiers des pratiquants.

Cette difficulté pour les enquêtés à définir le sport et à se souvenir de leurs séances les moins régulières explique que la mesure de la pratique sportive soit très dépendante des conditions de collecte (questionnaire, interrogation en ligne, par téléphone ou en face-à-face, etc.). Cela est d'autant plus vrai lorsqu'on mobilise une définition extensive de la pratique sportive, très sensible aux conditions de collecte (principe : « quand on cherche, on trouve ! »), mais beaucoup moins pour la pratique régulière, moins sensible aux conditions de collecte. Aussi, le Baromètre national des pratiques sportives, mis en place par l'Injep et le ministère des Sports en complément des enquêtes décennales pour disposer d'un suivi régulier de l'évolution de la pratique sportive, donne des taux de pratique différents ***(encadré 2)***.

▶ Encadré 2. Des enquêtes plus légères pour répondre à des questions plus conjoncturelles sur l'évolution de la pratique sportive

En 2018, l'Injep et le ministère des Sports ont souhaité se doter d'un outil de suivi plus léger que l'ENPPS, avec un Baromètre national des pratiques sportives, conduit tous les deux ans. Ces deux enquêtes sont complémentaires : l'ENPPS étudie de manière structurelle la pratique physique et sportive, tandis que le Baromètre assure un suivi régulier, pour évaluer, par exemple, l'atteinte de l'objectif de hausse de 3 millions du nombre de pratiquants d'APS.

Le Baromètre national des pratiques sportives est une enquête *Web* par quotas auprès de 4 000 répondants, contrairement à l'ENPPS qui repose sur une base de sondage et un échantillon probabiliste. Par ailleurs, le format plus léger du Baromètre ne propose pas un questionnement exhaustif comme dans l'ENPPS, ce qui conduit à sous-estimer certaines pratiques, lorsqu'elles sont peu fréquentes ou peu communes. Ainsi, le taux de pratique physique et sportive estimé dans l'ENPPS est plus élevé que dans le Baromètre, qui est lui-même plus élevé que celui des enquêtes généralistes non dédiées au sport. Ces dernières s'appuient sur une déclaration spontanée de la fréquence de pratique, sans préciser la nature de celle-ci, ni vérifier la pratique au cours des 12 derniers mois à l'aide d'une liste précise d'activités.

Cette sensibilité de la mesure de la pratique sportive aux conditions de collecte et d'enquête est connue et pose des difficultés pour effectuer des comparaisons internationales, ou dans le temps, lorsque les protocoles d'enquête ont changé (comme au Royaume-Uni en 2016). Néanmoins, au-delà des niveaux de pratique, ces enquêtes sont intéressantes lorsque l'on compare la pratique entre groupes sociaux, entre régions ou en fonction de l'activité pratiquée.

Des choix structurants pour l'enquête

La nomenclature présentée aux enquêtés, son degré de détails mais aussi sa structuration ne sont pas neutres dans les réponses récoltées par les statisticiens. Dans l'ENPPS, un soin particulier a été apporté à la constitution de cette nomenclature de pratiques, en veillant à être le plus près possible du langage commun, et sans volonté de correspondre aux classifications plus élaborées : en d'autres termes, il s'agit d'une nomenclature inspirée des rayonnages d'un magasin d'articles de sport plutôt que de ceux d'une bibliothèque universitaire ! Ainsi, comme souvent en matière de statistiques publiques, l'enjeu de la construction de la nomenclature est crucial, et on trouve souvent des enjeux similaires dans d'autres champs thématiques, comme les activités culturelles.

Un autre choix structurant correspond à la période de référence. L'ENPPS, comme d'autres enquêtes comparables dans d'autres pays, retient une période de 12 mois, indispensable pour capter l'ensemble des pratiques de vacances, ou encore les pratiques saisonnières (on ne pratique pas les mêmes sports en été ou en hiver), mais qui fait peser un risque de biais de mémoire. Alternativement, les enquêtes de santé publique interrogent les répondants sur leur pratique au cours d'une période très courte (jour ou semaine de référence), afin de pouvoir ensuite mesurer très précisément le nombre de minutes consacrées à telle ou telle activité physique, dont les activités sportives.

Ce choix de période de référence de 12 mois privilégie le concept de « séance » d'activité physique et sportive à celui de « nombre d'heures de pratique », dans la mesure où il est évidemment exclu de demander aux enquêtés de reconstituer sur une période aussi longue la durée cumulée de l'ensemble de leurs séances d'activités physiques et sportives (APS). Pour autant, les répondants sont invités à renseigner leur durée hebdomadaire habituelle d'exercice pour les sports pratiqués.

Choisir un indicateur de référence : la pratique régulière et son « halo »

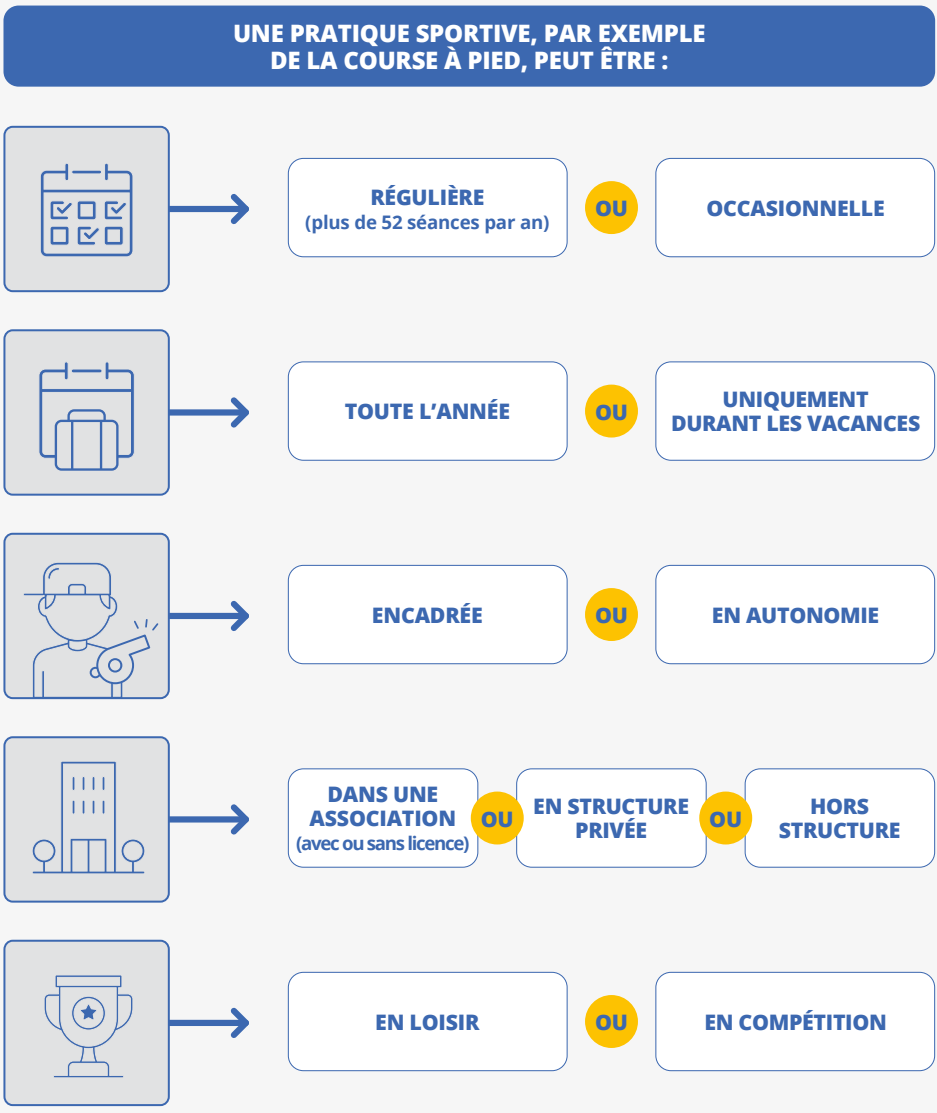
Dès lors, outillé de cette technologie d'enquête riche et complexe, quel indicateur principal retenir pour déterminer la part des répondants considérés comme sportifs ?

Pour répondre à cette question, l'attention des utilisateurs et utilisatrices des statistiques se focalise sur un concept de référence, mais restreint alors la focale à un seul indicateur, potentiellement réducteur. Le choix retenu dans la dernière édition de l'ENPPS est de mettre en avant la pratique sportive dite régulière, c'est-à-dire la part des répondants déclarant au moins 52 séances d'activités physiques et sportives durant l'année[12], en excluant les pratiques dites utilitaires (dans le cadre de déplacements domicile-travail notamment), ainsi que les activités de promenade, baignade et relaxation. Environ deux tiers des 15 ans et plus sont considérés comme sportifs réguliers selon cette définition, même s'ils ne pratiquent pas forcément leur APS principale toutes les semaines.

12 Les personnes déclarant réaliser une activité chaque semaine se voient attribuer 52 séances dans l'année, même s'ils ont pu interrompre leur pratique durant certaines périodes, par exemple en raison des vacances ou d'une blessure.

Pour autant, adopter un système de conventions n'empêche pas, bien au contraire, de valoriser la diversité et la variabilité des pratiques ***(figure 2)***, comme on peut le faire en matière de statistiques sur le marché du travail avec le chômage et le halo du chômage. Il est ainsi possible d'observer les différentes formes que prend l'activité physique et sportive en France. Puis, en croisant avec des questions précises sur le contexte, la fréquence, et les modalités de pratiques (en club ou non, etc.), de les qualifier et ainsi de constituer tout un éventail de la pratique sportive des Français, allant du « noyau dur » des personnes qui pratiquent plusieurs fois par semaine au « halo » des pratiquants occasionnels lors de leurs vacances (*Didier et alii, 2022*).

▶ Figure 2 - Typologie de la pratique sportive



Deux pistes peuvent potentiellement améliorer le système d'observation des pratiques sportives en France : exploiter les nouvelles données issues des pratiques sportives connectées et réaliser des diagnostics territoriaux reliant la pratique sportive locale et le maillage des clubs et équipements sportifs.

▶ Explorer le potentiel des « *big data* » et des pratiques connectées

Les objets connectés, à commencer par les smartphones que la grande majorité des adultes possèdent, ou les montres connectées, constituent potentiellement des outils particulièrement utiles pour quantifier la pratique physique et sportive.

Ils sont déjà largement mobilisés dans le sport de haut niveau, dans une perspective physiologique d'amélioration des performances des athlètes. À titre d'illustration, l'Institut national du sport, de l'expertise et de la performance (Insep) déploie un « *Sport data hub* » dans le cadre de la préparation des Jeux olympiques et paralympiques de Paris. Ces travaux s'appuient le plus souvent sur des capteurs spécifiques très précis que les sportifs de haut niveau acceptent de porter durant leur pratique. Ils ne sont donc pas reproductibles dans la population générale.

Toutefois, dans la mesure où les objets connectés sont capables de suivre les déplacements des personnes, ils pourraient s'avérer précieux dans le suivi du niveau d'activité physique ou de sédentarité de la population, par exemple pour connaître la part des Français suivant la recommandation de santé publique de réaliser au moins 10 000 pas par jour.

Dans un autre registre, les objets connectés sont prometteurs pour mesurer le nombre de spectateurs d'une manifestation sportive, notamment lorsqu'il s'agit d'événements en plein air ou en accès libre sans billetterie, comme une étape du Tour de France.

Enfin, dans les sports où les applications connectées sont fréquemment utilisées, comme la course à pied ou le vélo (sports dits « de nature »), des statistiques intéressantes peuvent être produites sur les tracés utilisés par les sportifs ***(figure 3)***, sur leurs horaires et durées de pratique, ainsi que sur le nombre de personnes présentes à la sortie du parcours. Cela a conduit le ministère des Sports à créer une plateforme (*Outdoorvision*) visant à réunir les informations issues de plusieurs applications, afin d'aider les collectivités locales dans leur politique d'aménagement et de protection des espaces naturels.

Pour autant, l'utilisation de ces données se heurte à plusieurs écueils. Tout d'abord, la mobilisation des seules données issues du bornage des téléphones portables sur les antennes relais, si elles ont l'avantage d'être exploitables pour l'ensemble des possesseurs d'un téléphone, ne permet pas une localisation suffisamment précise pour pouvoir établir des tracés, ou encore comptabiliser le nombre de participants à une manifestation sportive. Les données d'applications sportives dédiées sont, quant à elles, beaucoup plus précises, grâce à l'activation d'un suivi GPS, mais elles souffrent d'un biais de sélection, la pratique sportive connectée étant encore largement minoritaire, même dans les sports de nature. Ainsi, on peut s'intéresser à la hausse du nombre d'utilisateurs de telle ou telle application et à l'évolution de leurs pratiques, mais la difficulté est qu'il s'agit d'une population non représentative, évolutive d'une année à l'autre avec la popularité croissante des applications connectées.

▶ Figure 3 - Parcours des sportifs à Paris

Source : Nathan Yau, Flowing Data, 2014, where people run in major cities? (cf. https://flowingdata.com/2014/02/05/where-people-run/)

« L'année sportive », un retour en statistiques sur l'année passée publiées par l'entreprise Strava[13], illustre ces écueils. Quand on apprend sur Strava que « *le pourcentage d'athlètes ayant couru un marathon a presque doublé par rapport à 2021* » (et même triplé en France !), que peut-on en déduire ? Qu'il y a une hausse considérable du nombre de marathoniens (sans doute inspirés par l'exemple du britannique Gary McKee ayant couru un marathon par jour en 2022 !) ? Ou, plus probablement, que de plus en plus de coureurs de fond enregistrent leurs performances sur l'application Strava ?

À l'heure actuelle, les données issues d'applications connectées ne permettent donc pas un suivi quantitatif de la pratique sportive, y compris dans des sports où les applications sont déjà assez répandues. Ces applications peuvent cependant témoigner de phénomènes qualitatifs intéressants, comme récemment la hausse du nombre de séances de courses à pied ayant lieu en début d'après-midi plutôt qu'en soirée ou pendant la pause méridienne, probablement en lien avec le développement du télétravail[14].

13 Strava est un site internet et une application mobile utilisés pour enregistrer des activités sportives *via* GPS.
14 En 2021, les membres de l'application Running Heroes courent davantage entre 14 h et 18 h en semaine : 14 % contre 9 % en 2019 (L'Observatoire du Running 2022, Sport Heroes et UNION sport et cycle, 2022).

▶ Relier la pratique sportive et les équipements et clubs sportifs locaux

Les acteurs du monde du sport, notamment les collectivités territoriales, s'interrogent régulièrement sur le lien entre l'offre (équipements, clubs) et la pratique sportive locale. Des études qui feraient le lien entre de nouveaux équipements et l'évolution subséquente de la pratique sportive seraient particulièrement utiles. En France, l'existence d'un recensement des licences sportives, même s'il ne couvre pas toutes les formes de pratique sportive, loin de là (cf. *supra*), permettrait d'effectuer ce type d'analyse, par exemple autour du plan de 5 000 nouveaux équipements sportifs de proximité déployé récemment par l'Agence nationale du sport.

> ***Les données issues d'applications connectées ne permettent donc pas un suivi quantitatif de la pratique sportive, y compris dans des sports où les applications sont déjà assez répandues.***

Des nouvelles études seraient également précieuses pour mieux caractériser l'offre sportive territoriale, au-delà de l'indicateur habituel de densité du nombre d'équipements par habitant, qui présente à tort une vision très favorable des territoires ruraux peu denses, alors que les équipements sont souvent très éloignés du domicile des potentiels pratiquants ; ou de l'indicateur de densité du nombre d'équipements par km^2, qui présente à l'inverse les zones urbaines très denses sous un angle trop favorable. Un indicateur idéal tiendrait compte à la fois du temps d'accès aux équipements et de leur diversité, voire de l'âge de la population, et permettrait ainsi d'identifier des territoires sous-dotés en équipements et clubs sportifs.

Pour conclure, de nombreuses demandes sont adressées au service statistique ministériel du sport en cette période de préparation des Jeux olympiques et paralympiques de Paris, dont l'héritage doit générer à compter de 2024 une hausse durable de la pratique sportive, y compris pour les personnes actuellement éloignées de cette pratique. Le système d'observation statistique est particulièrement mobilisé pour quantifier la sportivité des personnes en situation de handicap ou celle des seniors. Ces demandes créent de nouveaux défis à la statistique publique, qui gagnerait à bénéficier d'un éclairage et d'échanges méthodologiques internationaux à développer dans le champ sportif.

▶ Bibliographie

- ANSES, 2022. Évaluation des risques liés aux niveaux d'activité physique et de sédentarité des adultes de 18 à 64 ans, hors femmes enceintes et ménopausées. [en ligne]. 18 janvier 2022. [Consulté le 1er août 2023]. Disponible à l'adresse : *https://www.anses.fr/fr/system/files/NUT2017SA0064-b.pdf.*

- BAKKER, Bart F. M. et DAAS, Piet J. H., 2012. Methodological challenges of register-based research. In : *Statistica Neerlandica*, Vol. 66, n°1, pp. 2-7. [en ligne]. [Consulté le 1er août 2023]. Disponible à l'adresse : *http://www.pietdaas.nl/beta/pubs/pubs/Bakker_Daas_Stat_Neerlandica.pdf.*

- BROMBERGER, Christian, 1995. De quoi parlent les sports ? In : *Terrain*, n°25, pp. 5-12. [en ligne]. [Consulté le 1er août 2023]. Disponible à l'adresse : *https://journals.openedition.org/terrain/2837.*

- CAILLE, Jean-Paul, 2020. Les pratiques sportives des collégiens sont très liées au rapport au sport de leurs parents et à leurs vacances d'été. In : *France, Portrait social*, Insee. 3 décembre 2020. [en ligne]. [Consulté le 1er août 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/4797662?sommaire=4928952*.

- CARLAC'H, Dominique et LE FUR, Marie-Amélie, 2023. Développer le parasport en France : de la singularité à l'universalité, une opportunité pour toutes et tous. In : *Rapport du CESE*. [en ligne]. [Consulté le 1er août 2023]. Disponible à l'adresse : *https://www.lecese.fr/travaux-publies/developper-le-parasport-en-france-de-la-singularite-luniversalite-une-opportunite-pour-toutes-et-tous.*

- CROUTTE, Patricia, MÜLLER, Jörg et DIETSCH, Bruno, 2019. La santé et le bien-être, premiers ressorts des pratiques sportives. In : *INJEP Analyses et synthèses* n°20. 17 janvier 2019. [en ligne]. [Consulté le 1er août 2023]. Disponible à l'adresse : *https://injep.fr/publication/la-sante-et-le-bien-etre-premiers-ressorts-des-pratiques-sportives/.*

- DIDIER, Mathilde, LEFEVRE, Brice et RAFFIN Valérie, 2022. Deux tiers des 15 ans ou plus ont une activité physique ou sportive régulière en 2020. In : *France, Portrait social*, Insee. 22 novembre 2022. [en ligne]. [Consulté le 1er août 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/6535289?sommaire=6535307.*

- DIETSCH, Bruno, 2022. Poids économique du sport en 2020. In : *Fiche-repères n° 52, INJEP*. 23 octobre 2020. [en ligne]. [Consulté le 1er août 2023]. Disponible à l'adresse : *https://injep.fr/publication/le-poids-economique-du-sport/.*

- GALEY, Catherine, VERDOT, Charlotte, SALANAVE, Benoît, CARUSO, Anthony, PELÉ, Tino, LEMONNIER, Fabienne et, RICHARD Jean-Baptiste, 2020. La pratique sportive chez les adultes en France en 2017 et évolutions depuis 2000 : résultats du Baromètre de Santé publique France. In : *Santé publique France*. 30 octobre 2020. [en ligne]. [Consulté le 1er août 2023]. Disponible à l'adresse : *https://www.santepubliquefrance.fr/determinants-de-sante/nutrition-et-activite-physique/documents/enquetes-etudes/la-pratique-sportive-chez-les-adultes-en-france-en-2017-et-evolutions-depuis-2000-resultats-du-barometre-de-sante-publique-france.*

- GUÉRANDEL, Carine et MARDON, Aurélia, 2022. Construction des féminités et des masculinités juvéniles dans le sport, In : *Agora débats/jeunesses* 2022/1, n° 90, Presses de sciences Po. [en ligne]. [Consulté le 1er août 2023]. Disponible à l'adresse : *https://injep.fr/publication/construction-des-feminites-et-des-masculinites-juveniles-dans-le-sport/.*

- HOFFMAN, Eivind, 1995. We must use administrative data for official statistics – but how should we use them? In : *Statistical Journal of the United Nations Economic Commission for Europe*, vol. 12, n°1, 1995. [en ligne]. [Consulté le 1er août 2023]. Disponible à l'adresse : *https://web-archive.oecd.org/2012-06-15/146108-36237589.pdf*.

- HULTEEN, Ryan M., SMITH, Jordan J., MORGAN, Philip J., BARNETT, Lisa M., HALLAL, Pedro C., COLYVAS, Kim et LUBANS, David R., 2017. Global participation in sport and leisure-time physical activities: A systematic review and meta-analysis. In : *Preventive medicine 95*. [en ligne]. [Consulté le 1er août 2023]. Disponible à l'adresse : *https://epidemio-ufpel.org.br/uploads/artigos/ypmed.pdf*.

- HURTIS, Muriel et SAUVAGEOT, Françoise, 2018. L'accès du plus grand nombre à la pratique d'activités physiques et sportives. In : *site du CESE*. [en ligne]. [Consulté le 1er août 2023]. Disponible à l'adresse : *https://www.lecese.fr/travaux-publies/lacces-du-plus-grand-nombre-la-pratique-dactivites-physiques-et-sportives*.

- INSERM (dir), 2008. Activité physique : Contextes et effets sur la santé. In : *Rapport d'expertise collective de l'Inserm*. [en ligne]. [Consulté le 1er août 2023]. Disponible à l'adresse : *https://www.ipubli.inserm.fr/handle/10608/80*.

- MICHOT, Thierry, 2021. La pratique d'activités physiques et sportives en France. Revue de la littérature et des données statistiques, 2021/n° 15. [en ligne]. In : *INJEP Notes & rapports/Revue de littérature*. [en ligne]. [Consulté le 1er août 2023]. Disponible à l'adresse : *https://injep.fr/publication/la-pratique-dactivites-physiques-et-sportives-en-france/*.

- MÜLLER, Jörg et LOMBARDO, Philippe, 2023. Comment l'après-Covid stimule l'élan sportif des Français. In : *INJEP Analyses et synthèses n°65*. [en ligne]. [Consulté le 1er août 2023]. Disponible à l'adresse : *https://injep.fr/publication/comment-lapres-covid-stimule-lelan-sportif-des-francais/*.

- PIOZIN, Éric, SÈVE, Carole et LEROY, Édouard, 2023. Le développement de la pratique sportive étudiante. In : *Rapport de l'IGESR, n°21-22 353 A*. [en ligne]. [Consulté le 1er août 2023]. Disponible à l'adresse : *https://medias.vie-publique.fr/data_storage_s3/rapport/pdf/288115.pdf*.

- RUSESKI, Jane E., HUMPHREYS, Brad R., HALLMAN, Kirstin, WICKER, Pamela et BREUER Christoph, 2014. Sport participation and subjective well-being: Instrumental variable results from German survey data. In : *Journal of Physical Activity and Health*, 11(2). [en ligne]. [Consulté le 1er août 2023]. Disponible à l'adresse : *https://www.researchgate.net/publication/235384751_Sport_Participation_and_Subjective_Well-Being_Instrumental_Variable_Results_From_German_Survey_Data*.

- SPORT ENGLAND, 2022. Active Lives Adult Survey. In : *site de Sport England*. November 2021-22 Report. [en ligne]. [Consulté le 1er août 2023]. Disponible à l'adresse : *https://www.sportengland.org/research-and-data/data/active-lives*.

- SPORT HEROES et UNION SPORT ET CYCLE, 2023. L'Observatoire du Running 2023 : la pratique du running se diversifie. In : *site de Union sport et cycle*. [en ligne]. [Consulté le 1er août 2023]. Disponible à l'adresse : *https://unionsportcycle.com/2023-04-03/observatoire-du-running-2023-la-pratique-du-running-se-diversifie*.

FINESS, le répertoire des établissements de santé

Johanna Bensoussan*, Joël Bizingre**, Nathalie Courvalin***

La santé fait intervenir une multitude d'acteurs au sein de différentes structures, constituées principalement par des personnes morales et leurs établissements, répartis sur le territoire. Un référencement complet de ces organisations est déterminant pour identifier de façon opposable les structures, cartographier l'offre territoriale de santé et objectiver les manques, informer les professionnels et le grand public, financer les établissements et les doter d'une identité numérique.

Depuis plus de 40 ans, le Fichier National des Établissements Sanitaires et Sociaux (FINESS) répond à ce besoin. Les structures y sont décrites à partir de décrets et d'arrêtés. Un numéro FINESS est attribué à chaque personne morale et établissement. Ce numéro est adossé aux numéros SIREN/SIRET du répertoire Sirene de l'Insee.

Ce fichier est au cœur des enjeux du numérique et des nombreux systèmes d'information du domaine de la santé, voire au-delà.

FINESS est un référentiel socle, garant d'identité et d'interopérabilité ; c'est un répertoire vivant avec plusieurs dizaines de milliers de mises à jour par an. Avec l'apparition de nouvelles formes d'organisation, une refonte est nécessaire (FINESS+).

There are many issues in the healthcare world, and facing them rely on a range of players operating within various types of structures, mainly made up of legal entities or establishments. A complete listing of these organisations is crucial to identify the structures in an opposable way, map the regional healthcare offer and objectify the shortcomings, inform professionals and the general public, finance the establishments and give them a digital identity.

For over 40 years, the National file of health and social establishments (FINESS) has been filling this need. The structures and their activities are described based on decrees and orders. Each establishment is assigned a FINESS registration number, which is linked to the SIREN/SIRET numbers in INSEE's Sirene register.

This file is at the centre of digital issues and numerous information systems in the healthcare sector and even beyond.

FINESS is a core repository, guaranteeing identity and interoperability; it is a living register, with tens of thousands of updates every year. Due to the emergence of new forms of organisation, an overhaul is needed (FINESS+).

* À la date de rédaction de l'article, Consultante Secteur Public, Sia Partners, Agence du Numérique en Santé (ANS), *johanna.bensoussan.ext@esante.gouv.fr*

** Consultant, Datassence, *joel.bizingre@datassence.fr*

*** Directrice du programme FINESS, Direction du Pilotage et de l'Efficience, Agence du Numérique en Santé (ANS), *nathalie.courvalin@esante.gouv.fr*

En première ligne face aux situations de crises (COVID par exemple), aux tensions (urgences, organisation territoriale, vocations), aux défis du vieillissement de la population et aux besoins d'accompagnement social, le monde de la santé est sous les feux de l'actualité.

Par ailleurs, le numérique, des systèmes de *e*-santé et plus largement des nombreux systèmes d'information des organismes de l'État en matière de santé, jouent un rôle particulièrement important.

Lancée en 2019 et reconduite pour la période 2023-2027, la feuille de route du numérique en santé vise à rassembler et coordonner les grandes orientations des politiques numériques de la *e*-santé, sanitaire, médico-sociale et sociale. Elle s'inscrit dans une vision d'ensemble portée par « Ma Santé 2022 », action qui aspire à « *favoriser une meilleure organisation des professionnels de santé qui devront travailler ensemble et mieux coopérer au service de la santé des patients*[1] ».

L'éditorial de la ministre des Solidarités et de la Santé présentant la première feuille de route dans le cadre de Ma Santé 2022 illustre les enjeux de la partie numérique :

« *Pourtant aujourd'hui, les professionnels de santé sont confrontés à une offre numérique morcelée qui complexifie leur pratique quotidienne, et les outils numériques mis à disposition des patients-usagers sont encore trop limités. Quant à nos systèmes numériques en santé, ils présentent une grande vulnérabilité face aux cyberattaques avec des risques associés considérables*. » *(Feuille de route « Ma Santé 2022 », 2019)*

Derrière l'ambition portée par cette feuille de route, invisible ou peu connue en dehors des professionnels de santé et pourtant qui nous concerne tous, il existe depuis plus de 40 ans un répertoire support à la description régalienne de l'organisation du monde de la santé et de son offre de soins. Ce répertoire, FINESS (Fichier National des Établissements Sanitaires et Sociaux), a un rôle de plus en plus fondamental, à la fois dans la prise en compte de nouvelles formes d'organisation nécessaires à la politique de santé et dans l'appui aux services numériques existants et à venir, parmi lesquels l'identité numérique des établissements.

Dans cet article, sont présentés l'histoire de ce répertoire, son caractère régalien, sa conception et son fonctionnement ainsi que ses limites et son avenir.

Le répertoire des établissements sanitaires et sociaux : FINESS

FINESS a été créé pour répondre à l'absence de système d'information décrivant les établissements sanitaires et sociaux, leur implantation géographique et leurs capacités.

Depuis, il est le répertoire national des structures à activités réglementées des domaines sanitaire, médico-social, social ainsi que de la formation aux professions sanitaires et sociales. Placé sous la responsabilité des directions du ministère de la Santé et du ministère des Solidarités, son rôle est fondamental dans la régulation, l'évaluation, le pilotage, le financement et l'identification électronique des structures qu'il recense.

1 *https://sante.gouv.fr/systeme-de-sante/masante2022/.*

> ***FINESS a été créé pour répondre à l'absence de système d'information décrivant les établissements sanitaires et sociaux, leur implantation géographique et leurs capacités.***

Utilisé par des institutionnels, des organismes publics comme privés ou encore par le grand public, et positionné au cœur de l'écosystème des Systèmes d'Information (SI) de santé, il constitue une source d'information de référence. Ses usages se diversifient sans cesse sous l'effet du développement des applications numériques et de l'évolution de l'offre de soins[2]. Cette offre correspond à l'ensemble des ressources organisationnelles, humaines, matérielles, logistiques (dont les systèmes d'information) et financières, mises à la disposition des populations, en vue de satisfaire la demande de santé.

FINESS décrit les personnes morales, leurs activités autorisées et leurs établissements d'exercice.

Les structures décrites dans FINESS couvrent quatre domaines d'activité :

- le domaine « sanitaire » c'est-à-dire les activités de soins, de pharmacie, de laboratoire, de dispensaire, etc., et les équipements matériels lourds (IRM, scanner, gamma-caméra, tomographes par émission de positons). Cela comprend notamment les centres hospitaliers et les laboratoires de biologie médicale. En septembre 2023, ce domaine représente 48 017 établissements soit 47,4 % du total ;
- le domaine « médico-social » c'est-à-dire les prises en charge de populations atteintes de déficiences ou d'incapacités liées à l'âge, au handicap, à la maladie longue ou chronique, ou à la dépendance, et pouvant nécessiter des soins médicalisés. Cela comprend notamment les établissements de services pour l'enfance et la jeunesse handicapée et les établissements et services pour personnes âgées. Ce domaine représente 39 108 établissements soit 38,6 % du total ;
- le domaine « social » c'est-à-dire les établissements et activités uniquement à caractère social s'adressant principalement à des personnes en difficulté sociale. Cela comprend notamment les établissements et services sociaux d'aide à la famille et les établissements sociaux d'hébergement et d'accueil. Ce domaine représente 12 883 établissements soit 12,7 % du total ;
- le domaine « enseignement » qui recouvre les formations préparant aux diplômes délivrés par l'administration sanitaire et sociale. Ce domaine représente 1 360 établissements soit 1,3 % du total.

Ces établissements sont implantés en France métropolitaine, dans les départements d'Outre-mer ou dans certaines collectivités d'Outre-mer.

2 *https://www.ars.sante.fr/loffre-de-soins-en-chiffres* onglet « Soigner ».

L'histoire commence en 1979

FINESS a été institué par la circulaire du 3 juillet 1979. Il s'agit d'un répertoire commun et normalisé, renseigné au niveau régional, pour faciliter la régulation de l'offre territoriale de santé par les services compétents (gestion des autorisations et de la carte sanitaire). Il répertorie sous un identifiant unique chaque établissement, équipement et activité du domaine de la santé.

Très tôt, dans l'arrêté du 15 septembre 1988, il est décidé de mettre en place un système de gestion en temps réel et de consultation par mode vidéotex[3], confortant la modernité numérique de FINESS.

Le périmètre des structures enregistrées dans FINESS est modifié par les réglementations successives, en particulier par l'arrêté du 13 novembre 2013 qui précise que seuls les établissements soumis à autorisation préalable conformément au Code de la santé ou au Code de l'action sociale et des familles doivent être enregistrés dans FINESS.

Enfin, l'arrêté du 23 septembre 2022 introduit une évolution importante sur la portée et la gouvernance du répertoire. Historiquement, seules les structures soumises à autorisation étaient enregistrées de manière obligatoire dans FINESS. Depuis cet arrêté, l'enregistrement facultatif des établissements des domaines sanitaire, médico-social et social non soumis à autorisation est permis, afin qu'ils puissent bénéficier de moyens d'identification pour accéder dans des conditions sécurisées aux services numériques de santé au sens de l'article L 1470-1 du Code de la santé publique. Le pilotage stratégique de FINESS, historiquement assuré par la Direction de la recherche, des études, de l'évaluation et des statistiques (DREES) est assuré désormais par la Délégation du numérique en santé (DNS) et la gestion opérationnelle par l'Agence du numérique en santé (ANS).

Quel cadre juridique pour l'enregistrement des données dans FINESS ?

FINESS est défini par un cadre juridique fort et régalien. Son alimentation est régie par des règles d'enregistrement des données requérant l'existence d'actes juridiques ou administratifs. Ces actes formalisent un engagement entre une structure et une autorité de régulation (AR). Ils correspondent le plus souvent à des arrêtés d'autorisation, émanant nécessairement de cette autorité de régulation qui habilite voire finance au moins en partie ces structures. L'agrément peut également prendre la forme d'une contractualisation, matérialisée par un contrat pluriannuel d'objectifs et de moyens (CPOM).

> ***Son alimentation est régie par des règles d'enregistrement des données requérant l'existence d'actes juridiques ou administratifs.***

Tout ce qui est enregistré dans le répertoire doit l'être dans un cadre juridique : autorités, actes administratifs, arrêtés, contrats, etc.

3 Vidéographie dans laquelle la transmission des demandes d'informations des usagers et des messages obtenus en réponse est assurée par un réseau de télécommunications, en particulier le réseau téléphonique. Le Minitel est un terminal vidéotex.

Les autorités de régulation varient selon les secteurs d'activité et les catégories d'établissements. Les Agences régionales de santé (ARS) régulent les structures des champs sanitaire et médico-social, les Conseils départementaux (CD) régulent le champ social. Et les conseils départementaux régulent, soit exclusivement, soit conjointement avec les ARS, certaines structures du champ médico-social.

Les autorités de régulation transmettent les engagements (arrêtés, contrats, etc.) aux autorités dites « d'enregistrement » (AE[4]) pour inscrire les structures et activités correspondantes dans le répertoire. Ces autorités d'enregistrement varient selon les secteurs d'activité et les catégories d'établissements.

Depuis l'arrêté du 23 septembre 2022, les autorités de régulation, telles que les conseils départementaux, ont la possibilité d'enregistrer elles-mêmes les personnes morales, les établissements et les activités relevant de leur champ de compétence.

▶ Comment le contenu du répertoire est-il actualisé ?

L'actualisation du répertoire est liée aux événements juridiques et administratifs des structures : création de la structure, changement de lieu d'exercice, nouvelle activité autorisée, changement de capacité pour une activité, transfert d'activités, cessation de tout ou partie de ses activités, etc., jusqu'à la fermeture d'une entité. Elle s'effectue selon des processus bien précis.

On peut l'illustrer avec le cas du processus d'instruction d'une demande de création d'un établissement dans FINESS (***figure 1***), qui met en évidence un certain nombre d'étapes obligatoires. À la fin du processus, la structure est enregistrée dans FINESS, traduisant le cadre légal d'exercice ou d'exploitation de son activité.

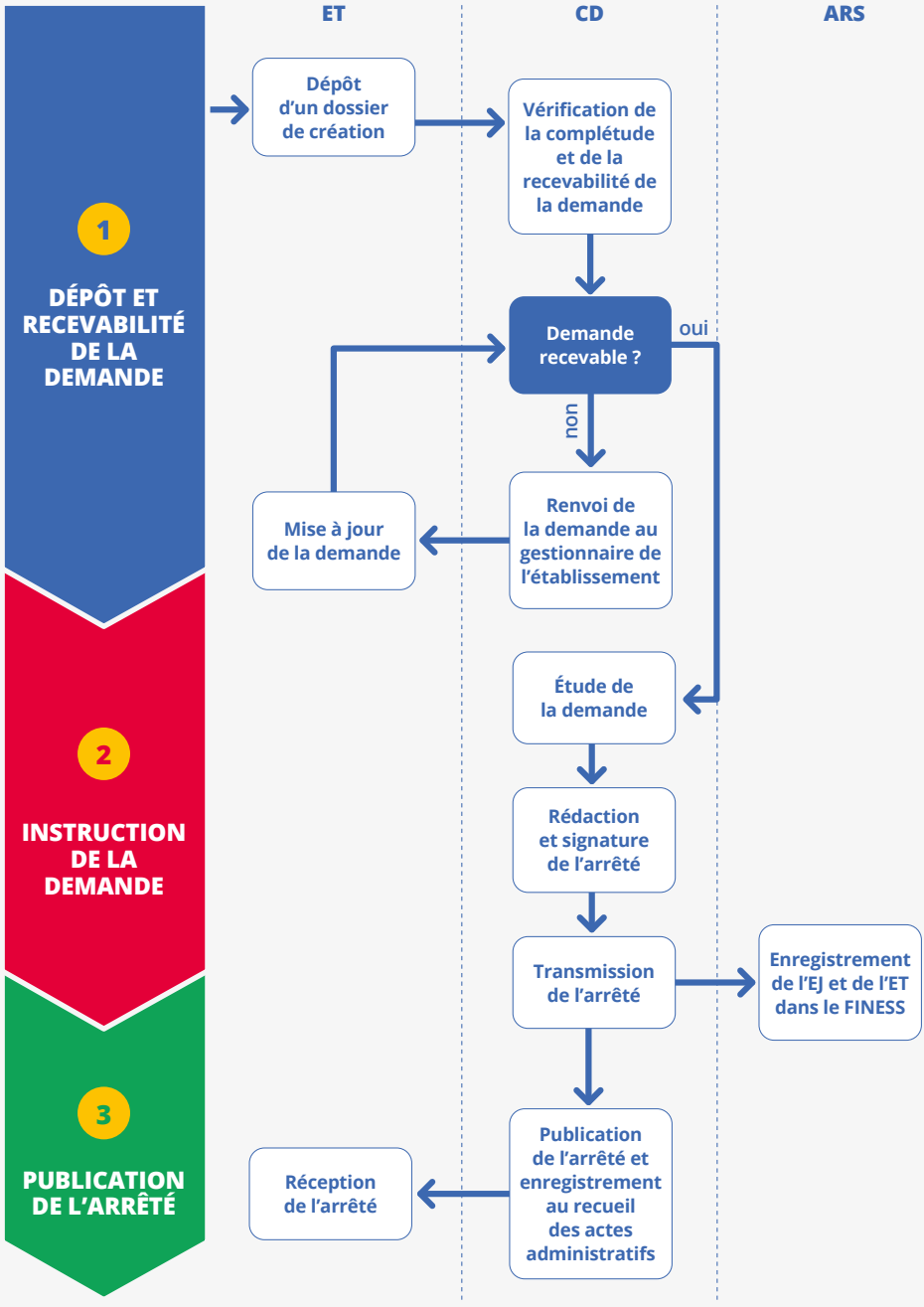
▶ Les nombreuses finalités du répertoire

La fonction de FINESS s'est élargie au fur et à mesure des nouvelles réglementations et de ses usages. Il contribue aujourd'hui à sept finalités :

- décrire l'offre en matière de structures et activités des domaines sanitaire, médico-social, social et de la formation aux professions sanitaires et sociales *(DREES, 2022)* ;
- constituer le référentiel opposable des personnes morales et des établissements acteurs du système de santé et soumis à autorisation préalable ; le numéro FINESS est l'identifiant clé des structures pour de nombreux SI de santé ;
- cartographier l'offre territoriale de santé et objectiver les manques ;
- informer les professionnels et le grand public sur l'offre de santé. FINESS est au cœur du réseau des SI de santé, alimentant par exemple des portails tels que santé.fr et pour-les-personnes-agees.gouv.fr, dédié aux personnes âgées de la Caisse nationale de solidarité pour l'autonomie (CNSA) ;
- attribuer aux établissements une identité numérique afin qu'ils puissent utiliser les services numériques en santé ;

4 AE : ARS, DREETS (Directions régionales de l'économie, de l'emploi, du travail et des solidarités), DRIHL (Direction régionale et interdépartementale de l'Hébergement et du Logement).

Figure 1 - Macro-processus de création d'un établissement



ET : Établissement
CD : Conseil départemental
ARS : Agence régionale de santé
EJ : Entité juridique

- financer les établissements par dotation ou à l'acte. La majorité des établissements ne peuvent pas être financés sans avoir été préalablement enregistrés dans le répertoire ;
- répondre à un enjeu de gestion de crise : la qualité et l'exhaustivité des données de santé revêtent une importance capitale en période de crise. En effet, FINESS est le premier point d'identification des établissements potentiellement concernés.

▶ FINESS, socle de l'écosystème des Systèmes d'Information de santé

Outre son caractère de référentiel régalien, avec de nombreuses années d'existence et des usages multiples, FINESS joue un rôle majeur dans l'écosystème des Systèmes d'Information de santé.

Outre son caractère de référentiel régalien, avec de nombreuses années d'existence et des usages multiples, FINESS joue un rôle majeur dans l'écosystème des Systèmes d'Information de santé. Cet écosystème est complexe avec des SI critiques comme celui de la Caisse nationale de l'assurance maladie (Cnam). Pour cette dernière, la liquidation de factures et les 1,5 milliard d'opérations associées nécessitent le contrôle des établissements et des autorisations à partir des données FINESS.

FINESS tient le rôle d'un référentiel de données de portée nationale (principes du Cadre Commun d'Architecture des référentiels de données de l'État[5]).

Il est le référentiel socle – central, unique et garant des identifications des entités et de leurs activités (*Bizingre et alii, 2013*), de la même façon que le SNGI pour l'identification des individus dans le domaine de la protection sociale (*Préveraud de Vaumas, 2022*) ou le RNIPP en matière d'état-civil (*Espinasse et alii, 2022*). Dès lors qu'il s'agit de décrire une structure d'une offre de santé, FINESS doit être pris en compte de façon systématique par les systèmes du monde de la santé, voire au-delà et jusque dans un cadre d'usages ouverts *(open data)*.

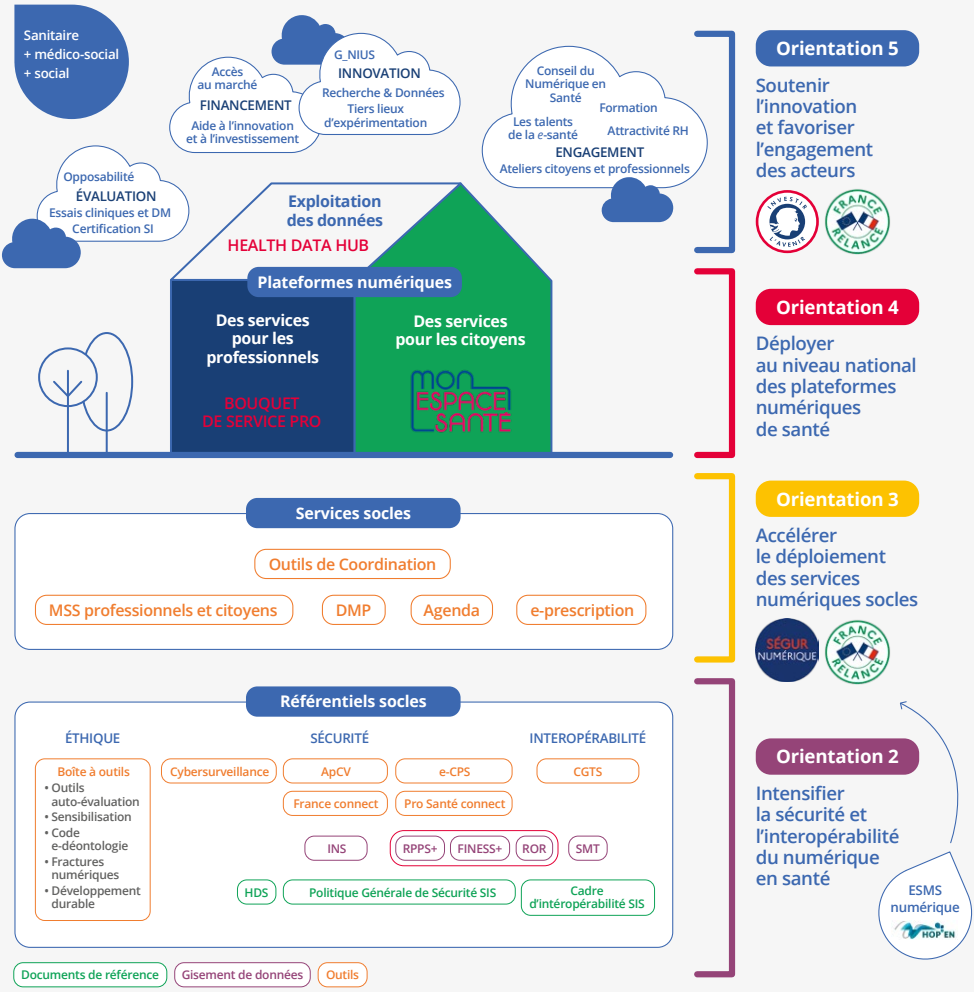
Ce rôle est confirmé par la feuille de route du numérique en santé[6].

Cette feuille de route décrit les chantiers prioritaires pour les années à venir. Elle fixe le cadre utile à leur mise en œuvre, et notamment les aspects architecturaux des systèmes d'information concernés, pour lesquels FINESS se positionne comme référentiel socle (***figure 2***).

Plus précisément, la place et le rôle de FINESS dans cette feuille de route se situent au sein des politiques générales des Systèmes d'Information de santé. En tant que garant légal d'identité, il intervient dans la politique générale de sécurité pour l'identification électronique des acteurs de santé et contribue à l'annuaire national d'identification des utilisateurs de services de *e*-santé. Dans la politique d'interopérabilité, l'utilisation et la circulation des identifiants FINESS sont préconisées au sein des systèmes et entre ces systèmes pour automatiser les rapprochements de données ou le suivi du processus concernant les établissements (exemple de l'allocation et du suivi budgétaire des établissements par la Direction générale de l'offre de soins (DGOS)).

5 *https://docplayer.fr/192521-Cadre-commun-d-architecture-des-referentiels-de-donnees.html.*
6 *https://esante.gouv.fr/actualites/lancement-de-la-feuille-de-route-du-numerique-en-sante-2023-2027.*

▶ Figure 2 - Place de FINESS dans les référentiels socles



ApCV : Application Carte Vitale
CGTS : Centre de Gestion des Terminologies de Santé
DMP : Dossier Médical Partagé
DM : Dispositif médical
ESMS : Établissements et Services Médico-Sociaux
E-CPS : Version électronique Carte de Professionnel de Santé
HDS : Hébergeur de Données de Santé
INS : Identité Nationale de Santé
MSS professionnels et citoyens : Messageries sécurisées de santé professionnels et citoyens
RH : Ressources Humaines
ROR : Répertoire Opérationnel des Ressources
RPPS : Répertoire Partagé des Professionnels de Santé
SI : Système d'Information
SIS : Systèmes d'Information de Santé
SMT : Serveur Multi-Terminologies

▶ FINESS au centre d'un système de répertoires

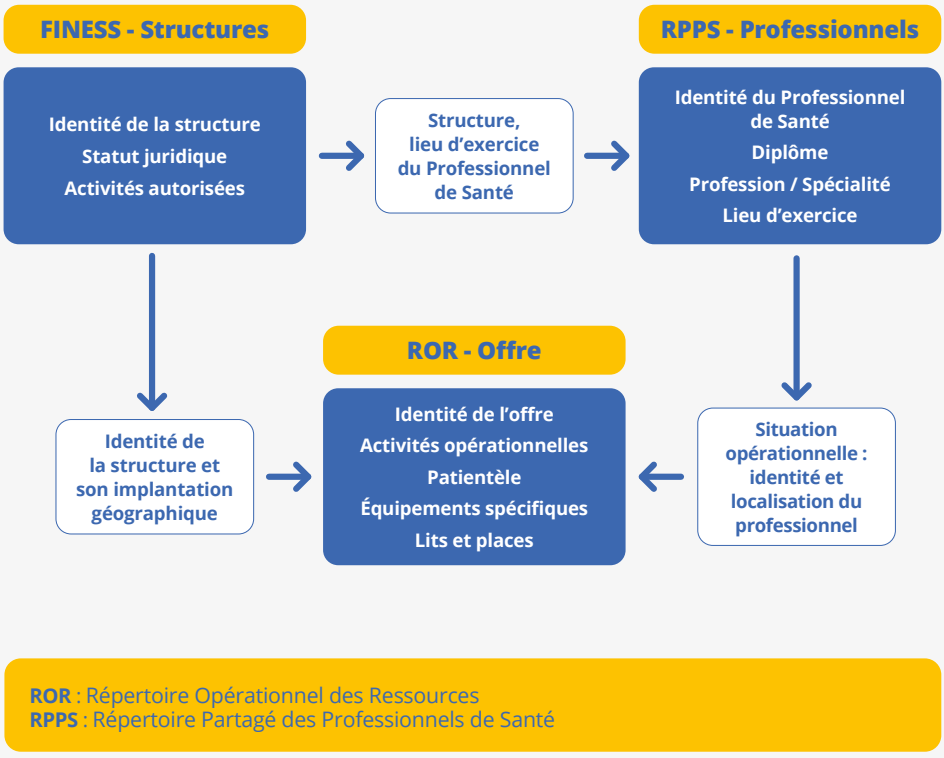
Le rôle de référentiel socle de FINESS est renforcé par son articulation avec les deux autres référentiels socles qui décrivent l'organisation et les acteurs associés du monde de la santé : le Répertoire Partagé des Professionnels de Santé (RPPS)[7] et le Répertoire Opérationnel des Ressources (ROR).

FINESS et les répertoires ROR et RPPS forment un trio au cœur de l'organisation du système de santé (personnes morales, établissements et professionnels de santé).

Le RPPS est le répertoire unique de référence permettant d'identifier les professionnels de santé. Il rassemble et publie les informations des professionnels de santé, à partir d'un numéro RPPS attribué au professionnel toute sa vie.

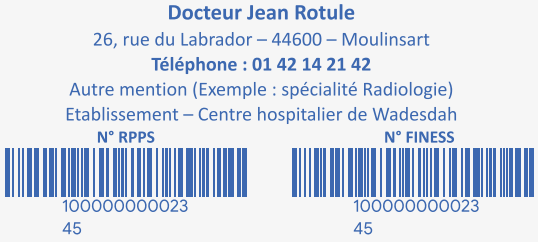
Le ROR décrit, de façon normalisée, la partie opérationnelle des établissements, c'est-à-dire l'ensemble des soins et des services qui y sont dispensés. Le ROR facilite l'orientation des patients, la communication auprès des familles, la coordination entre les acteurs du parcours de santé, de soins et de vie (***figure 3***).

▶ Figure 3 - Articulation entre les 3 répertoires FINESS - RPPS - ROR



7 *https://esante.gouv.fr/produits-services/repertoire-rpps.*
https://www.omedit-nag.fr/actualites/double-identification-rppsfiness.

La coordination entre les trois répertoires est fondamentale. L'appariement n° FINESS – n° RPPS – ROR permet de garantir l'interopérabilité entre et au sein de nombreux systèmes. Elle permet de répondre à la question : à quelle entité FINESS tel professionnel de santé est-il rattaché et quelle est l'offre de soins de cette entité ? Le ROR intègre les données identifiantes du RPPS et du FINESS. De la même façon, le RPPS enregistre le lien entre un professionnel de santé et sa structure de rattachement FINESS.



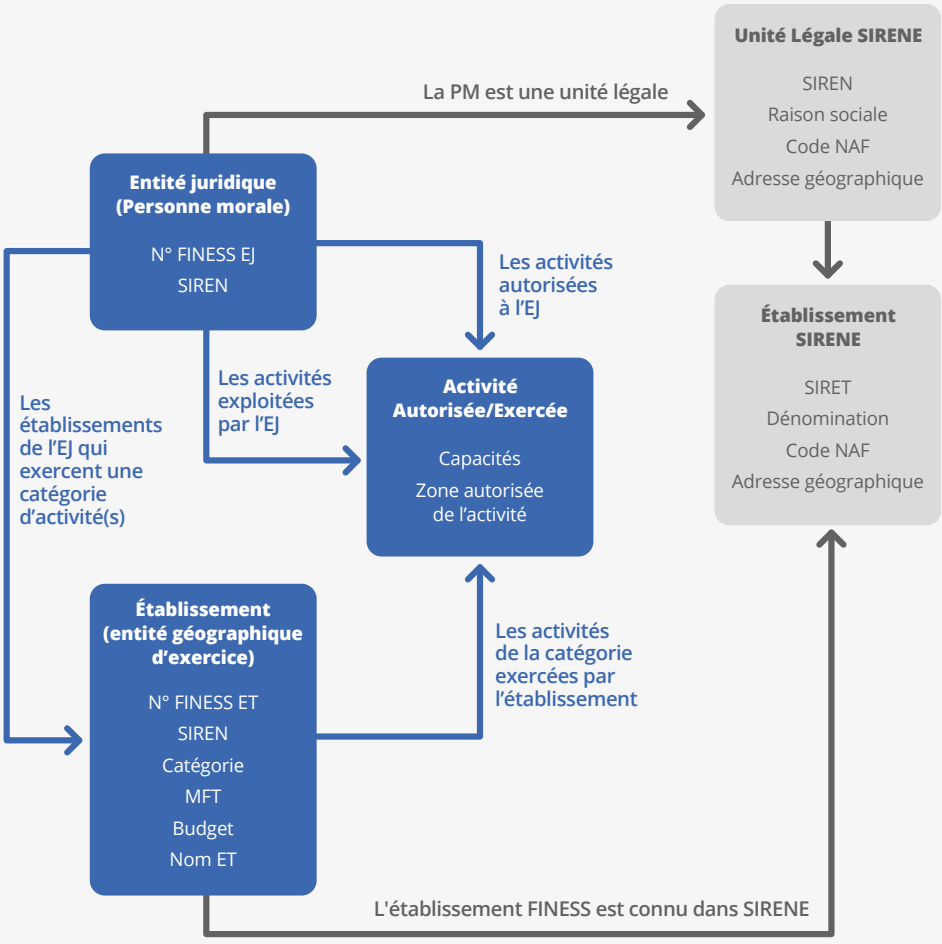
Quelles sont les données de référence du répertoire FINESS ?

Le répertoire FINESS s'organise autour de trois types d'objets : entité juridique, établissement et activité (***figure 4***) :

> ***Le répertoire FINESS s'organise autour de trois types d'objets : entité juridique, établissement et activité.***

- les entités juridiques (EJ) identifiées par un numéro FINESS juridique et le numéro SIREN associé du répertoire Sirene. Elles correspondent aux personnes morales au sens juridique. Elles sont responsables juridiquement et elles détiennent des droits d'activité sous forme d'autorisations, d'agréments, de conventions ou de tout type d'acte administratif de nature à autoriser ;
- les établissements identifiés par un numéro FINESS géographique et le numéro SIRET associé. Ils dépendent d'une EJ et traduisent un groupe d'activités de même nature exercée sur un lieu géographique donné. Si deux groupes d'activités de nature distincte (désignés par des codes catégorie différents) sont exercés sur un même lieu, deux établissements sont créés dans FINESS (***encadré 1***) ;
- les activités sont décrites selon les domaines, par des activités de soins, des équipements matériels lourds (IRM, scanner), des disciplines d'équipement social ou médico-social ou encore des disciplines d'enseignement.

Figure 4 - Schéma des objets de référence FINESS



Un établissement exerce des activités autorisées au niveau de son entité juridique de rattachement.

Une entité juridique peut avoir plusieurs établissements qui lui sont rattachés.

EJ : Entité juridique
ET : Établissement
MFT : Mode de Fixation des Tarifs
NAF : Nomenclature d'activité française
PM : Personne Morale

▶ Encadré 1. Les données de référence de FINESS

Les entités juridiques (EJ) sont décrites dans FINESS par les données suivantes :

- le n° FINESS juridique ;
- le n° SIREN ;
- le statut juridique ;
- la raison sociale ;
- l'adresse d'implantation ;
- des coordonnées de contact ;
- les dates marquant les grands événements de sa vie (création et fermeture notamment).

Les établissements (ET) sont décrits dans FINESS par les données suivantes :

- le n° FINESS géographique ;
- le n° SIRET ;
- la catégorie ;
- le n° FINESS de l'EJ auquel il est rattaché ;
- une dénomination ;
- une adresse d'implantation ;
- des coordonnées de contact ;
- les dates marquant les grands événements de sa vie (création et fermeture notamment) ;
- un mode de fixation tarifaire (tarif libre, tarif conventionné, etc.).

Les activités du domaine sanitaire sont exprimées sous la forme « activité / modalité / forme », par exemple « Médecine d'urgence / SMUR Antenne / Non saisonnier ». D'autres activités sanitaires sont enregistrées en tant que « disciplines d'équipement sanitaire », par exemple la discipline « 086 − Activité de vaccination gratuite ».

Les activités du domaine médico-social sont exprimées sous la forme « discipline / mode de fonctionnement / clientèle », par exemple : « Accueil au titre de la Protection de l'Enfance / Accueil de jour / Enfants, adolescents et jeunes majeurs ASE (Aide Sociale à l'Enfance) ». Et les activités de formation sont exprimées sous forme de disciplines d'enseignement, par exemple : « Formation DE Infirmier ».

Les activités sont autorisées pour une capacité donnée. Deux types de capacités sont enregistrés dans FINESS : les capacités « autorisées » dans l'arrêté et les capacités réellement « installées ».

Par ailleurs, les activités peuvent être circonscrites à une zone d'intervention précisée dans l'acte administratif d'autorisation. Cette zone d'intervention enregistrée dans FINESS peut être définie à l'échelle régionale, départementale et/ou communale.

À noter que FINESS fait appel à plus d'une centaine de nomenclatures, ce qui témoigne de sa richesse en termes de capacité de représentation, tout en nécessitant une certaine maîtrise.

▶ L'immatriculation : une fonction socle du répertoire

L'immatriculation est la fonction principale de FINESS pour créer une structure. Les activités exercées au sein de cette structure lui sont ensuite rattachées.

À partir de l'immatriculation[8], FINESS doit garantir à ses clients l'exactitude des résultats suivants :

- l'identification, c'est-à-dire retrouver à partir de traits d'identité, les numéros FINESS correspondants et les activités associées ;

8 Voir l'article de Séverine Bidet-Caulet et Christian Burel, au sujet de l'immatriculation, sur le répertoire académique et ministériel sur les établissements du système éducatif Ramsese, dans ce même numéro.

- la consultation, c'est-à-dire l'opération inverse qui consiste à retrouver à partir des numéros FINESS, les caractéristiques des entités correspondantes et de leurs activités ;
- et plus classiquement, la recherche et la vérification de l'existence de structures et d'activités à partir de caractéristiques (raison sociale, adresse, code catégorie d'activité, etc.).

▶ Qu'est-ce que le numéro FINESS ?

L'immatriculation est déclenchée par un événement juridique (arrêté d'autorisation, convention, contrat pluriannuel d'objectif et de moyens, etc.) émanant d'une autorité de régulation qui justifie l'enregistrement d'une structure dans FINESS (entité juridique, établissement). L'opération d'immatriculation est réalisée par l'autorité d'enregistrement compétente. Elle se matérialise par l'attribution d'un numéro FINESS. Sans l'immatriculation, l'entité ne peut pas exercer.

Deux types de numéros FINESS existent. Le numéro FINESS juridique qui est attribué à toute entité juridique enregistrée dans FINESS. Et le numéro FINESS géographique qui est attribué à tout établissement enregistré dans FINESS.

La structure du numéro FINESS est la même, qu'il s'agisse d'un numéro FINESS juridique ou d'un numéro FINESS géographique. Elle est composée comme suit :

N° FINESS de l'EHPAD du Centre Hospitalier d'Is-sur-Tille (Côte d'Or)

* *Calculée automatiquement – permet de s'assurer de la validité du numéro*[9].

L'immatriculation dans FINESS est régie par plusieurs règles.

Le numéro FINESS est immuable dans le temps sauf dans le cas particulier d'un changement de département. Le numéro FINESS n'est jamais recyclé.

Plusieurs numéros FINESS géographiques peuvent être associés à la même adresse. Par exemple, l'entité juridique « Assistance Publique Hôpitaux de Paris », FINESS juridique 750 712 184 détient 2 établissements situés à la même adresse 1 rue Georges Clémenceau 91 750 Champcueil. Le premier établissement correspond à un établissement de soins de longue durée (USLD) nommé « USLD HENRI MONDOR SITE CLEMENCEAU APHP » identifié par le n° FINESS géographique 910 021 963. Et le deuxième établissement correspond au Centre Hospitalier Régional nommé « HU HENRI MONDOR SITE CLEMENCEAU APHP » identifié par le n° FINESS géographique 910 100 015.

9 *https://fr.wikipedia.org/wiki/Formule_de_Luhn.*

L'immatriculation s'appuie sur un ensemble limité de traits d'identité qui vont définir l'unicité de la structure : raison sociale, dénomination, adresse géographique, n° SIREN/SIRET, et la catégorie d'activité selon une nomenclature au centre de FINESS.

Appariement avec le répertoire Sirene

> ***Le rapprochement entre FINESS et le répertoire Sirene est légitime de par leur périmètre commun en matière d'objets : entités juridiques et établissements.***

Le rapprochement entre FINESS et le répertoire Sirene est légitime de par leur périmètre commun en matière d'objets : entités juridiques et établissements, mais il existe quelques différences (***encadré 2***).

Lors de l'enregistrement, un appariement est réalisé entre le numéro FINESS juridique et un numéro SIREN et entre le numéro FINESS géographique d'un établissement et un numéro SIRET.

Cet appariement renforce l'assise juridique de FINESS. Comme pour le répertoire FINESS, le répertoire Sirene (*Alviset, 2020*) dispose d'une assise juridique (décret n° 73-314 du 14 mars 1973) qui renforce celle de FINESS.

Encadré 2. Évolution et comparaison avec le répertoire Sirene des notions d'établissement et d'entité juridique

À l'origine du fichier FINESS, les établissements n'hébergeaient que des activités de même nature. Or, au cours du temps, les activités des établissements se sont diversifiées et la notion d'établissement telle qu'elle avait été définie ne permettait plus de représenter la diversité des types d'activités exercées sur un site géographique donné.

Aujourd'hui, un établissement désigne un site géographique où sont exercées des activités autorisées de même nature. Sa catégorie dans FINESS dépend donc du type d'activité exercé.

La notion et la place des établissements dans la construction du fichier FINESS sont différentes de celles de l'Insee. Il n'y a pas de correspondance un pour un. En effet, à un établissement au sens Sirene avec son numéro SIRET, peut correspondre plusieurs établissements au sens FINESS identifiés chacun par numéro FINESS géographique propre.

Par ailleurs, les règles de gestion divergent parfois. Dans le cas d'un changement de lieu d'exercice, le répertoire Sirene fermera l'ancien établissement et en ouvrira un nouveau alors que dans FINESS l'immatriculation reste la même dès lors que l'établissement reste au sein d'un même département.

Face à l'ensemble de ces divergences de concepts et de règles de gestion, il est parfois difficile d'avoir le bon numéro SIRET en face du numéro FINESS géographique.

Demain, dans le futur FINESS (FINESS+), la dénomination « établissement » (ET) évoluera vers la dénomination « entité géographique d'exercice » (EGE), pour se séparer plus clairement du concept d'établissement de la base Sirene*.

En revanche, concernant la représentation des entités juridiques, la correspondance entre le répertoire FINESS et le répertoire Sirene est « un pour un ». Le répertoire FINESS peut être vu comme un sous-ensemble du champ de Sirene limité au secteur de la santé.

Dans le futur FINESS (FINESS+), la dénomination d'entité juridique évoluera au profit du libellé Personne Morale Sanitaire/Médico-social/Social/Enseignement en correspondance avec la terminologie du dernier arrêté FINESS de 2022 et également pour souligner la correspondance de concept avec Sirene (à une entité juridique – Personne Morale FINESS correspond une unité légale de la base Sirene avec son numéro SIREN**).

* *https://www.insee.fr/fr/metadonnees/definition/c1377.*

** *https://www.insee.fr/fr/metadonnees/definition/c1044.*

Il permet de consolider la qualité des données *via* des contrôles de cohérence de présence et d'état des structures communes entre les deux répertoires (coordination sur les immatriculations – nouvelles entrées dans les répertoires, et sur le statut – actif/inactif – sorties des répertoires).

Enfin il répond aux enjeux d'interopérabilité entre systèmes de différents organismes, *via* la capacité à rapprocher des données connues à partir d'un numéro SIREN/SIRET et des données connues à partir d'un numéro FINESS.

▶ Une visibilité et des usages toujours plus vastes

Les données du fichier FINESS sont consommées par une diversité d'acteurs pour des usages qui ne cessent de se diversifier et de s'accroître. Parmi les clients, se trouvent des acteurs de la santé mais également des acteurs plus éloignés (***figure 5***).

Les acteurs de la santé qui utilisent FINESS sont les administrations centrales et régionales d'État[10], les organismes ou partenaires extérieurs (Insee, Assurance maladie), certains institutionnels (le Samu Social de Paris), les établissements eux-mêmes et le grand public.

Les acteurs hors du champ de la santé incluent notamment la Direction Générale des Finances Publiques (alimentation d'un infocentre financier avec des données FINESS), la RATP et la SNCF, ces derniers gérant en propre des centres de santé déclarés dans FINESS.

À partir d'extractions FINESS, les administrations centrales croisent, filtrent, trient des données (catégories, clientèles, activités, etc.) à différents niveaux géographiques (national, régional, départemental, communal) pour cartographier l'offre, identifier des manques, définir et suivre le déploiement des politiques publiques.

Par exemple, l'Agence Technique de l'Information sur l'Hospitalisation (ATIH) utilise FINESS pour le Programme de Médicalisation des Systèmes d'Information (PMSI), outil de gestion du financement des établissements de santé (tarification à l'activité) et d'organisation de l'offre de soins (planification).

L'Insee et la DREES utilisent les données FINESS pour alimenter des enquêtes et produire des statistiques, notamment les enquêtes Statistique Annuelle des Établissements (SAE) ou d'autres enquêtes[11].

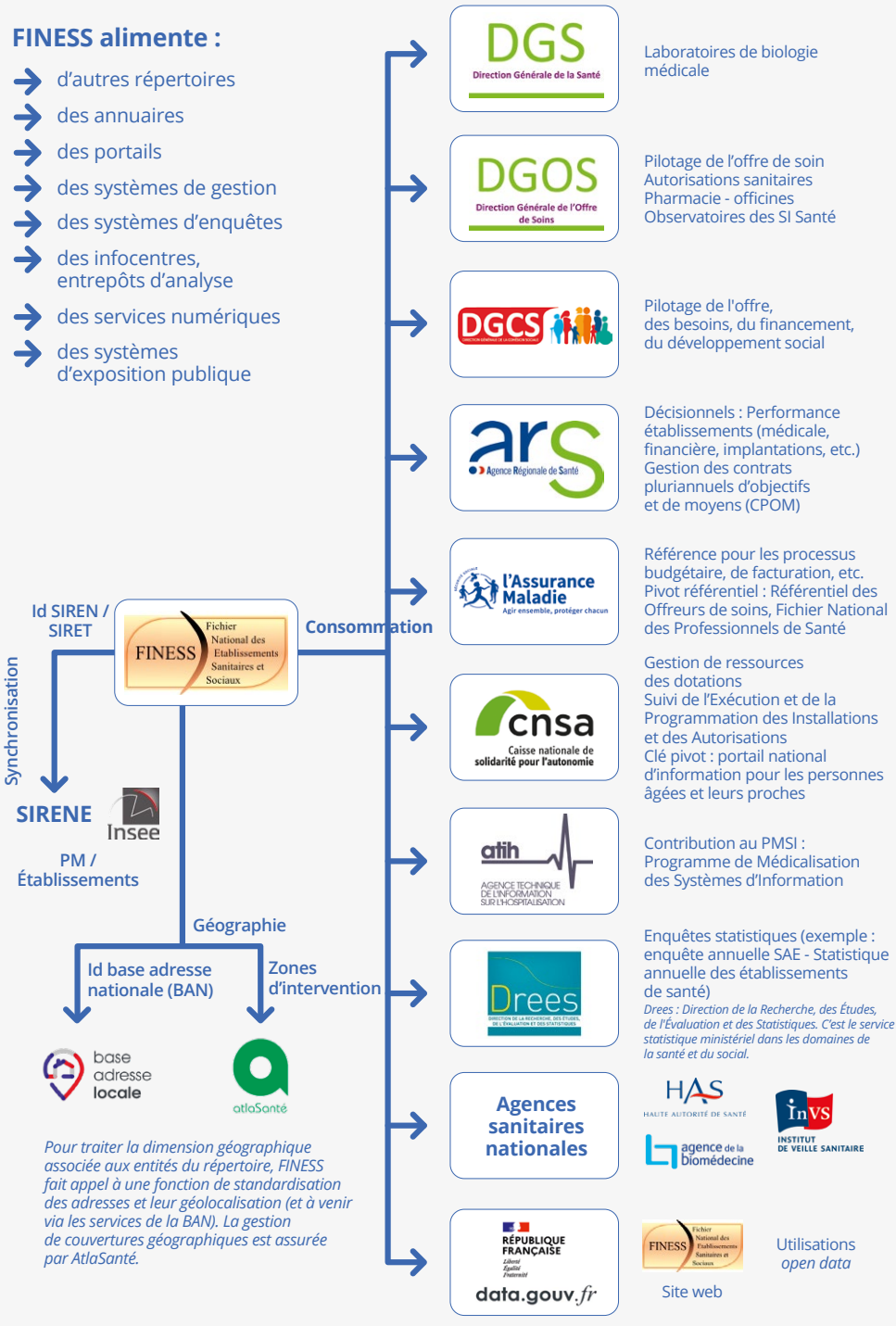
Certains partenaires utilisent FINESS pour informer le public sur l'offre sanitaire et sociale, à travers leur portail pour connaître les établissements et les lieux de prise en charge. Par exemple, le portail de la CNSA[12] délivre aux usagers la liste des EHPAD sur un territoire donné.

10 Direction Générale de la Cohésion Sociale, Direction Générale de l'Offre de Soins, Direction Générale de la Santé, Agences Régionales de Santé, Directions Régionales de l'Économie de l'Emploi du Travail des Solidarité, Directions Régionales et Interdépartementales de l'Hébergement et du Logement.

11 SAE *https://www.sae-diffusion.sante.gouv.fr/sae-diffusion/accueil.htm*
https://drees.solidarites-sante.gouv.fr/sources-outils-et-enquetes/00-la-statistique-annuelle-des-etablissements-sae.

12 *https://www.pour-les-personnes-agees.gouv.fr/.*

▶ Figure 5 - FINESS au service de nombreux clients



FINESS alimente également le système de gestion d'identité numérique des établissements de santé : cela leur permet d'accéder aux services du numérique en santé. Par exemple, l'immatriculation récente des Maisons Départementales pour les Personnes Handicapées (MDPH) dans FINESS a permis de les doter de moyens d'identification électronique pour accéder au service de messageries sécurisées MSSanté[13].

La maîtrise de la qualité des données du répertoire

Le cadre juridique du fichier FINESS implique l'enregistrement systématique des structures devant être immatriculées. Cependant, ce cadre juridique ne suffit pas à assurer la qualité de toutes les données saisies dans FINESS. Des pratiques hétérogènes de rédaction des arrêtés juridiques (première source de données FINESS) ainsi que des échanges insuffisamment normalisés entre les autorités de régulation et d'enregistrement conduisent à des défauts de qualité des données. La complexité de la description des activités ajoute une difficulté supplémentaire concernant la fiabilité de FINESS.

> ***Près de 35 000 entités juridiques et/ou établissements sont créés ou modifiés chaque année.***

Face à ces facteurs de manque de qualité des données, compte tenu de la multiplicité des clients, des usages du fichier FINESS et du caractère central du répertoire, des dispositifs de maîtrise de la qualité de données ont été mis en place. Sachant que près de 35 000 entités juridiques et/ou établissements sont créés ou modifiés chaque année, ces dispositifs reposent sur différents contrôles automatisés ainsi que sur des interventions humaines.

Des outils de paramétrage des activités existent également pour encadrer et guider la saisie des activités. Lorsqu'un gestionnaire d'une autorité d'enregistrement enregistre une activité, les interfaces de saisie lui proposent un « champ des codes nomenclatures possibles ».

Enfin, les enquêtes statistiques récurrentes auprès des établissements sont l'occasion de vérifier leurs données et de les corriger si besoin.

Un répertoire ouvert et diffusé largement

Les données présentes dans FINESS sont publiques et le répertoire offre actuellement de multiples modalités d'accès à celles-ci. À l'origine, les données FINESS étaient diffusées uniquement à la demande et *via* des fichiers spécifiques. Puis, un ensemble de fichiers standards a été défini ; les données sont désormais publiées quotidiennement. Les données FINESS sont actuellement consultables sur un portail web[14] avec des fonctions de recherche et d'extraction. Et les fichiers de données sont publiés en *open data* sur la plateforme ouverte des données publiques françaises (*data.gouv.fr*), avec une publication bimestrielle : données des entités juridiques[15], données des établissements[16], données des autorisations[17], etc.

13 MSSanté est un système de messageries électroniques, réservé aux professionnels de santé.
14 *https://finess.esante.gouv.fr.*
15 *https://www.data.gouv.fr/fr/datasets/finess-extraction-des-entites-juridiques/.*
16 *https://www.data.gouv.fr/fr/datasets/finess-extraction-du-fichier-des-etablissements/.*
17 *https://www.data.gouv.fr/fr/datasets/finess-extraction-des-autorisations-dactivites-de-soins/.*

Enfin, les données FINESS sont accessibles depuis cette année par une *Application Programming Interface* (API)[18] ; elles sont restituées à la norme internationale *Fast Healthcare Interoperability Resource* (FHIR)[19].

▶ FINESS +, une refonte nécessaire…

Pour bénéficier de l'apport de nouvelles technologies et pallier certaines limites, la DREES a engagé en 2021 un projet de refonte applicative, repris par l'ANS en 2022 et qui donnera lieu à une nouvelle version du fichier FINESS, nommée FINESS+.

Dans le fichier FINESS, il est impossible de décrire certains types de groupes et leur composition sous forme de relations collaboratives ou juridiques. Ainsi, la composition de groupes privés (une *holding* et ses membres comme le groupe DOMOTYS dans le cadre des EHPAD) ou d'un Groupement Hospitalier de Territoire (GHT) ne peut pas être représentée. Elle ne permet pas non plus de décrire des fédérations d'associations, des groupements de coopération ou des dispositifs, pour faciliter la coordination territoriale de structures privées et publiques autour de parcours complexes de patients ou pour la mise en commun entre différentes structures d'une activité (cancérologie par exemple). Dans le cadre de FINESS+, les différentes formes de composition pourront être retracées ainsi que les événements s'y rapportant.

▶ … qui répond à un besoin de retracer l'historique des structures…

Les besoins des clients du répertoire ne se limitent pas à disposer d'un état des structures et activités à un instant *t*. La connaissance des événements et des modifications apportées aux caractéristiques des établissements dans le passé est essentielle pour comprendre l'historique des structures.

> ***La connaissance des événements et des modifications apportées aux caractéristiques des établissements dans le passé est essentielle pour comprendre l'historique des structures.***

Le projet FINESS+ introduit explicitement dans son modèle cette gestion des événements actuellement absente et qui permettra de disposer de la vision du passé (***figure 6***).

Les types d'événements considérés sont :

- les événements qui conditionnent l'existence de la structure : fusion, scission, suppression, etc. ;
- les événements sur les relations qu'il peut y avoir entre structures (affectation d'un ET à une EJ, évolution des relations entre ET) ;

18 Une API (*application programming interface* ou « interface de programmation d'application ») est une interface logicielle qui permet de « connecter » un logiciel ou un service à un autre logiciel ou service afin d'échanger des données et des fonctionnalités.

19 FHIR (*Fast Healthcare Interoperability Resource*) est une norme d'interopérabilité développée par HL7 (l'organisme de normalisation du niveau de santé 7) conçue pour permettre l'échange de données de santé par voie électronique entre différents systèmes du secteur de la santé.

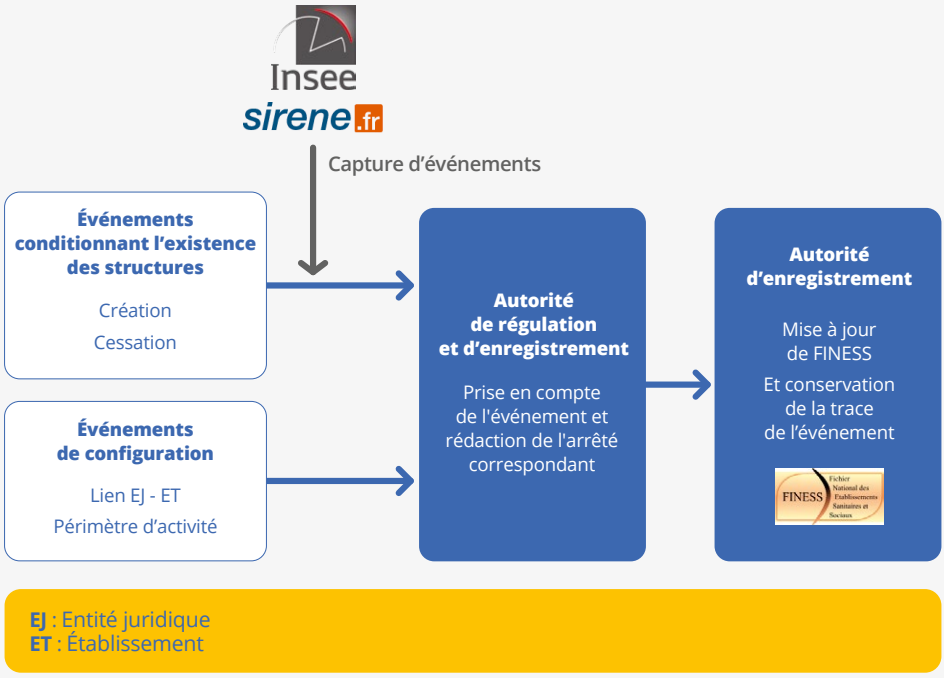
- les événements sur les activités (reconfiguration du périmètre d'activités d'un ET, nouvelles activités, récupération d'activités d'un autre ET, et réciproquement cession d'activités à un autre ET, etc.).

Les événements sont décrits par un ensemble de caractéristiques : type d'événement, structure et activité concernée, date de l'événement en accord avec les éléments fournis par les autorités, impact sur les relations entre structures et activités, nouvel état à la suite de l'événement.

La connaissance des événements va répondre à de nombreux besoins :

- Étudier l'évolution des activités d'une entité. Reconstituer la trajectoire individuelle de chaque entité en tenant compte des événements ayant eu un impact (fusion, scission). Comprendre l'origine d'écarts de mesures pour une entité, du fait de la prise en charge de nouvelles activités ;
- Fournir des statistiques agrégées par territoire (région, département) sur les dispositifs de santé dans le temps. Expliquer comment a évolué le nombre d'entités de telle catégorie, année par année, comparer la configuration des entités à N, N-1, [...], N-5 (par exemple : baisse en raison de la fusion d'entités, du transfert d'activités entre entités) ;
- Surveiller des concentrations, des monopoles sur une activité donnée, sur un territoire donné. Surveiller la fragilité des territoires en offres de santé (décision par un groupe de se défaire d'une structure, seule sur un territoire donné). Suivre une entité (une clinique par exemple) absorbée par un groupe.

▶ Figure 6 - Gestion des événements dans FINESS



Dans les événements marquants, la fin de vie des entités est essentielle. Mais cela ne signifie pas qu'on le supprime du fichier FINESS. Par exemple pour la Cnam : « Si un établissement ferme, on doit pouvoir accepter les prescriptions liées à cette entité pendant 27 mois ».

Enfin dans le cadre de sa refonte, FINESS intégrera une gestion classique d'historisation en masse des changements de valeurs de la grande majorité des attributs du répertoire (conservation de la trace de la date de modification, de la valeur avant, de la valeur après).

▶ … avec une aide à la saisie pour obtenir des données plus fiables

La version actuelle de l'application permet aux Agences régionales de santé (ARS) et aux Directions régionales de l'économie, de l'emploi, du travail et des solidarités (DREETS) d'enregistrer des structures. D'autres autorités, telles que les conseils départementaux doivent pouvoir le faire, conduisant ainsi à une meilleure qualité du référentiel.

La notion d'engagement entre une autorité et une structure sera décrite explicitement permettant ainsi de rapprocher plus facilement les données enregistrées et les différents cadres juridiques associés.

L'adossement à Sirene sera complété par une vérification de l'existence de l'entité juridique et/ou de l'établissement en cours d'enregistrement par appel direct à la base Sirene (contrôle des SIREN et SIRET). La communication entre Sirene et FINESS permettra de capter des événements de vie de type « cessation d'activité » ou « fermeture d'activité » en lien avec les structures et de signaler ces événements pour prise en compte par les gestionnaires.

Enfin l'aide à la saisie sera renforcée afin d'homogénéiser les pratiques d'enregistrement entre les différentes autorités.

Des contrôles automatisés seront ajoutés pour renforcer la qualité du référentiel, notamment ceux portant sur l'adresse. Dès qu'une adresse sera renseignée, les données saisies seront vérifiées auprès de la Base Adresse Nationale et corrigées automatiquement.

En définitive, en plus de quarante années d'existence, à travers de nombreuses réformes réglementaires, des adaptations successives de FINESS ont été nécessaires sans que sa structure et son fonctionnement ne soient modifiés en profondeur. Cependant, aujourd'hui, FINESS doit se transformer pour répondre à la forte évolution du paysage des systèmes numériques et à la prise en compte de toutes les nouvelles formes d'organisation, d'exercice et de contractualisation. L'arrivée prochaine de FINESS+ est une véritable transformation tant stratégique que fonctionnelle du fichier FINESS qui cherche à renforcer sa capacité à répondre à l'histoire à venir.

▶ Fondements juridiques

- Code de la santé publique. Article L1470-1. In : *site de Légifrance*. [en ligne]. [Consulté le 18 octobre 2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000043497477*.

- Code de l'Action Sociale et des Familles. Paragraphe 2. In : *site de Légifrance*. [en ligne]. [Consulté le 7 décembre 2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/codes/id/LEGISCTA000033642841/2017-02-01*.

- Décret n° 73-314 du 14 mars 1973 portant création d'un système national d'identification et d'un répertoire des entreprises et de leurs établissements. In : site de Légifrance. [en ligne]. [Consulté le 18 octobre 2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000306534*.

- Arrêté du 15 septembre 1988 relatif à la mise en place d'un système de gestion en temps réel et de consultation par mode vidéotex du fichier national des établissements sanitaires et sociaux (FINESS). In : site de Légifrance. [en ligne]. [Consulté le 18 octobre 2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000498981*.

- Arrêté du 13 novembre 2013 relatif à la mise en place d'un répertoire national des établissements sanitaires et sociaux. In : site de Légifrance. [en ligne]. [Consulté le 18 octobre 2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000028218434*.

- Arrêté du 23 septembre 2022 relatif à la mise en œuvre du « Répertoire national des établissements sanitaires, médico-sociaux et sociaux » (FINESS). In : site de Légifrance. [en ligne]. [Consulté le 18 octobre 2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000046349894*.

- Circulaire DAGPB/DOMI n° 79-1 du 3 juillet 1979 relative au fichier national des établissements sanitaires et sociaux. In : Bulletin Officiel n° 2002 – 14. [en ligne]. [Consulté le 18 octobre 2023]. Disponible à l'adresse : *https://sante.gouv.fr/fichiers/bo/2002/02-14/a0141329.htm*.

Bibliographie

- ALVISET, Christophe, 2020. La troisième refonte du répertoire Sirene : trop ambitieuse ou pas assez. In : *Courrier des statistiques*. [en ligne]. 29 juin 2020. Insee, N° N4, pp. 101-121. [Consulté le 6 décembre 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4497083/courstat-4-7.pdf*.

- BIZINGRE, Joël, PAUMIER, Joseph, RIVIÈRE, Pascal, 2013. *Les référentiels du système d'information - Données de référence et architectures d'entreprise*. Juillet 2013. Collection : InfoPro, Dunod. EAN : 9782100598748.

- DREES, 2022. *Les établissements de santé*. Édition 2022. [en ligne]. [Consulté le 18 octobre 2023]. Disponible à l'adresse : *https://drees.solidarites-sante.gouv.fr/publications-documents-de-reference-communique-de-presse/panoramas-de-la-drees/les-etablissements*.

- ESPINASSE, Lionel et ROUX, Valérie, 2022. Le Répertoire national d'identification des personnes physiques (RNIPP) au cœur de la vie administrative française. In : *Courrier des statistiques*. [en ligne]. 29 novembre 2022. Insee. N° N8, pp. 72-92. [Consulté le 6 décembre 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6665188?sommaire=6665196*.

- MA SANTÉ, 2022. *Dossier d'information. Feuille de route "accélérer le virage numérique"*. 25 avril 2019. [en ligne]. [Consulté le 18 octobre 2023]. Disponible à l'adresse : *https://sante.gouv.fr/IMG/pdf/190425_dossier_presse_masante2022_ok.pdf.*

- MINISTÈRE DE LA SANTÉ ET DE LA PRÉVENTION, 2022. Ma santé 2022 : un engagement collectif – *Dossier de presse*. 18 septembre 2018. [en ligne]. [Consulté le 18 octobre 2023]. Disponible à l'adresse : *https://sante.gouv.fr/IMG/pdf/ma_sante_2022_pages_vdef_.pdf*.

- PRÉVERAUD DE VAUMAS, Joseph, 2022. Un référentiel des identités pour les besoins de la sphère sociale. Le système national de gestion des identifiants (SNGI). In : *Courrier des statistiques*. [en ligne]. 29 novembre 2022. Insee. N° N8, pp. 93-114. [Consulté le 6 décembre 2023]. Disponible à l'adresse : *www.insee.fr/fr/information/6665190?sommaire=6665196*.

Le répertoire d'établissements Ramsese au service des acteurs du système éducatif

Séverine Bidet-Caulet*, Christian Burel**

Le répertoire académique et ministériel sur les établissements du système éducatif, Ramsese, est au cœur du système administratif et statistique du ministère de l'Éducation nationale. Il désigne à la fois un outil de collecte, de gestion et de mise à disposition de données sur les établissements du système éducatif français. Il répond au besoin du ministère de disposer d'un répertoire de référence sur les établissements avec des données de qualité.

Ce répertoire comprend tous les établissements qui assurent une activité de formation initiale générale, technique ou professionnelle, de la maternelle à l'enseignement supérieur, du secteur public ou privé, que ces établissements soient sous tutelle du ministère de l'Éducation nationale ou non. Environ 90 000 établissements sont recensés avec un identifiant national unique.

Il constitue l'un des référentiels indispensables aux statistiques produites par la Direction de l'évaluation, de la prospective et de la performance (Depp), que ce soit pour le lancement d'enquêtes ou les remontées de sources administratives et leur contrôle. La gestion de ce référentiel est fondamentale dans le positionnement de la Depp au sein du ministère, son accès à des données administratives et la qualité des statistiques qu'elle produit.

The academic and ministerial register of educational establishments, Ramsese, is the core of the Ministry of Education's administrative and statistical system. It is a tool aimed at collecting, managing and making available data on schools in the French education system. It meets the Ministry's need for a reference register of schools with high-quality data.

This register includes all establishments providing initial general, technical or vocational training, from pre-school to higher education, in the public or private sector, whether these establishments are under the supervision of the Ministry of National Education or not. Some 90,000 establishments are listed with a unique national identifier.

It is one of the essential registers for the statistics produced by the Statistical Office of the Ministry of Education, the Department of Evaluation, Forecasting and Performance (DEPP), whether for launching surveys or transferring and checking administrative sources. The management of this repository is fundamental to DEPP's positioning within the Ministry, its access to administrative data and the quality of statistics produced.

* À la date de rédaction de l'article, cheffe de projet au bureau des nomenclatures et répertoires, Depp, Ministère de l'Éducation nationale et de la jeunesse,
severine.bidet-caulet@education.gouv.fr

** Chef du bureau des nomenclatures et répertoires, Depp, Ministère de l'Éducation nationale et de la jeunesse,
christian.burel@education.gouv.fr

Avant 1977, l'absence d'identification des établissements du système éducatif par un numéro unique dans un fichier centralisé rendait, au sein du ministère de l'Éducation nationale, les opérations de gestion et les opérations statistiques complexes.

Inscrit dans le schéma directeur de l'informatique du ministère, le premier « fichier des établissements », consolidé à partir des fichiers rectoraux des établissements, avait été mis en place en 1977 avec vocation à servir de fichier de référence (avec un identifiant unique d'établissement) pour l'ensemble des applications informatiques du ministère, et obligation pour les rectorats de transférer ces informations par le biais des services statistiques académiques (SSA).

À partir de 1996, le ministère a choisi de renforcer la responsabilité des rectorats en leur confiant le soin de la gestion de leurs établissements et en palliant également l'obsolescence technique du dispositif en place. La gestion du répertoire a été organisée en trente bases académiques, chaque académie disposant d'un outil de gestion dédié. La base nationale consolidée était constituée mensuellement par un transfert de fichiers vers l'administration centrale.

Le répertoire académique et ministériel sur les établissements du système éducatif, Ramsese, existe sous cette dénomination depuis 1996, à la suite d'une importante refonte du système d'information relatif à la gestion des établissements du système éducatif.

Aujourd'hui, la gestion décentralisée du répertoire est toujours assurée par les SSA, mais l'outil informatique est mutualisé au sein de Ramsese, avec une base de données nationale et une interface gestionnaire qui sont les mêmes pour tous.

Ramsese est un répertoire vivant et couvre un large périmètre d'établissements dont l'identification est nécessaire aux systèmes d'information du ministère de l'Éducation nationale et de la Jeunesse ainsi qu'à ceux de l'Enseignement supérieur et de la Recherche. Mais quel est vraiment son rôle au sein de ces deux ministères ? Comment est-il alimenté ? Qui assure la gestion et la mise à jour de ses données ? Comment s'assurer de la qualité des données ? Quels en sont les usages et qui peut avoir accès à ses données ?

▶ Des besoins très variés

Né du besoin de disposer d'un répertoire centralisé d'établissements, permettant d'identifier des entités propres, avec des caractéristiques spécifiques aux ministères de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche, Ramsese s'est construit sans prendre appui sur un cadre juridique. Il s'est doté d'informations que l'on ne retrouve pas nécessairement dans d'autres répertoires centraux externes, tels que le répertoire Sirene (*Alviset, 2020*). Sa notoriété au sein de l'éducation nationale s'est accrue au fil du temps avec le développement et la multiplication des échanges entre systèmes d'information.

Dans les académies, les services statistiques doivent disposer de données sur les établissements pour répondre à des besoins locaux, alimenter les études académiques, les outils d'aide au pilotage, ainsi que les transferts d'informations statistiques à fournir à la Depp dans le cadre de ses enquêtes.

Au niveau national, la Depp doit répondre aux besoins statistiques et de pilotage du ministère (*Repères et références statistiques, 2023*). Ainsi, elle réalise des enquêtes ou échantillonne des

établissements, elle calcule des indicateurs au niveau de l'établissement (*Les établissements scolaires, 2023*), ou tout simplement elle compte les élèves scolarisés dans les écoles du premier degré ou dans les établissements du 2nd degré, qu'ils soient publics ou privés, et ce à chaque rentrée scolaire. Pour cela la Depp doit pouvoir disposer d'un référentiel d'établissements sur lequel s'appuyer.

De son côté, la direction du numérique doit mener des projets d'urbanisation des applicatifs pour répondre aux besoins métiers et assurer la gouvernance du système d'information (SI). Ce travail implique de recenser et de capitaliser l'ensemble des informations qui sont disponibles. Si la majorité des SI dispose d'informations sur les établissements, la présence de données semblables mais non identiques dans les bases ou les fichiers est une source de complexité, qui exige de construire des « points de vérité » uniques sur les données du SI. Le répertoire Ramsese constitue un point de vérité. En retour la Depp doit accéder à des données issues des systèmes de gestion du ministère.

▶ Un positionnement particulier dans le paysage éducatif

Le répertoire des établissements Ramsese permet de répondre à une grande variété de besoins.

Ainsi, de nombreuses applications académiques développées localement pour les besoins propres des académies ou pour le besoin national s'appuient directement ou indirectement sur les données du répertoire[1].

> ***Ramsese facilite le partage des données dans le cadre des projets d'urbanisation des applicatifs.***

Le répertoire permet de répondre à de nombreux besoins locaux, tels que la réponse à la demande (recherche d'adresses, comptage d'établissements, etc.) ou la constitution de listes d'établissements (établissements avec BTS[2], liste d'établissements soumis à la contribution de vie étudiante et de campus CVEC, etc.).

Ramsese facilite le partage des données dans le cadre des projets d'urbanisation des applicatifs. Le référentiel étant commun et centralisé, il contribue fortement au dialogue au sein des écosystèmes de ressources humaines ou de scolarité. Une cinquantaine d'applications nationales du ministère sont ainsi utilisatrices des données du répertoire. Cela concerne les systèmes d'information d'administration tels que ceux utilisés pour le suivi des élèves et des étudiants, ceux qui permettent l'orientation des élèves (dont Parcoursup[3]), ou le suivi des concours et des examens, ainsi que les systèmes d'information pour la santé scolaire, pour le suivi des personnels de l'éducation nationale et pour la gestion budgétaire et financière.

Les données de Ramsese sont aussi utilisées pour établir le constat de rentrée (recenser les élèves à chaque rentrée scolaire), une des opérations statistiques « phare » de la Depp à chaque rentrée, ou dans les enquêtes telles que l'enquête Système d'information de la

1 Voir l'article de Johanna Bensoussan, Joël Bizingre et Nathalie Courvalin, sur le répertoire des établissements de santé FINESS, dans ce même numéro, en ce qui concerne la position centrale du répertoire.
2 BTS : brevet de technicien supérieur.
3 Parcoursup est une plateforme web destinée à recueillir et gérer les vœux d'affectation des futurs étudiants de l'enseignement supérieur français.

formation des apprentis (SIFA) qui recense l'ensemble des apprentis inscrits au 31 décembre de chaque année dans un centre de formation d'apprentis (CFA) quel que soit le niveau d'études. L'unité interrogée est l'établissement. Autre illustration, le dispositif InserJeunes diffuse au niveau de chaque établissement différents indicateurs pour toutes les formations professionnelles du CAP[4] au BTS, pour mieux informer les jeunes sur l'insertion des sortants de l'établissement, le taux de poursuite d'études, le taux d'interruption en cours de formation ou la valeur ajoutée de l'établissement sur le taux d'emploi[5] et fournir des outils de pilotage aux acteurs de la voie professionnelle.

Par ailleurs, les indicateurs de valeur ajoutée des lycées (IVAL) permettent de mesurer l'action propre de chaque lycée (*Evain, 2020*), en prenant en compte la réussite des élèves au baccalauréat et leur parcours scolaire dans l'établissement. Ils sont établis au niveau établissement, pour chaque lycée public ou privé sous contrat, à partir des données de la session antérieure du baccalauréat.

Les tirages d'échantillons de certaines enquêtes utilisent les données du répertoire, comme par exemple l'enquête internat, cadre de vie (ICV) ou l'enquête programme international pour le suivi des acquis des élèves (PISA), qui mesure l'efficacité des systèmes éducatifs en évaluant les acquis scolaires d'un élève de 15 ans.

Les chefs d'établissement du premier et second degrés ou les services des rectorats et inspections académiques utilisent également les données au niveau établissement grâce à une application de recherche et de choix d'indicateurs de pilotage des établissements et écoles (Archipel) qui met à disposition différents indicateurs sur les établissements.

De plus, Ramsese est aussi utilisé à l'extérieur des ministères de l'Éducation nationale et de l'enseignement supérieur et de la recherche, par exemple dans les espaces numériques de travail (ENT[6]). Dans ce cadre, les ENT qui utilisent comme clé d'identification unique le numéro attribué à chaque établissement par Ramsese ont la possibilité de récupérer un ensemble d'informations, soit directement dans le répertoire Ramsese (comme les regroupements pédagogiques intercommunaux), soit dans d'autres SI qui utilisent également ce numéro (comme les formations dispensées dans l'établissement).

> ***En étendant à l'ensemble du ministère l'utilisation d'un référentiel commun sur les établissements, la Depp contribue ainsi à l'amélioration de la qualité des systèmes d'information.***

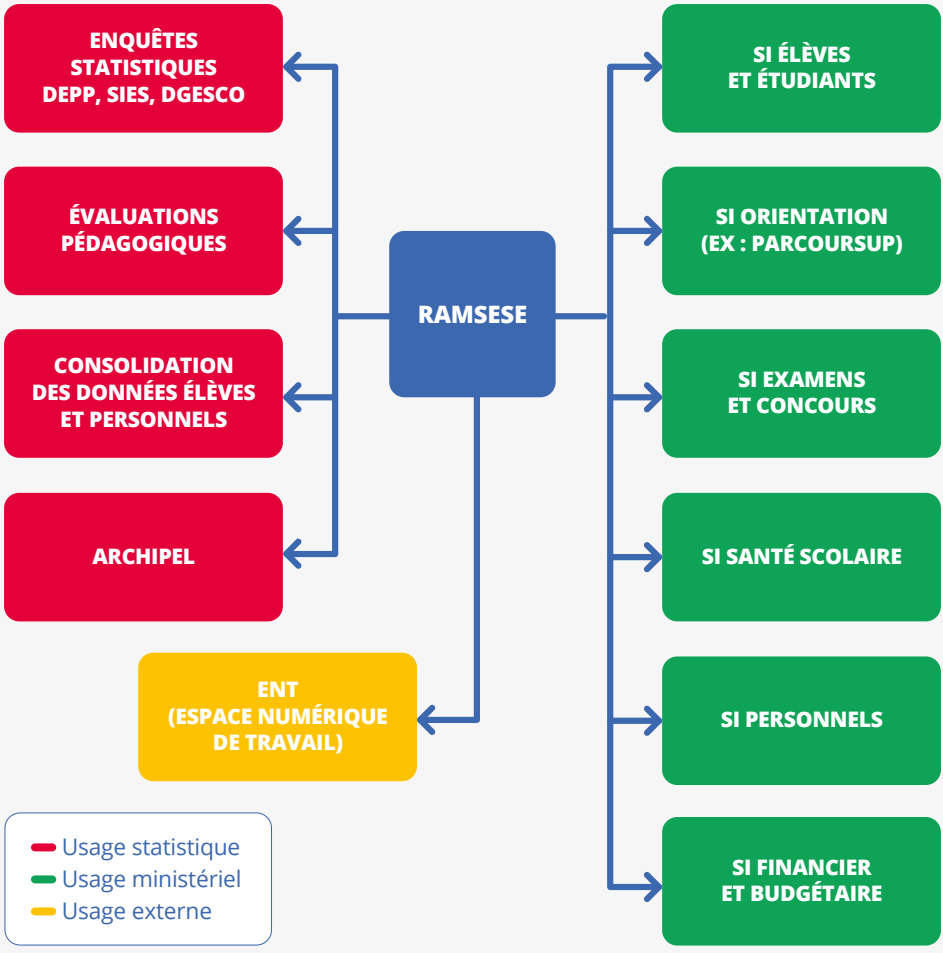
En étendant à l'ensemble du ministère l'utilisation d'un référentiel commun sur les établissements, la Depp contribue ainsi à l'amélioration de la qualité des systèmes d'information, favorise l'interopérabilité de différents SI (*Bizingre et alii, 2013*) et en conséquence facilite leur gestion.

4 CAP : Certificat d'aptitude professionnelle.
5 La valeur ajoutée d'un établissement sur le taux d'emploi est un indicateur qui compare de façon pertinente le taux d'emploi des élèves sortants de cet établissement au taux d'emploi d'établissements similaires.
6 ENT : un espace numérique de travail (ENT) désigne un ensemble intégré de services numériques choisis et mis à disposition de tous les acteurs de la communauté éducative d'une ou plusieurs écoles ou d'un ou plusieurs établissements scolaires, dans un cadre de confiance défini par un schéma directeur des ENT et par ses annexes. Il constitue un point d'entrée unifié permettant à l'utilisateur d'accéder, selon son profil et son niveau d'habilitation, à ses services et contenus numériques. Il offre un lieu d'échange et de collaboration entre les usagers, et avec d'autres communautés en relation avec l'école ou l'établissement.

Ramsese constitue l'un des référentiels indispensables aux statistiques produites par la Depp, que ce soit pour le lancement d'enquêtes ou le transfert des sources administratives (sur la scolarité des élèves, les ressources humaines…) et leur contrôle.

La position centrale du répertoire ***(Figure 1)*** dans le système d'information est essentielle pour garantir la cohérence d'ensemble des SI des ministères. Cela implique une veille constante des réformes afin de faire évoluer le répertoire en conséquence et dans les délais impartis[7].

▶ Figure 1 - Ramsese au cœur des systèmes d'information (SI)



DEPP : Direction de l'évaluation, de la prospective et de la performance
SIES : Sous-direction des systèmes d'information et des études statistiques
DGESCO : Direction générale de l'enseignement scolaire
ARCHIPEL : Application de recherche et de choix d'indicateurs de pilotage des établissements et écoles
PARCOURSUP : plateforme nationale de préinscription en première année de l'enseignement supérieur

7 Voir l'article de Johanna Bensoussan, Joël Bizingre et Nathalie Courvalin sur FINESS dans ce même numéro.

▶ Chaque structure du système éducatif est immatriculée

Afin de répondre à la diversité des usages, décrits ci-dessus, le répertoire couvre un large champ :

- les établissements qui assurent une activité de formation initiale générale, technique ou professionnelle, de la maternelle à l'enseignement supérieur, du secteur public ou privé, que ces établissements soient sous tutelle du ministère de l'Éducation nationale ou non. On trouve ainsi dans le répertoire les écoles maternelles ou élémentaires, les collèges, les lycées, les universités, les écoles d'ingénieurs, de commerce, etc. ;
- les établissements de formation continue de l'Éducation nationale, les GRETA[8], les centres de formation d'apprentis (CFA) ;
- les établissements d'administration du système éducatif public tels que les rectorats ou les inspections académiques ;
- les établissements de formation médico-sociale.

Les entités du répertoire sont des établissements au sens juridique, tels qu'un lycée, mais peuvent être des parties d'établissement n'ayant pas de personnalité juridique, telles que des annexes[9], pour autant que l'identification de ces structures soit nécessaire à plusieurs systèmes d'information des ministères de l'Éducation nationale ou de l'Enseignement supérieur et de la Recherche. Il peut aussi s'agir d'une composante possédant une autonomie juridique au sein d'un établissement, tout particulièrement dans l'enseignement supérieur, telle qu'un IUT[10], un UFR[11] ou un GRETA.

Il peut ainsi y avoir plusieurs entités à la même adresse, un lycée et sa section d'enseignement professionnel par exemple. Cette distinction permet notamment d'appliquer à la section d'enseignement professionnel, dans les systèmes d'information, les mêmes règles que pour les lycées professionnels et se justifie par une direction d'établissement, un profil d'élèves et d'enseignants ainsi que des moyens attribués différents du lycée auquel la section est rattachée. De même, deux immatriculations à une même adresse sont nécessaires pour un collège qui dispose d'une section d'enseignement général et professionnel adapté (SEGPA). La SEGPA est immatriculée et correspond à une structure rattachée au collège assurant la formation d'élèves présentant des difficultés scolaires.

> ***La notion d'unité administrative immatriculée (UAI) a été retenue comme unité de base du répertoire de la Depp.***

Compte tenu de ces spécificités, la notion d'unité administrative immatriculée (UAI) a été retenue comme unité de base du répertoire de la Depp, une UAI se définissant ainsi comme un établissement au sens juridique, une partie d'établissement sans personnalité juridique ou une composante disposant d'une certaine autonomie juridique au sein d'un établissement.

8 GRETA : les groupements d'établissements publics locaux d'enseignements (GRETA) sont les structures de l'éducation nationale qui organisent des formations pour adultes dans pratiquement tous les domaines professionnels.
9 Annexe géographique rattachée pédagogiquement à un établissement principal.
10 IUT : institut universitaire de technologie.
11 UFR : unité de formation et de recherche.

Le champ géographique couvert correspond à la France métropolitaine, aux départements d'Outre-mer (DOM), aux collectivités d'Outre-mer (COM) et aux pays étrangers dans lesquels sont implantés des établissements français ayant reçu un agrément conjoint du ministère des Affaires étrangères et du ministère de l'Éducation nationale ou de celui de l'Enseignement supérieur et de la Recherche.

Certaines structures ne sont cependant pas répertoriées. Ainsi, les établissements privés qui ne proposent que des cours de soutien scolaire ou d'entraînement complémentaires à un cursus scolaire ou universitaire ne sont pas immatriculés. Il en est de même de certains lieux de formation, tels que les salles de sport ou les piscines dans lesquelles se déroulent des formations.

Le répertoire recense ainsi environ 90 000 UAI ouvertes aujourd'hui sur 150 000 UAI ouvertes et fermées au total. L'écart entre les 90 000 UAI ouvertes et les 150 000 UAI gérées par le répertoire correspond aux UAI fermées depuis la création du répertoire en 1977.

▶ Un répertoire avec des données d'identification

Les entités du répertoire sont identifiées par leur numéro UAI qui constitue une clé d'identification unique sur huit positions.

Les traits d'identité sont constitués de données distinctes de l'identifiant, mais permettant d'identifier l'entité. Dans Ramsese il s'agit du sigle, de la dénomination principale qui correspond à une appellation normalisée permettant de savoir de quel type de structure il s'agit (par exemple : une école maternelle publique), ainsi que de la dénomination complémentaire qui correspond au nom de la structure (par exemple : Marie Curie) ou de l'appellation officielle qui permet de décliner dans son intégralité le nom officiel de l'UAI. L'adresse postale permet de savoir où se trouve l'UAI ; elle est enregistrée dans Ramsese selon la norme postale.

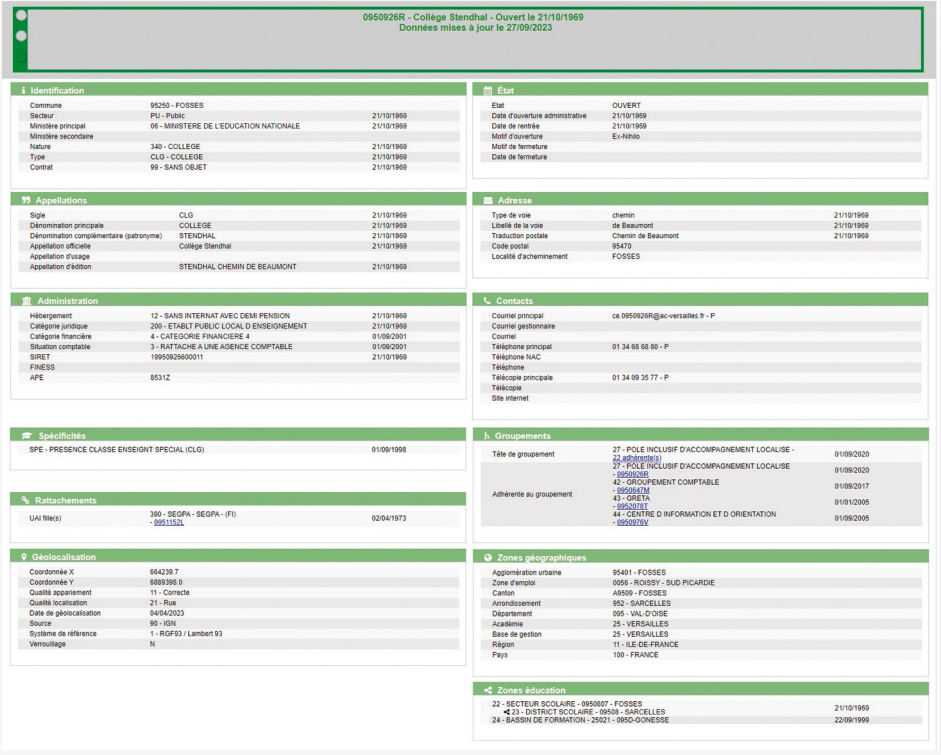
Ramsese est un répertoire socle (*Rivière, 2022*) qui ne devrait contenir que les données intrinsèques à l'UAI (la dénomination, la localisation, la date d'ouverture, etc.), les sous-ensembles de données métier plus spécialisées (relatives aux formations dispensées par les établissements, aux classes, aux effectifs, etc.) étant reportés dans d'autres référentiels adossés au référentiel socle.

Un contenu qui ne se réduit pas aux seules données référentielles

Ramsese propose des données de nature référentielle mais également des données de gestion telles que le numéro de téléphone d'une UAI ou des variables indispensables aux autres SI, permettant de décrire chaque structure mais limitées à des caractéristiques stables dans le temps. Une UAI contient les informations dont la liste a été arrêtée en 1992 lors de la décision de refonte de l'application et dont le contenu est resté quasi stable. Dans Ramsese, une vingtaine de variables sont associées à un numéro UAI ***(figure 2)***.

Les variables constituant les traits d'identité sont recensées mais aussi des caractéristiques de classification, des données géographiques, des données de zonage d'éducation, des données de groupement, des données de rattachement, des données de gestion administrative et des données de contact.

Figure 2 - Exemple pour le collège Stendhal de Fosses



0950926R - Collège Stendhal - Ouvert le 21/10/1969
Données mises à jour le 27/09/2023

Identification

Commune	95250 - FOSSES	
Secteur	PU - Public	21/10/1969
Ministère principal	06 - MINISTERE DE L'EDUCATION NATIONALE	21/10/1969
Ministère secondaire		
Nature	340 - COLLEGE	21/10/1969
Type	CLG - COLLEGE	21/10/1969
Contrat	99 - SANS OBJET	21/10/1969

État

Etat	OUVERT
Date d'ouverture administrative	21/10/1969
Date de rentrée	21/10/1969
Motif d'ouverture	Ex-Nihilo
Motif de fermeture	
Date de fermeture	

Appellations

Sigle	CLG	21/10/1969
Dénomination principale	COLLEGE	21/10/1969
Dénomination complémentaire (patronyme)	STENDHAL	21/10/1969
Appellation officielle	Collège Stendhal	21/10/1969
Appellation d'usage		
Appellation d'édition	STENDHAL CHEMIN DE BEAUMONT	21/10/1969

Adresse

Type de voie	chemin	21/10/1969
Libellé de la voie	de Beaumont	21/10/1969
Traduction postale	Chemin de Beaumont	21/10/1969
Code postal	95470	
Localité d'acheminement	FOSSES	

Administration

Hébergement	12 - SANS INTERNAT AVEC DEMI PENSION	21/10/1969
Catégorie juridique	200 - ETABLT PUBLIC LOCAL D ENSEIGNEMENT	21/10/1969
Catégorie financière	4 - CATEGORIE FINANCIERE 4	01/09/2001
Situation comptable	3 - RATTACHE A UNE AGENCE COMPTABLE	01/09/2001
SIRET	19950926600011	21/10/1969
FINESS		
APE	8531Z	

Contacts

Courriel principal	ce.0950926R@ac-versailles.fr - P
Courriel gestionnaire	
Courriel	
Téléphone principal	01 34 68 68 80 - P
Téléphone NAC	
Téléphone	
Télécopie principale	01 34 09 35 77 - P
Télécopie	
Site internet	

Spécificités

SPE - PRESENCE CLASSE ENSEIGNT SPECIAL (CLG)	01/09/1998

Groupements

Tête de groupement	27 - POLE INCLUSIF D'ACCOMPAGNEMENT LOCALISE - (2 adhérent(s))	01/09/2020
Adhérente au groupement	27 - POLE INCLUSIF D'ACCOMPAGNEMENT LOCALISE - 0950926B	01/09/2020
	42 - GROUPEMENT COMPTABLE - 0950647M	01/09/2017
	43 - GRETA - 0952078T	01/01/2005
	44 - CENTRE D INFORMATION ET D ORIENTATION - 0950976V	01/09/2005

Rattachements

UAI fille(s)	390 - SEGPA - SEGPA - (FI) - 0951152L	02/04/1973

Géolocalisation

Coordonnée X	664239.7
Coordonnée Y	6889398.0
Qualité appariement	11 - Correcte
Qualité localisation	21 - Rue
Date de géolocalisation	04/04/2023
Source	90 - IGN
Système de référence	1 - RGF93 / Lambert 93
Verrouillage	N

Zones géographiques

Agglomération urbaine	95401 - FOSSES
Zone d'emploi	0056 - ROISSY - SUD PICARDIE
Canton	A9509 - FOSSES
Arrondissement	952 - SARCELLES
Département	095 - VAL-D'OISE
Académie	25 - VERSAILLES
Base de gestion	25 - VERSAILLES
Région	11 - ILE-DE-FRANCE
Pays	100 - FRANCE

Zones éducation

22 - SECTEUR SCOLAIRE - 0950807 - FOSSES	
◀ 23 - DISTRICT SCOLAIRE - 09508 - SARCELLES	21/10/1969
24 - BASSIN DE FORMATION - 25021 - 095D-GONESSE	22/09/1999

Des caractéristiques permettent de catégoriser chaque UAI : les plus importantes sont la nature de l'UAI ***(Encadré 1)***, le secteur, le ministère de tutelle et les spécificités. On distingue secteur public et secteur privé, et à l'intérieur de ce dernier, le type de contrat. Les spécificités permettent d'apporter une information complémentaire à la nature. Par exemple, une école élémentaire qui accueille des élèves de maternelle aura la spécificité

▶ Encadré 1. La nature de l'unité administrative immatriculée (UAI) : une typologie détaillée de l'activité

La nature permet de caractériser de la façon la plus fine possible l'activité ou la mission de l'UAI.

Les natures sont regroupées en huit grandes sous-catégories, décrites dans le tableau ci-dessous. Ces ensembles peuvent représenter des niveaux d'enseignement (1, 3, 4 et 5), des types de formation (6 et 7) ou encore correspondre à l'organisation administrative (8). Le premier caractère du code nature précise la sous-catégorie à laquelle appartient l'UAI.

On trouve ainsi :

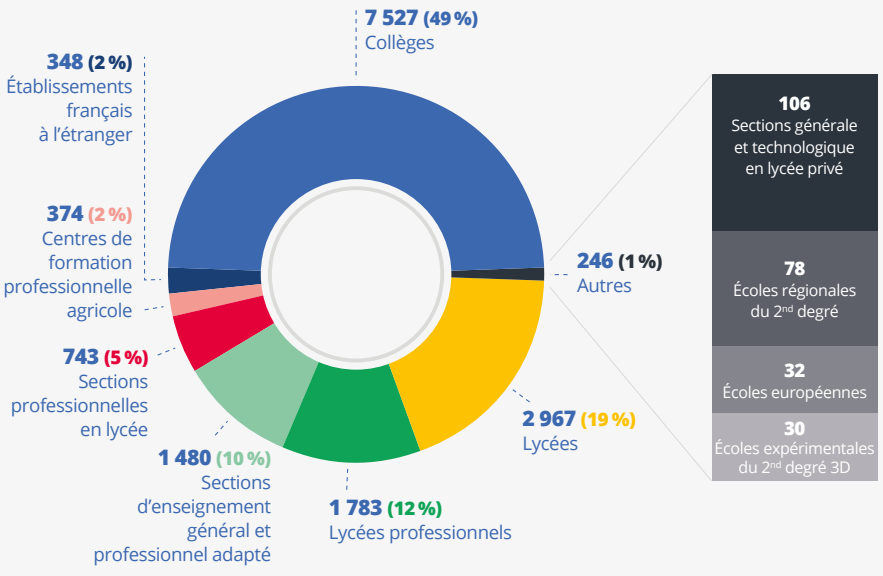
- Sous-catégorie 1 : les UAI du premier degré ;
- Sous-catégorie 2 : les UAI médico-éducatifs et socio-éducatifs ;
- Sous-catégorie 3 : les UAI d'enseignement du second degré ;
- Sous-catégorie 4 : les UAI de formation spécialisée post-secondaire et supérieure non universitaire ;
- Sous-catégorie 5 : les UAI d'enseignement supérieur publics et privés et les écoles d'ingénieurs ;
- Sous-catégorie 6 : les UAI des centres de formation d'apprentis ;
- Sous-catégorie 7 : les UAI de formation continue ;
- Sous-catégorie 8 : les UAI des établissements et services d'administration de l'éducation nationale, de l'enseignement supérieur et de la recherche.

Une UAI appartient à un groupe plus ou moins vaste de structures semblables : les écoles élémentaires (code nature 151) constituent un groupe de près de 37 000 UAI alors que la nature 568 permet d'identifier les 32 instituts nationaux supérieurs du professorat et de l'éducation (INSPE).

Actuellement, 179 natures actives existent dans Ramsese.

Par exemple, la répartition des natures dans la sous-catégorie 3 (2nd degré) (voir figure).

Dans le 2nd degré, les UAI constituent un ensemble homogène, avec une variété limitée de natures. La scolarisation dans un collège unique a conduit à une création importante de collèges ainsi qu'à une augmentation du nombre de lycées, accentuée à partir de 1985 avec la création du bac professionnel.



CLP[12] qui indique la présence de ces élèves. Dès lors, si dans une enquête tous les élèves de maternelle sont concernés, les UAI de nature « école maternelle » seront sélectionnées ainsi que toutes les UAI d'une autre nature ayant la spécificité CLP.

Les données géographiques regroupent les zonages administratifs disponibles dans le code officiel géographique[13] (codes Insee associés à chaque territoire administratif : régions, départements) ainsi que le découpage en académie fait par le ministère et une géolocalisation précise de l'IGN.

On trouve aussi des caractéristiques de groupements et de rattachements. La notion de groupement est fortement liée aux politiques mises en place au sein du ministère. Par exemple, avec la politique d'éducation prioritaire, des regroupements d'UAI correspondant à des réseaux d'éducation prioritaire (REP) ont été créés dans Ramsese. La notion de rattachement est liée à l'organisation des UAI entre elles. Par exemple, une section d'éducation professionnelle sera généralement rattachée à un lycée qui en assure la gestion administrative.

Les données de gestion administrative sont les dates de création, de fermeture, de mise à jour d'une UAI.

Le répertoire permet donc de caractériser les UAI, leurs composantes ainsi que toutes les relations fondamentales existant entre les différentes structures et leur environnement pédagogique et administratif.

Les différents événements qui ont pu intervenir depuis la création d'une structure, tels qu'un changement de nature, un déménagement, une absorption ou une fusion, sont archivés, ce qui permet d'avoir un historique complet.

Les différents événements qui ont pu intervenir depuis la création d'une structure, tels qu'un changement de nature, un déménagement, une absorption ou une fusion, sont archivés, ce qui permet d'avoir un historique complet.

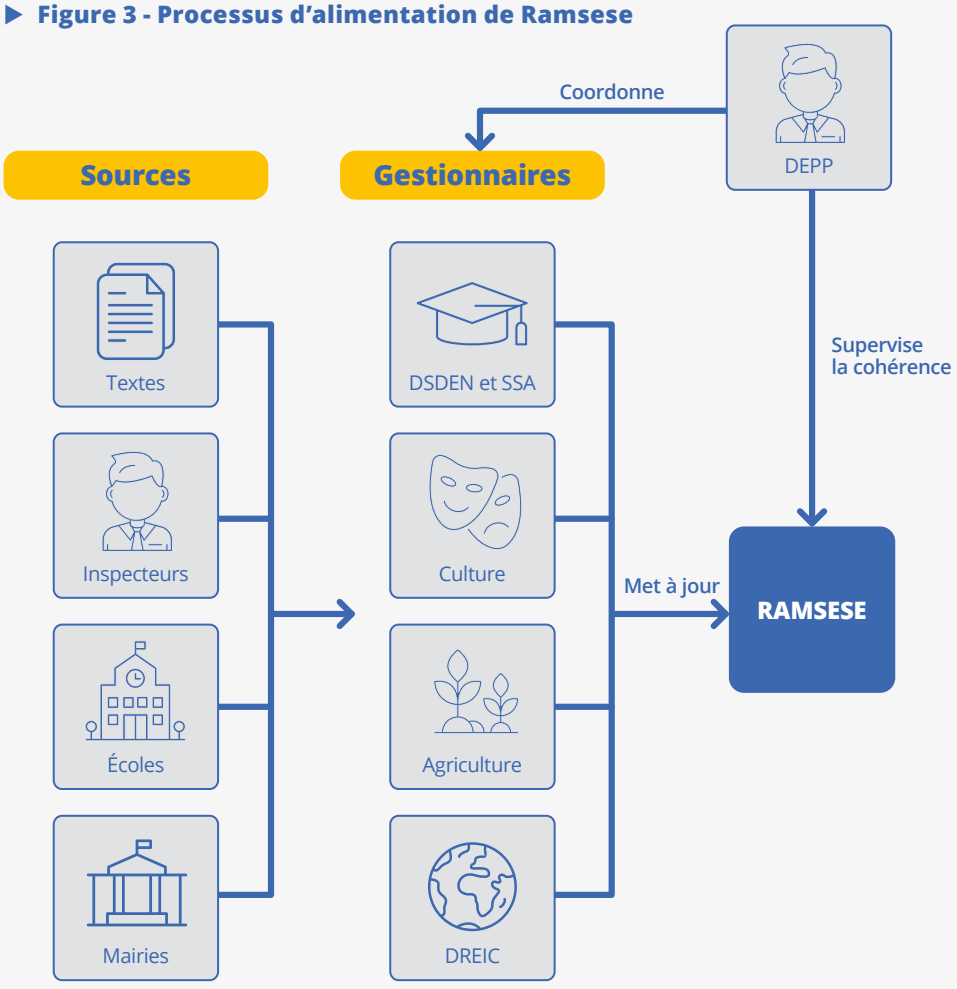
Afin de garantir la stabilité du répertoire et ne pas l'alourdir inutilement, certaines données sont reportées dans d'autres SI. Il s'agit notamment de données qui concernent les bâtiments, les équipements, les formations ou encore des données quantitatives telles que la superficie de l'UAI, le nombre d'élèves ou le nombre de classes. De même certaines informations présentes dans le répertoire ne délivrent qu'une information synthétique. Par exemple, la variable sur l'offre d'hébergement ou de restauration permet de savoir qu'une UAI propose une telle offre, sans préciser si elle est disponible au sein de l'UAI même ou externalisée, l'information plus détaillée relevant d'un autre système d'information.

12 CLP : Présence de classe(s) préélémentaire(s) dans une école élémentaire.
13 *https://www.insee.fr/fr/information/6800675.*

Pourquoi Ramsese a-t-il évolué ?

La collecte des données se fait à partir de nombreuses sources ***(Figure 3)***. Pour une grande part, l'alimentation se fait à partir des textes réglementaires. Mais les données peuvent aussi provenir directement des services ministériels ou des services déconcentrés, tels que les inspecteurs de l'éducation nationale chargés notamment de la carte scolaire de l'enseignement primaire. Ou encore directement des UAI elles-mêmes. Les mouvements au sein des UAI du secteur privé hors contrat doivent, quant à eux, faire l'objet d'une veille sur internet et les annuaires postaux, veille menée par les gestionnaires académiques ou départementaux de Ramsese.

Figure 3 - Processus d'alimentation de Ramsese



DSDEN : Direction des services départementaux de l'éducation nationale
SSA : Services statistiques académiques

Les réformes administratives ou la refonte d'un système d'information du ministère ont un impact sur la gestion de Ramsese : par exemple, celui de la loi pour la liberté de choisir son avenir professionnel sur la liste des centres de formation en apprentissage à immatriculer dans Ramsese. Les règles de gestion et les données gérées ont dû évoluer et s'adapter à la nouvelle réglementation. Avant cette réforme, le nombre de CFA était quasi stable : la réforme ayant assoupli et simplifié les règles de création des CFA, le nombre de demandes d'immatriculation pour ces structures, dont le référencement n'était pas toujours nécessaire aux systèmes d'information des ministères de l'Éducation nationale et de l'Enseignement supérieur et de la Recherche, a fortement augmenté. Bien que le numéro UAI soit un identifiant interne à ces deux ministères, la présence historique et obligatoire du numéro UAI sur le formulaire Cerfa du contrat d'apprentissage a également complexifié cette gestion.

Autre exemple, une contribution de vie étudiante et de campus (CVEC), instituée au profit des établissements d'enseignement supérieur, a été créée par l'article 12 de la loi du 8 mars 2018 relative à l'orientation et à la réussite des étudiants. Elle est destinée à favoriser l'accueil et l'accompagnement social, sanitaire, culturel et sportif des étudiants et à conforter les actions de prévention et d'éducation à la santé réalisées à leur intention. Cette contribution est acquittée auprès du centre régional des œuvres universitaires et scolaires (CROUS) par les étudiants inscrits en formation initiale dans un établissement d'enseignement supérieur public ou privé. Depuis la rentrée universitaire 2022-2023, délégation est donnée aux rectorats pour la constitution des listes des établissements assujettis à la CVEC. Ramsese constitue la source à partir de laquelle est établie la liste des établissements entrant dans le champ de la CVEC, et permet de s'appuyer sur la codification détaillée de la nature des établissements d'enseignement supérieur définie avec le ministère de l'Enseignement supérieur et de la Recherche, ou sur la codification des catégories juridiques. La liste ainsi constituée est adressée par les rectorats au réseau des œuvres universitaires et scolaires en charge de la gestion de la CVEC.

▶ Le travail des gestionnaires, indispensable pour la qualité de Ramsese

> ***Les gestionnaires du répertoire ont un rôle déterminant, la qualité du contenu reposant sur la qualité de leur travail, réalisé à partir d'échanges de documents, par mail ou téléphoniques avec les différents acteurs, en particulier les établissements, mais également les autres services de l'administration.***

Le répertoire ne dispose pas de processus de mise à jour complètement automatisé : il n'existe qu'une automatisation partielle. Ainsi, les gestionnaires du répertoire ont un rôle déterminant, la qualité du contenu reposant sur la qualité de leur travail, réalisé à partir d'échanges de documents, par mail ou téléphoniques avec les différents acteurs, en particulier les établissements, mais également les autres services de l'administration.

Les gestionnaires Ramsese interviennent en continu, tout au long de l'année scolaire sur les données du répertoire. Ils sont localisés dans les services statistiques académiques (SSA) des rectorats et dans les directions des services départementaux de l'éducation nationale (DSDEN) pour la gestion des UAI du 1er degré. Ils ont la responsabilité des données relevant de leur périmètre académique. À ce titre, ils effectuent la mise à jour

au fil de l'eau des informations sur les UAI de leur ressort géographique, telle que la mise à jour des CFA avant l'ouverture du portail apprentissage[14] ou la création et la mise à jour prévisionnelle[15] des établissements du 2nd degré public et privé sous contrat pour permettre le mouvement des personnels de direction.

La gestion des établissements éducatifs dépendant des ministères chargés de la culture ou de l'agriculture, ainsi que des établissements français à l'étranger, est confiée à des correspondants nationaux localisés dans les directions et ministères concernés.

La Depp procède, de façon ponctuelle, au chargement centralisé de certaines données, notamment celles relevant de dispositifs arrêtés au niveau ministériel, tels que l'identification des réseaux d'éducation prioritaire et le rattachement des écoles à ces réseaux.

Un cadrage national par la Depp est indispensable dans de nombreuses situations afin d'harmoniser les pratiques et de garantir la cohérence et la qualité des données. Créés par la loi pour une « école de la confiance », les pôles inclusifs d'accompagnement localisés (PIAL) ont pour principal objectif de permettre la coordination des moyens d'accompagnement humain (aides pédagogiques, éducatives et à terme thérapeutiques) au sein des écoles et des établissements scolaires de l'enseignement public et de l'enseignement privé sous contrat. Leur mise en place a aussi nécessité une analyse au niveau national afin de déterminer la meilleure manière de les prendre en compte dans le répertoire et a conduit à la transmission de consignes aux académies.

Les établissements publics locaux d'enseignement internationaux (EPLEI) qui dispensent un enseignement majoritairement au sein de sections internationales et/ou européennes, ont la particularité de regrouper au sein d'une entité administrative unique des classes du premier et du second degrés. Les systèmes d'information ne sont pas adaptés pour intégrer ces structures singulières, et n'évoluent pas toujours en parallèle de la réglementation ; ainsi, les académies consultent le niveau national afin de disposer de solutions spécifiques pour leur création dans le répertoire.

Chaque année, la Depp réunit à Paris durant une journée le réseau des gestionnaires du répertoire afin de faire un bilan des actions de l'année écoulée et pour échanger sur les consignes et règles d'immatriculation des UAI.

Par ailleurs, chaque année, la Depp réunit à Paris durant une journée le réseau des gestionnaires du répertoire afin de faire un bilan des actions de l'année écoulée et pour échanger sur les consignes et règles d'immatriculation des UAI.

Les gestionnaires du répertoire et la Depp peuvent également être mobilisés sur des opérations ponctuelles mobilisant des moyens importants au sein du ministère : à chaque nouvelle réforme sur les établissements par exemple, ou plus spécifiquement lors de certaines opérations telles que les élections professionnelles. Ainsi, l'organisation des élections professionnelles en 2022 a impliqué une

14 Le portail apprentissage permet aux CFA de transmettre les données individuelles de leurs apprentis et des formations suivies. En recensant les apprentis inscrits au 31 décembre de chaque année dans un CFA , l'enquête SIFA permet de connaître l'état de l'apprentissage en France et d'élaborer des prévisions de court terme.

15 Un établissement peut être créé, dans Ramsese, jusqu'à un ou deux ans avant son ouverture effective. Il y apparaît alors avec une date future d'ouverture.

logistique importante. La qualité des données du répertoire Ramsese était essentielle, tout particulièrement pour constituer le référentiel des adresses permettant l'envoi des notices de vote aux électeurs dans les écoles, les établissements et les services académiques.

Un travail au niveau national a été réalisé pour déterminer quelles étaient les structures à inclure dans l'extraction de Ramsese et pour constituer le fichier des adresses professionnelles des électeurs. La distribution dans les DOM et les COM a été réalisée par un canal propre à ces territoires.

La remise du matériel de vote sur site par les transporteurs nécessitait d'actualiser certaines adresses, en particulier lorsque celles-ci contenaient uniquement les mentions de boîte postale, de quartier ou de zone qui pouvaient s'avérer insuffisantes pour la distribution, dans un campus universitaire par exemple.

Pour mener ces actions, le réseau des gestionnaires Ramsese a été sollicité durant toute la phase de préparation de l'opération, avec des consignes nationales transmises afin de fiabiliser les extractions de données.

La volumétrie du répertoire Ramsese est limitée et les sources de mises à jour sont très diverses. Ainsi, il n'existe pas à ce jour de flux de données entrantes automatisé. La mise à jour des informations dans le répertoire résulte soit d'une action directe d'un gestionnaire, soit d'une mise à jour en masse décidée par un gestionnaire sur son académie ou par la Depp au niveau national. Les mises à jour en masse peuvent être réalisées par l'importation d'un fichier au format csv.

De plus, l'organisation territoriale, au plus près des acteurs de terrain et des établissements, peut parfois dégrader l'homogénéité des données du répertoire, par l'application de règles de gestion différentes sur le territoire, en fonction de contraintes locales. Par exemple les unités de formation par apprentissage (UFA) ne sont immatriculées que dans certaines académies. Pour éviter ces situations, la Depp, de par son rôle de pilotage en tant que maîtrise d'ouvrage du répertoire, transmet aux gestionnaires des consignes de gestion et veille à la cohérence de l'ensemble du répertoire.

L'exploitation annuelle des données issues des opérations statistiques sur les effectifs d'élèves ou d'étudiants, ou sur le parc immobilier, permet de renforcer la qualité des données du répertoire sur les champs concernés.

L'exploitation annuelle des données issues des opérations statistiques sur les effectifs d'élèves ou d'étudiants, ou sur le parc immobilier, permet de renforcer la qualité des données du répertoire sur les champs concernés.

Ainsi, chaque année, un appariement avec les données issues des constats de rentrée dans le premier et le second degrés permet de mettre à jour des informations sur les UAI concernées comme l'existence d'unités localisées pour l'inclusion scolaire (ULIS) au sein d'un établissement. Cela permet de garantir la fraîcheur de certaines données en particulier des spécificités.

▶ Des contrôles pour assurer la qualité du répertoire

Le niveau de qualité des données est une préoccupation constante dans la gestion du répertoire, d'autant plus que le contexte évolue constamment. Disposer d'un répertoire fiable, de qualité et exhaustif est primordial pour répondre aux besoins des utilisateurs.

Afin d'y répondre, des contrôles de cohérence des données présentes dans le répertoire sont réalisés sur les flux, lors de la mise à jour en continu pour éviter des erreurs de saisie. Par exemple, un établissement du secteur public n'aura pas de type de contrat, variable réservée aux UAI du privé.

Cependant, le niveau de qualité des données n'est pas le même selon le champ considéré. Certaines catégories d'UAI, plus fréquentes et indispensables aux systèmes d'information du ministère de l'Éducation nationale et de l'Enseignement supérieur et de la Recherche, sont mieux maîtrisées et connues par les gestionnaires du répertoire, voire disposent de circuits d'échanges d'informations mieux formalisés que d'autres. Les données sur les UAI des 1er et 2nd degrés sont de meilleure qualité car mieux contrôlées que celles sur les centres de formation en apprentissage.

Dans le cadre d'une convention, l'Institut national de l'information géographique et forestière (IGN) fournit plusieurs fois par an à la Depp les coordonnées de géolocalisation[16] des UAI, accompagnées d'indicateurs sur la qualité des adresses ce qui permet aux gestionnaires de réaliser des vérifications complémentaires.

Par ailleurs, des opérations ponctuelles de mise en qualité, portées au niveau national, sont effectuées régulièrement. Par exemple, des opérations de contrôle et de correction d'adresse, de vérification des doublons.

D'autres éléments peuvent être récupérés et corrigés dans le répertoire à la suite d'enquêtes ou de signalements liés à des usages plus administratifs des données du répertoire. Par exemple, les informations sur les cités scolaires sont complétées par les résultats de l'enquête Internat, cadre de vie (ICV). Des incohérences sont corrigées dans des périodes de validité erronées : elles sont signalées par l'application de gestion de ressources humaines SIRHEN qui gère le dossier administratif et financier des 1 100 000 agents du ministère ainsi que les moyens alloués aux établissements pour la préparation de la rentrée scolaire.

Enfin, à l'Insee, le pôle chargé de Sirene[17] pour le secteur public, ainsi que les directions régionales de l'Insee pour les structures relevant du secteur privé, constituent des partenaires qui fournissent l'immatriculation Siret des UAI, enregistrée dans Ramsese. Pour s'assurer de la qualité du Siret des structures publiques d'enseignement et administratives, le bureau chargé des nomenclatures et répertoires à la Depp pilote les demandes d'immatriculation auprès de Sirene ***(encadré 3)***.

16 Ces coordonnées sont directement exploitables par les utilisateurs des données du répertoire, sans qu'ils aient à géocoder les établissements à partir des adresses.
17 Sirene : Système national d'identification et du répertoire des entreprises et de leurs établissements.

▶ Encadré 3. Comparaison avec le répertoire Sirene

Contrairement au répertoire Ramsese, connu et reconnu de par son usage, le répertoire Sirene dispose d'une assise juridique (décret n° 73-314 du 14 mars 1973).

Les deux répertoires ne couvrent pas le même champ car Ramsese se limite aux établissements dont l'identification est utile aux systèmes d'information des ministères de l'Éducation nationale et de l'Enseignement supérieur et de la Recherche.

Le concept d'établissement est également différent entre les deux répertoires. Par exemple, dans Sirene, n'existe qu'un seul Siret pour un lycée et sa section d'enseignement professionnelle alors que dans Ramsese deux numéros UAI sont créés.

Par ailleurs, les règles de gestion divergent parfois. Dans le cas d'un déménagement d'un établissement, le répertoire Sirene fermera l'ancien établissement et en ouvrira un nouveau alors que dans Ramsese le numéro UAI reste le même dès lors que l'établissement reste au sein d'un même département.

Sirene fonctionne selon une logique d'événements, quand Ramsese s'appuie sur une veille permanente de la part de ses gestionnaires.

Face à l'ensemble de ces divergences de concepts et de règles de gestion, il est parfois difficile dans Ramsese d'avoir le bon numéro Siret en face du numéro UAI.

Afin de contribuer à l'amélioration de la qualité des deux répertoires, en 2022, la Depp et ses services académiques ont débuté une mise en cohérence des données enregistrées dans les deux répertoires sur le champ des établissements du 1er et du 2nd degrés publics avec le pôle secteur public de Sirene. L'opération a consisté à corriger les divergences, puis à mettre en place des échanges en flux afin de conserver cette cohérence. Des opérations qualité entre les deux répertoires seront ensuite mises en place afin de veiller au maintien de cette cohérence.

Les données actualisées à l'occasion de ces échanges sont la dénomination, l'adresse, la catégorie juridique, l'activité principale et le Siret des établissements.

▶ La diffusion du répertoire, de plus en plus *via* des API

Les données de Ramsese constituant la source de données sur les établissements pour les applications de gestion, de ressources humaines, de paie, d'affectation des élèves et des étudiants, une offre de service a été développée (*Biehl, 2016*) donnant aux applications la possibilité d'interroger le répertoire et de s'alimenter par l'intermédiaire d'une API (solution informatique permettant à des applications de communiquer et d'échanger des services ou des données).

Ainsi, les multiples systèmes d'information de ressources humaines font partie des applications qui utilisent les API proposées par Ramsese : leurs rénovations se font en s'appuyant sur le répertoire. La démarche interministérielle de modernisation des systèmes d'information relatifs aux ressources humaines et à la paye de l'État, se traduit par une réurbanisation des applicatifs de ressources humaines et la migration vers l'offre interministérielle de progiciel de gestion intégré RenoiRH[18]. Ramsese s'intègre parfaitement dans cette démarche. Ainsi, un agent de la fonction publique est affecté administrativement et opérationnellement sur une unité organisationnelle (UO). Ces unités sont gérées dans un répertoire (RSP-Référentiel des Structures Partagées), qui s'alimente en amont à partir de Ramsese et permet de gérer l'organisation au sein des établissements.

L'API est indispensable afin de faciliter l'accès des utilisateurs aux données du répertoire et contribue à la séparation de sa gestion et de sa diffusion.

18 RenoiRH (RENouveau des Outils Informatiques relatifs aux Ressources Humaines) est le système d'information de gestion des ressources développé par le CISIRH dans le but de mettre à disposition de ses partenaires une solution informatique mutualisée en matière de gestion des ressources humaines. RenoiRH s'adresse à tous les ministères ou entités de la Fonction Publique d'État.

Cela concourt à réduire la charge de conception, pour les fonctionnalités de gestion qui peuvent ainsi évoluer sans impact sur la diffusion, et pour les produits de diffusion en évitant de devoir faire du « sur mesure » pour chaque système d'information qui utilise les données. La communication et le partage de données sont réalisés sans connaître les détails de la mise en œuvre des autres systèmes d'information.

Cette séparation contribue à une simplification et à une amélioration du service rendu aux utilisateurs, objectif essentiel, auquel les API contribuent en exposant et rendant les données plus accessibles et mieux documentées.

Un gestionnaire d'API permet de centraliser les offres de service. L'offre de service de Ramsese se substitue ainsi progressivement au mode de mise à disposition historique par extraction sur un serveur sécurisé sous la forme de fichiers, au moyen d'un module de l'application de gestion.

Par ailleurs, une copie quotidienne de la base de gestion de Ramsese est mise à disposition des services académiques pour les besoins locaux. Elle permet d'alimenter des applications développées localement.

▶ Une visibilité toujours plus étendue

La Depp met à disposition sur Internet un outil de consultation et de cartographie des établissements du système éducatif français dénommé ACCÉ (Application de consultation et de cartographie des établissements). Il permet de consulter l'information disponible sur les établissements du répertoire Ramsese à partir d'une recherche *via* un formulaire et il intègre également une présentation cartographique des établissements, cette dernière proposant différentes vues (carte, satellite, etc.) et permettant une sélection d'établissements autour d'un point donné, afin par exemple d'obtenir rapidement une liste d'établissements dans un périmètre donné pour répondre à une situation de crise. ACCÉ est disponible sous une forme professionnelle, avec des fonctionnalités étendues, pour les acteurs de la sphère éducative ou pour le grand public[19].

> ***La Depp met à disposition sur Internet un outil de consultation et de cartographie des établissements du système éducatif français dénommé ACCÉ (Application de consultation et de cartographie des établissements).***

Dans le processus de pilotage des établissements, les données de Ramsese sur les écoles et les établissements du 2nd degré public et privé sous contrat sont extraites une fois par an pour alimenter l'application Archipel, qui constitue ainsi une photo des données en dehors de Ramsese pour une exploitation plus statistique. Archipel met à la disposition des directeurs et des chefs d'établissements des informations et des indicateurs sur leurs établissements.

19 La version professionnelle permet une recherche d'établissement à partir d'un formulaire intégrant un plus grand nombre de critères de recherche (une vingtaine au lieu de sept). Elle permet aussi de réaliser des extractions et d'enregistrer des recherches. Dans la version professionnelle les établissements ouverts et fermés sont présentés alors que seuls les établissements ouverts le sont dans la version grand public.

La loi pour une République numérique s'est traduite par la mise à disposition en *Open data* d'une partie des données du répertoire, ainsi que d'indicateurs calculés au niveau établissement, tels que les indicateurs de valeur ajoutée des lycées ou les effectifs d'élèves par niveau et nombre de classes par école. L'*Open data*[20] a considérablement accru la visibilité du répertoire et réduit les demandes d'informations d'utilisateurs externes.

Un répertoire qui s'adapte aux évolutions de l'éducation nationale et de la société

L'un des objectifs de la Depp est d'améliorer l'intégration du répertoire dans un environnement/écosystème urbanisé, en poursuivant notamment les échanges avec le répertoire Sirene ou en automatisant les échanges avec le ministère chargé de l'agriculture pour l'alimentation directe du répertoire à partir de leur système d'information.

Un recentrage au niveau national de certains traitements est parfois indispensable pour garantir l'homogénéité des traitements sur l'ensemble du territoire.

> ***L'un des objectifs de la Depp est d'améliorer l'intégration du répertoire dans un environnement/ écosystème urbanisé, en poursuivant notamment les échanges avec le répertoire Sirene.***

Certaines variables présentes dans le répertoire pour des raisons historiques seront supprimées du répertoire : par exemple, à court terme des données telles que les arrondissements ou les unités urbaines qui ne sont plus actualisées, notamment en raison de la disponibilité de ces informations dans une base de nomenclatures. La suppression des données qui ne sont pas des données de référence est un chantier à l'horizon plus lointain, en raison de la diversité des usages selon le champ (scolarité des élèves, ressources humaines, financier, statistique) et avec des implications non négligeables sur les systèmes d'information, qui s'alimentent à partir du répertoire, et leur nécessaire évolution coordonnée.

La démarche d'amélioration de la qualité du répertoire, qui s'est traduite récemment par une modélisation du processus d'immatriculation des établissements et par une analyse des risques, se poursuit avec la mise en place d'un plan d'action et d'une organisation adaptée permettant son suivi. L'objectif principal recherché est encore et toujours une amélioration de la qualité et du service rendu.

20 *https://data.education.gouv.fr/explore/dataset/fr-en-adresse-et-geolocalisation-etablissements-premier-et-second-degre/table/. https://data.education.gouv.fr/explore/dataset/fr-en-etablissements-fermes/table/.*

▶ Fondements juridiques

- Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique. In : Légifrance [en ligne]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000033202746&categorieLien=id*.
- Loi n° 2018-727 du 10 août 2018 pour un État au service d'une société de confiance. In : Légifrance [en ligne]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000037307624&categorieLien=id*.
- Loi n° 2019-791 du 26 juillet 2019 pour une école de la confiance. In : Légifrance [en ligne]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000038829065&categorieLien=id*.
- Loi n° 2018-771 du 5 septembre 2018 pour la liberté de choisir son avenir professionnel. In : Légifrance [en ligne]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000037367660&categorieLien=id*.
- Décret n° 2017-331 du 14 mars 2017 relatif au service public de mise à disposition des données de référence. In : Légifrance [en ligne]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000034194946&categorieLien=id*.

▶ Bibliographie

- ALVISET, Christophe, 2020. La troisième refonte du répertoire Sirene : trop ambitieuse ou pas assez. In : *Courrier des statistiques*. [en ligne]. 29 juin 2020. Insee, N° N4, pp. 101-121. [Consulté le 1er août 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4497083?sommaire=4497095*.

- BIEHL Matthias, 2016. *RESTful API Design. APIs your consumers will love*. [en ligne]. Août 2016. API-University Press, série API-University, Vol3. [Consulté le 1er août 2023]. Disponible à l'adresse : *https://api-university.com/wp-content/uploads/2015/05/api-design-toc.pdf*.

- BIZINGRE, Joël, PAUMIER, Joseph, RIVIÈRE, Pascal, 2013. *Les référentiels du système d'information - Données de référence et architectures d'entreprise*. Juillet 2013. Collection : InfoPro, Dunod. EAN : 9782100598748.

- EVAIN, Franck, 2020. Indicateurs de valeur ajoutée des lycées. Du pilotage interne à la diffusion grand public. In : *Courrier des statistiques*. [en ligne]. 31 décembre 2020. Insee. N° N5, pp. 74-94. [Consulté le 1er août 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/5008703?sommaire=5008710*.

- Les établissements scolaires, 2023. *https://www.education.gouv.fr/panorama-scolaire*.

- Repères et références statistiques. 2023. *https://www.education.gouv.fr/reperes-et-references-statistiques-2023-378608*.

- RIVIÈRE, Pascal, 2022. Qu'est-ce qu'un répertoire ? De multiples exigences pour un système complexe. In : *Courrier des statistiques*. [en ligne]. 29 novembre 2022. Insee. N° N8, pp. 52-71. [Consulté le 1er août 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6665186?sommaire=6665196*.

Peut-on se fier aux sondages empiriques ?

Pascal Ardilly[*]

Les enquêtes par sondage s'appuient sur un échantillon probabiliste ou sur un échantillon empirique. Dans l'approche empirique, la probabilité d'être enquêté, pour un individu donné, est généralement dépendante de la valeur de la variable que l'on collecte auprès de cet individu. Cela produit une erreur particulière appelée « biais de sélection ». Dans la méthode empirique dite « des quotas », on limite ce biais en structurant l'échantillon selon certaines variables expliquant le phénomène mesuré. Néanmoins, un biais subsiste si ces variables ne suffisent pas à en appréhender toute la variabilité. Pour justifier pleinement la méthode, on fait appel à une hypothèse de comportement des individus, appelée modélisation. D'autres méthodes de sélection empiriques existent, comme la méthode des unités-type – traduisant la perception que l'on a communément de la « représentativité » – ou l'échantillonnage de volontaires, particulièrement développé ces dernières années au travers des « Access panels ». Dans ce dernier cas, le biais de sélection peut être important, voire considérable. Malheureusement, on ne réduit pas le biais en augmentant la taille de l'échantillon. Deux exemples spectaculaires – l'un portant sur le taux de couverture vaccinale contre le coronavirus, l'autre sur les élections présidentielles de 1936 aux États-Unis – illustrent ce phénomène, dit « paradoxe des big data ».

Sample surveys are based on either a probability sample or a non-probability sample. In the non-probability approach, the probability of a given individual being included in the sample generally depends on the value of the variable collected from that individual. This produces a particular error known as 'selection bias'. In the non-probability 'quotas' method, this bias is limited by structuring the sample according to certain variables that explain the measured phenomenon. However, a bias remains if those variables fail to account its whole variability. In order to fully justify the method, one appeals to an assumed behaviour of individuals, known as modelling. Other non-probability selection methods exist, such as the purposive selection method – reflecting the common perception of 'representativeness' - or volunteer sampling, particularly developed in recent years through 'Access panels'. In this last case, the selection bias can be significant, even considerable. Unfortunately, the bias cannot be reduced by increasing the size of the sample. Two striking examples – one relating to the vaccine uptake rate against the coronavirus, the other to the 1936 presidential elections in the United States – illustrate this phenomenon, known as the 'big data paradox'.

[*] Expert, Département des méthodes statistiques, DMCSI, Insee,
pascal.ardilly@insee.fr

Le statisticien est par nature confronté à des problèmes d'estimation. Il cherche en effet à approcher au plus près différentes « grandeurs » dont personne ne connaît *a priori* la valeur exacte. Ces grandeurs, qualifiées de « paramètres d'intérêt », sont définies dans une population (individus, entreprises, articles de vente, etc.) généralement de très grande taille, à partir de variables individuelles quantitatives ou qualitatives que l'on appelle « variables d'intérêt ». La plupart des paramètres sont des moyennes ou se construisent à partir de moyennes (totaux, proportions, dispersions). Par exemple, on s'intéresse au revenu moyen des personnes résidant en Bretagne à une date donnée, au chiffre d'affaires total annuel des boulangeries parisiennes, ou encore à l'évolution des prix moyens de l'alimentaire entre deux mois consécutifs. Dans ce contexte, la statistique la plus précise relève d'une collecte exhaustive, donc d'un recensement. Le coût d'un recensement étant généralement dissuasif, on utilise, en pratique, des techniques de sondage consistant à restreindre la collecte à une sous-population, l'échantillon. Les techniques d'échantillonnage distinguent deux grandes familles : les sondages probabilistes et les sondages empiriques (*Ardilly & Lavallée, 2017*). Ces derniers s'appuient parfois sur des échantillons de volontaires – les fameux « *Access panels*[1] » – et utilisent très souvent la célèbre méthode de sondage « par quotas ». Ils séduisent du fait de la rapidité de la mise en œuvre et de l'économie de moyens, tandis que le respect des quotas a un côté rassurant. Certes, utiliser un sondage est toujours une prise de risque, mais avec ces méthodes, une prudence toute particulière est requise. Pourquoi, et que peut-on leur reprocher ?

Dans cet article, on cherche à répondre à cette question en insistant sur les erreurs que produisent le plus souvent les échantillonnages empiriques. En particulier, on ne peut pas les réduire en se contentant d'augmenter la taille de l'échantillon. En revanche, elles peuvent disparaître si on accepte certaines hypothèses portant sur les variables d'intérêt considérées. Un changement de paradigme permet de construire un cadre théorique généralement utilisé pour l'expliquer.

▶ Sondage probabiliste et sondage empirique : un schisme méthodologique

Dans la réalisation d'une enquête par sondage, le statisticien distingue quatre étapes : l'échantillonnage, la collecte, l'estimation, et le calcul de précision. L'échantillonnage – sauf cas très particulier des unités-type (cf. *infra*) – constitue une source majeure d'aléa : il s'agit de désigner les unités auprès desquelles on va collecter l'information. L'étape suivante est celle de la collecte, qui doit respecter de nombreuses consignes et qui produit presque toujours de la non-réponse. La non-réponse introduit une seconde source d'aléa, qu'il est souhaitable de minimiser. Vient ensuite l'estimation, étape calculatoire qui agrège de façon adéquate les données individuelles collectées afin d'estimer le paramètre d'intérêt. L'opération s'achève par la mesure d'erreur, communément appelée « calcul de précision ».

Dans une population donnée, la sélection d'un échantillon quelconque peut être aléatoire ou non, et si elle est aléatoire, on peut être capable ou non de calculer la probabilité d'obtenir l'échantillon en question. Le contexte dans lequel l'échantillonnage permet une maîtrise des probabilités de sélection est celui de l'échantillonnage probabiliste. Par « maîtrise », il faut

1 Il s'agit de bases rassemblant un grand nombre de personnes volontaires pour participer, sous conditions, à des enquêtes portant sur des thèmes variés.

comprendre que la méthode d'échantillonnage mise en œuvre autorise un calcul théorique de ces probabilités. Dans le cas contraire, on a affaire à un échantillonnage empirique.

Les fondements de l'échantillonnage probabiliste attribuent un rôle central à la base de sondage et à l'algorithme de sélection. La base de sondage est la liste exhaustive et sans double compte des individus formant la population d'intérêt. Sur cette base, on applique un algorithme, c'est-à-dire une règle objective (sans intervention humaine) et entièrement codifiée, de sélection aléatoire des individus de l'échantillon. Dans ces conditions, on peut connaître, pour chaque individu de la base de sondage, la probabilité qu'il appartienne à l'échantillon. On en déduit un poids de sondage, facteur déterminant qui traduit le nombre d'unités de la population que l'individu échantillonné représente. On multiplie le poids de sondage de chaque individu par ses réponses au questionnaire, et la résultante est sommée sur l'échantillon pour produire les estimations attendues. En pratique, on est confronté à la non-réponse, que l'on traite généralement en corrigeant les poids. L'ampleur numérique de cette correction est importante : elle consiste, dans l'approche la plus fruste, à multiplier les poids par l'inverse de la proportion de répondants dans l'échantillon tiré. On ajoute presque toujours une étape finale dite de redressement (ou de calage) qui consiste à modifier de nouveau les poids – de manière marginale cette fois – pour améliorer la qualité de l'estimation (*Ardilly, 2006* ; *Lohr, 2021*).

> ***L'échantillonnage empirique, lorsqu'il est aléatoire, relève d'une pratique de sélection qui ne permet pas le calcul de la probabilité de sélection des échantillons ni celle des individus de la population.***

À l'inverse, l'échantillonnage empirique, lorsqu'il est aléatoire, relève d'une pratique de sélection qui ne permet pas le calcul de la probabilité de sélection des échantillons ni celle des individus de la population. Non pas qu'il s'agisse d'une impuissance mathématique des statisticiens, mais parce que cette sélection résulte par nature d'un processus en partie subjectif. En pratique, on confie ce rôle à des enquêteurs ou on s'en remet au volontariat des participants, perdant ainsi la maîtrise des probabilités de tirage : on peut parfaitement imposer et superviser la façon dont fonctionne un programme informatique de sélection, mais ce contrôle n'est plus possible lorsque la sélection résulte en partie du comportement humain !

Les enquêtes par quotas, méthodologie standard du sondage empirique

La sélection empirique la plus commune est faite « sur le terrain » par des enquêteurs, en face à face ou par téléphone, en s'appuyant sur un ensemble de consignes qui tendent à reproduire autant que possible un mécanisme probabiliste uniforme où tous les individus ont exactement la même chance d'être tirés. L'objectif consiste à rendre cette sélection aléatoire autant que possible, en évitant de privilégier certaines catégories de population. Une façon naturelle de réduire ce risque passe par le respect de quotas – d'où le nom de « méthodes par quotas ». Il s'agit de définir des sous-populations à partir des modalités d'un jeu de variables qualitatives ou quantitatives (les « variables de quotas ») découpées en tranches, et de demander à chaque enquêteur de constituer un échantillon dont les

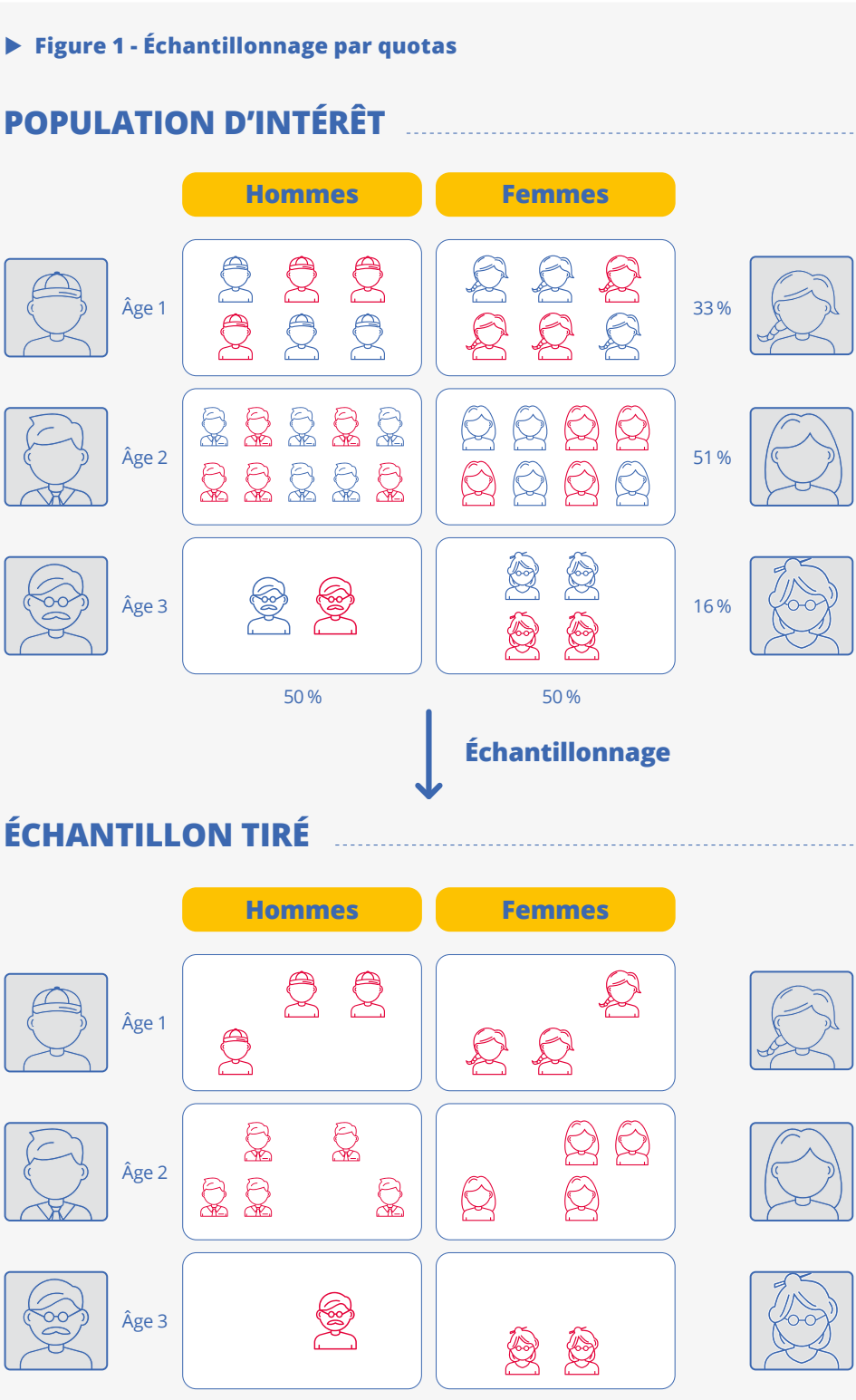
effectifs appartenant à ces différentes sous-populations – les quotas – soient égaux à ce que produirait « en moyenne » un échantillonnage probabiliste à probabilités égales (dit « équiprobable »). Par exemple, on demande à ce que l'échantillon empirique comprenne moitié d'hommes et moitié de femmes, parce qu'il s'agit de la structure par sexe « moyenne » résultant d'un échantillonnage aléatoire équiprobable. C'est aussi la vraie structure de la population française selon cette variable. On évitera ainsi qu'un enquêteur ne produise un échantillon trop déséquilibré sur la variable sexe, ce qui éloignerait le processus de sélection d'un processus équiprobable. La plupart du temps, on impose un jeu de quotas construits en croisant plusieurs variables – par exemple simultanément le sexe, l'age et le diplôme ***(figure 1)***. Ainsi, on dispose au final d'un échantillon qui a les caractéristiques d'une photo-réduction de la population d'intérêt en ce qui concerne les variables de quotas. La méthode permet de se passer d'une base de sondage : c'est un avantage considérable, car les bases sont souvent coûteuses à acquérir et elles peuvent être couvertes en amont par la confidentialité des données individuelles.

▶ Face aux échantillonnages probabilistes, un double handicap quant à la qualité

Le respect des quotas est une condition nécessaire mais non suffisante pour qu'on puisse assimiler l'échantillonnage empirique à du tirage aléatoire équiprobable. Pour apprécier la nature du risque, imaginons une enquête sur l'emploi du temps et imposons des quotas construits à partir du sexe et de l'âge. L'échantillon final respectera donc la structure de la population d'intérêt selon ces deux critères. Les enquêteurs, sur le terrain, en mode face-à-face ou en mode téléphone, vont sélectionner des individus consentants, et *a priori* ils vont le faire durant la journée, aux horaires durant lesquels la plupart des actifs exercent leur profession. Ainsi, il est fort probable que l'échantillon soit déficitaire en certaines catégories d'actifs, ceux que l'on peut contacter tôt le matin, tard le soir, voire parfois seulement la nuit. À l'inverse, on « surcharge » l'échantillon en personnes sans emploi, plus faciles à contacter en journée. Évidemment, dans le cas présent, la ficelle est assez grosse, et on limitera ce risque en agissant ici dans trois directions au moins : on enrichira les quotas en ajoutant au moins une variable liée à l'activité, on demandera aux enquêteurs d'élargir leurs horaires de collecte en semaine et de travailler le week-end, et on étendra la période de collecte. On enrichira les quotas… à condition qu'on puisse déterminer les variables en rapport suffisamment étroit avec l'emploi du temps, à condition qu'on connaisse la structure de la population selon les modalités de ces variables, à condition aussi que les contraintes générées par l'accumulation des quotas ne rendent pas la collecte insupportable pour les enquêteurs. Par ailleurs, on ne pourra probablement élargir les horaires de collecte que jusqu'à un certain point. Ainsi, même si on parvient à réduire sensiblement les risques, il n'y aura aucune garantie qu'il ne reste pas une ou plusieurs variables cachées explicatives de l'emploi du temps mais gérées (en toute inconscience) de manière déséquilibrée par le réseau d'enquêteurs. *A contrario*, l'échantillonnage probabiliste à probabilités égales dispose

> ***Il n'y aura aucune garantie qu'il ne reste pas une ou plusieurs variables cachées explicatives de l'emploi du temps mais gérées (en toute inconscience) de manière déséquilibrée par le réseau d'enquêteurs.***

▶ **Figure 1 - Échantillonnage par quotas**



d'un avantage fort en éliminant ce type de risque, car il produit un échantillon équilibré « en moyenne » sur n'importe quelle variable.

Dans la pratique, le phénomène de non-réponse joue comme une phase d'échantillonnage supplémentaire – non-contrôlée par le statisticien – qui vient réduire la qualité des estimations. Les deux types d'échantillonnage sont affectés par la non-réponse mais la collecte qui suit l'échantillonnage probabiliste impose une multiplication des tentatives de contact auprès de chaque individu échantillonné impossible à joindre, cela jusqu'à atteindre un seuil de renoncement. En revanche, dans un sondage empirique, un individu échantillonné mais non-répondant est définitivement ignoré si sa variable d'intérêt n'est pas immédiatement collectée. L'approche empirique prend un avantage très substantiel en termes de coût, mais à l'issue de la collecte, la non-réponse a sensiblement moins déséquilibré un échantillon probabiliste qu'elle ne le fait avec un échantillon empirique.

Ce phénomène insidieux est souvent ignoré parce que la non-réponse est occultée dans les approches empiriques : non quantifiée, il n'en est à peu près jamais fait état et elle semble même inexistante pour les utilisateurs des données puisque l'échantillon final a toujours par construction la taille initialement requise. Sur ce point, l'échantillonnage probabiliste offre un avantage comparatif parce qu'il est possible d'estimer par modèle les probabilités de réponse et d'apporter des corrections qui limitent l'effet négatif de la non-réponse ; néanmoins, les imperfections (inévitables) de cette phase corrective produisent *in fine* un biais d'estimation.

▶ Les erreurs dans les enquêtes par sondage

Différents types d'erreur affectent les enquêtes par sondage (*Blog Insee, 2022*). On peut en distinguer quatre.

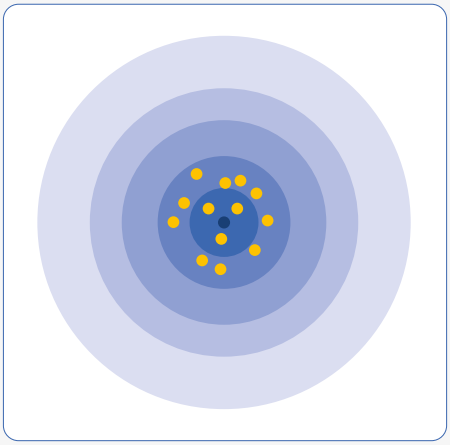
La première erreur est celle du *défaut de couverture*, qui survient lorsque certains individus de la population d'intérêt ne peuvent pas être échantillonnés. Dans les enquêtes probabilistes, c'est dû à un éventuel défaut d'exhaustivité de la base de sondage. Dans les enquêtes empiriques, en l'absence de base de sondage, ce type d'erreur est plus difficile à cerner, mais on en imagine facilement les manifestations. En particulier, pour une collecte en face à face auprès de personnes physiques, il est fort probable que certains individus soient consciemment rejetés par l'enquêteur – parce qu'ils sont par exemple d'accès difficile ou simplement dissuadent *a priori* l'enquêteur de par leur apparence ou leur comportement peu engageant. En effet, lorsqu'on a le choix de l'enquêté, on a naturellement tendance à se porter vers des individus qui semblent « faciles » à aborder.

> ***Si la moyenne de toutes les estimations obtenues à partir de tous ces échantillons diffère de la vraie valeur, on dit qu'il y a un biais d'estimation.***

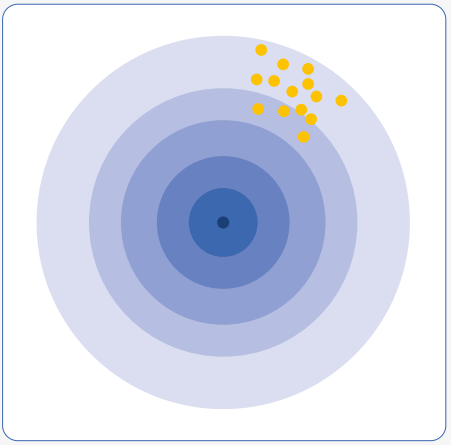
La seconde erreur est l'*erreur d'échantillonnage*. Cette erreur traduit le fait que les estimations produites sont sensibles à la composition de l'échantillon et qu'elles ne sauraient donc coïncider avec la valeur « exacte » du paramètre d'intérêt. Deux composantes sont identifiables : le biais et la variance ***(figure 2)***.

▶ Figure 2 - Biais et variance

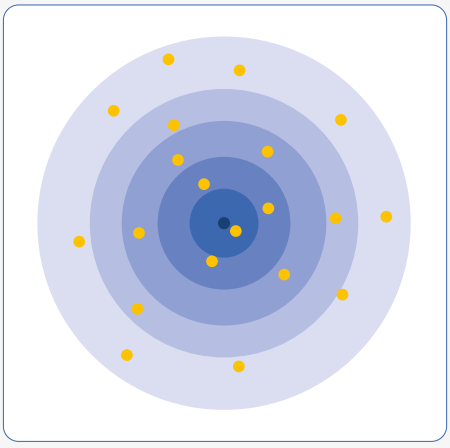
SCÉNARIO 1
Pas de biais et faible variance



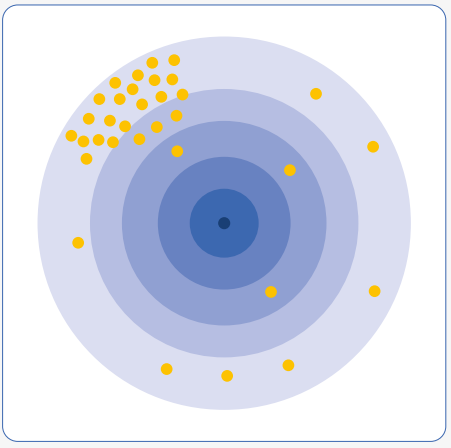
SCÉNARIO 2
Biais et faible variance



SCÉNARIO 3
Pas de biais et forte variance



SCÉNARIO 4
Biais et forte variance



Légende
La cible couvre l'ensemble des estimations possibles.
Le centre de la cible représente la vraie valeur.
Chaque point jaune matérialise une estimation et correspond à un échantillon particulier.

Supposons que l'on tire un grand nombre d'échantillons, selon une méthode donnée, et que chaque échantillon produise son estimation. Si la moyenne de toutes les estimations obtenues à partir de tous ces échantillons diffère de la vraie valeur, on dit qu'il y a un biais d'estimation. Cela peut provenir par exemple de déséquilibres systématiques dans la composition de l'échantillon. Par ailleurs, l'hétérogénéité des différentes estimations peut être formalisée au travers d'un indicateur appelé variance d'échantillonnage : plus il y a de dispersion des estimations, plus la variance est grande, et moins la qualité d'estimation sera bonne. Une grande variance signifie donc que l'estimation dépend fortement de l'échantillon, ce qui n'est évidemment pas souhaitable. De façon générale, la variance d'une estimation dépend de trois facteurs : le système de pondération utilisé – qui essentiellement reflète la méthode d'échantillonnage adoptée – la taille de l'échantillon tiré, et l'ampleur de la non-réponse. Quelle que soit la méthode d'échantillonnage, la variance diminue lorsque la taille de l'échantillon répondant augmente.

La troisième erreur, déjà abordée, est due à la *non-réponse*, qui provient essentiellement de problèmes de localisation (dans les enquêtes probabilistes en face-à-face), de comportements de refus de l'enquêté et d'impossibilité à joindre l'enquêté (dans les enquêtes par téléphone par exemple).

Pour terminer, citons l'*erreur d'observation*, qui survient chaque fois que l'information collectée n'est pas conforme à la réalité (par exemple à la suite d'une mauvaise déclaration de l'enquêté – consciemment ou non – ou d'erreur de saisie de l'enquêteur, voire d'une mauvaise formulation des questions). Sur ce type d'erreur, il n'y a pas lieu de penser que la nature de l'échantillonnage ait un effet particulier. L'erreur d'observation est distinguable des précédentes en ce sens où elle traduit une véritable « faute » humaine : les autres erreurs sont plutôt de la nature d'imperfections dont la responsabilité revient au contexte, ou tout simplement au hasard.

▶ Le problème spécifique du biais dans les enquêtes empiriques

On critique parfois la méthode des quotas parce qu'elle produit des estimations biaisées. L'origine fondamentale de ce biais est l'impossibilité de construire un système de pondération qui corresponde à l'échantillonnage pratiqué et à la non-réponse (on rappelle ici que la formulation du poids dépend théoriquement de la probabilité de sélection). De fait, puisqu'on ne maîtrise pas les probabilités de sélection et que l'on ne sait rien de la non-réponse, les estimations issues des échantillonnages empiriques sont toujours construites à partir de poids constants. Ainsi, pour estimer des moyennes dans la population d'intérêt, on calcule des moyennes simples dans l'échantillon : c'est faute de pouvoir faire autrement, et c'est là le point faible majeur des enquêtes par quotas ! Formellement, on peut montrer ***(annexe)*** que l'ampleur du biais est conditionnée par la corrélation existante entre la variable d'intérêt et la probabilité de sélection des individus de la population. C'est assez intuitif : si on reprend l'exemple de l'enquête Emploi du temps, on peut craindre que la probabilité d'enquêter un individu soit plus forte si cet individu travaille moins. Un échantillonnage probabiliste n'aura pas ce défaut, parce que le processus de sélection ne sera en rien influencé par la nature de l'activité de l'enquêté – ou si

> ***L'ampleur du biais est conditionnée par la corrélation existante entre la variable d'intérêt et la probabilité de sélection des individus de la population.***

tel est le cas, ce sera d'une manière parfaitement contrôlée. Néanmoins, dans une enquête probabiliste la non-réponse génère *a priori* un biais ; il est donc important de chercher à avoir le taux de réponse le plus élevé possible.

Annuler la corrélation entre deux variables qui n'ont aucune raison de ne pas être corrélées de manière « naturelle » revient à créer les conditions pour rendre constante l'une des deux variables en question. La première façon d'y parvenir est d'agir afin que la probabilité de sélection soit constante. C'est précisément ce que l'échantillonnage empirique ne parvient pas à faire de manière rigoureuse en pratique, mais on cherche naturellement à se rapprocher de cette situation idéale − qui est évidemment celle de l'échantillonnage aléatoire équiprobable. C'est pourquoi il est absolument essentiel de donner des instructions aux enquêteurs pour rendre ce processus aléatoire au maximum, afin que la collecte soit la moins sélective possible. Ce sont en pratique des instructions de bon sens, consistant à parcourir des zones variées, ne pas interroger son voisinage exclusivement, varier les horaires de contact ainsi que les jours de collecte… La seconde façon d'annuler la corrélation, c'est de faire en sorte que la variable d'intérêt soit constante. Cette piste semble absurde de prime abord, mais c'est néanmoins celle qui produit la meilleure justification de la méthode des quotas : sa philosophie est portée par l'approche par modèle, exposée ci-après.

La meilleure façon de justifier statistiquement les enquêtes par quotas

La spécificité des enquêtes par sondage tient à la nature de l'aléa du sondage. Dans son approche historique, qui est aussi l'approche par défaut adoptée de nos jours, c'est en effet la composition de l'échantillon qui est aléatoire, et non les variables d'intérêt. On considère ainsi que les données collectées sont déterministes, c'est-à-dire fixées, connues et fournies par l'enquêté (sauf en cas d'erreur d'observation). Et dans ce cas, l'estimation est entachée de biais et de variance. En parallèle, la théorie statistique a développé une autre approche – dite stochastique − qui traite les données collectées auprès d'un individu donné comme la résultante d'un phénomène aléatoire, exactement comme si une loterie avait décidé de leurs valeurs. C'est une autre façon d'aborder la question de l'estimation de paramètres, qui offre un cadre théorique confortable pour justifier l'approche par quotas. L'idée sous-jacente consiste à relier la valeur de la variable d'intérêt aux modalités des variables de quotas, la première étant une fonction simple des secondes, en la circonstance une somme de valeurs caractérisant chaque modalité. Par exemple, on postulera que le temps passé aux tâches domestiques est une fonction du sexe (homme / femme), de l'âge (enfant / âge actif / personne âgée) et du statut d'activité (actif occupé / autre), en retenant ces trois variables et leurs modalités comme variables de quotas pour constituer l'échantillon. Ainsi, en connaissant le sexe, la tranche d'âge et le statut d'activité, on est « presque » en mesure de déterminer le temps passé aux tâches domestiques. Dans ces conditions, il est assez intuitif que seules ces variables sont importantes pour déterminer la composition de l'échantillon : puisque les autres critères ne comptent pas, ou très peu, un éventuel déséquilibre sur ces derniers n'aura pas de conséquence sur l'estimation. En la circonstance, il faut certes que les proportions de femmes, d'enfants, de personnes âgées, et d'actifs occupés au sein de l'échantillon soient celles de la population, mais pour le reste peu importe : si l'échantillon est constitué par ailleurs essentiellement de ruraux peu diplômés et célibataires, y compris de manière grossièrement excessive, il ne faudra pas s'en émouvoir puisque ni le type de commune, ni le diplôme, ni l'état matrimonial ne sont des critères qui influent sur le temps passé aux tâches domestiques.

> ***La façon de sélectionner l'échantillon – dès lors qu'il respecte les contraintes de quotas – n'a pas d'importance.***

Un tel état d'esprit fait donc entière confiance à une relation entre variables, ce qui constitue une hypothèse simplificatrice de la réalité : exactement ce qu'on dénomme un « modèle » en statistique.

Les modèles stochastiques proposent par ailleurs un cadre très pratique pour calculer des erreurs (*Deville, 1991*). Il ne s'agit plus d'erreurs d'échantillonnage mais d'erreurs d'une autre nature puisque l'aléa est celui qui affecte les valeurs des variables d'intérêt. On part toujours du principe que le modèle est juste en ce sens où la valeur de la variable d'intérêt est en moyenne égale à la somme des valeurs caractérisant les modalités des variables de quotas ***(encadré 1)***. Il s'ensuit un principe fondamental : la façon de sélectionner l'échantillon – dès lors qu'il respecte les contraintes de quotas – n'a pas d'importance, et comme corollaire direct, il apparaît dans l'approche stochastique qu'on n'a pas besoin de pondérer les individus échantillonnés (*Smith, 1983*). L'utilisation d'une moyenne simple pour estimer une vraie moyenne inconnue trouve donc là sa pleine justification.

▶ Encadré 1. La justification par modèle de la méthode des quotas

L'utilisation d'un modèle – donc d'une hypothèse de comportement – permet de s'affranchir de la composition de l'échantillon. Un nouveau paradigme particulièrement pratique se présente.

Pour simplifier, le contexte met ici en jeu deux variables de quotas : le sexe et l'activité (actif ou non). La modalité i du sexe contribue à former la quantité Y_k – par exemple le temps hebdomadaire consacré aux tâches ménagères – *en moyenne* à hauteur a_i et la modalité j de l'activité y contribue pour une valeur b_j *en moyenne*. Soit pour tout individu k de la cellule (i,j) :

$$Y_k = a_i + b_j + \epsilon_k$$

où ϵ_k traduit le fait que la connaissance de la cellule (i,j) ne suffit pas à déterminer numériquement Y_k, en tout cas pas précisément car si les variables de quotas sont bien choisies, alors ϵ_k aura vocation à être petit (on parle de 'résidu'). La variable ϵ_k est en moyenne nulle, ce qui constitue l'hypothèse fondamentale faite ici, justifiant le terme de « modèle ». De fait, la situation idéale (ϵ_k petit) est celle où, connaissant le sexe et le statut d'activité, on peut « presque parfaitement » prédire le temps consacré par tout individu aux tâches ménagères.

Dans ce modèle dit « additif simple », Y_k est une variable aléatoire, tout comme ϵ_k, mais les termes a_i et b_j ne sont pas aléatoires. La moyenne définie par rapport à l'aléa du modèle est appelée « espérance ».

La taille d'échantillon dans la cellule (i,j) est $n_{i,j}$. On note $n_{i,.}$ et $n_{.,j}$ (respectivement $N_{i,.}$ et $N_{.,j}$) les tailles d'échantillon (respectivement les tailles de population) marginales, qui sont aussi les 'quotas'. Le respect des quotas s'avère essentiel, puisqu'il s'agit d'imposer

$$\frac{n_{i,.}}{n} = \frac{N_{i,.}}{N} \quad \text{et} \quad \frac{n_{.,j}}{n} = \frac{N_{.,j}}{N}$$

pour tout (i,j). En la circonstance, cela impose que les proportions respectives d'hommes et de femmes dans la population et dans l'échantillon soient égales. Et il en est de même pour les proportions associées aux deux modalités activité / non-activité distinguées. On montre que dans ces conditions, et quelque soit l'échantillon tiré (ce qui est essentiel !), en espérance la différence entre la moyenne simple $\bar{y}$ dans l'échantillon et la moyenne vraie dans la population complète $\bar{Y}$ est nulle, traduisant l'absence de biais de $\bar{y}$ dans ce contexte spécifique de modélisation.

On remarque que le modèle standard des quotas consiste à considérer comme constante la variable d'intérêt – à de petits écarts aléatoires près – au sein de chaque sous-population définie par le croisement des modalités des variables de quotas. Il s'agit d'une hypothèse fortement contraignante, d'autant que les variables de quotas sont généralement en nombre limité et que la forme de leur relation avec la variable d'intérêt doit être spécifique (en l'occurrence additive).

Mais ce cadre crée *de facto* des corrélations (à peu près) nulles entre probabilité de sélection et variable d'intérêt soit, comme signalé précédemment, les conditions d'un biais d'échantillonnage (très) faible[2] – ainsi la boucle est bouclée !

Puisqu'un modèle est la formalisation d'une hypothèse, le risque est évidemment celui d'une hypothèse fausse, qui aurait pour sanction immédiate un biais d'estimation au sens de l'aléa du modèle.

▶ Échantillonnage probabiliste ou échantillonnage par quotas ?

C'est évidemment une question opérationnelle centrale. Lorsqu'on ne dispose pas de base de sondage, nécessité fait loi, car il n'est pas possible de procéder à un échantillonnage probabiliste. C'est une situation assez fréquente, parce que les bases de sondage sont très souvent des fichiers confidentiels constitués et détenus par des organismes publics, qu'il n'est pas possible de diffuser. Pour les personnes physiques, c'est par exemple le cas du Recensement de la population, ou des fichiers fiscaux. Pour les entreprises en revanche, le répertoire Sirene est accessible à tout utilisateur. Il faut ensuite tenir compte des budgets d'enquête : l'enquête probabiliste est sensiblement plus coûteuse puisqu'elle impose des unités échantillonnées. Cela nécessite davantage de tentatives de contact, et des coûts de déplacement plus élevés si le mode de collecte est le face-à-face.

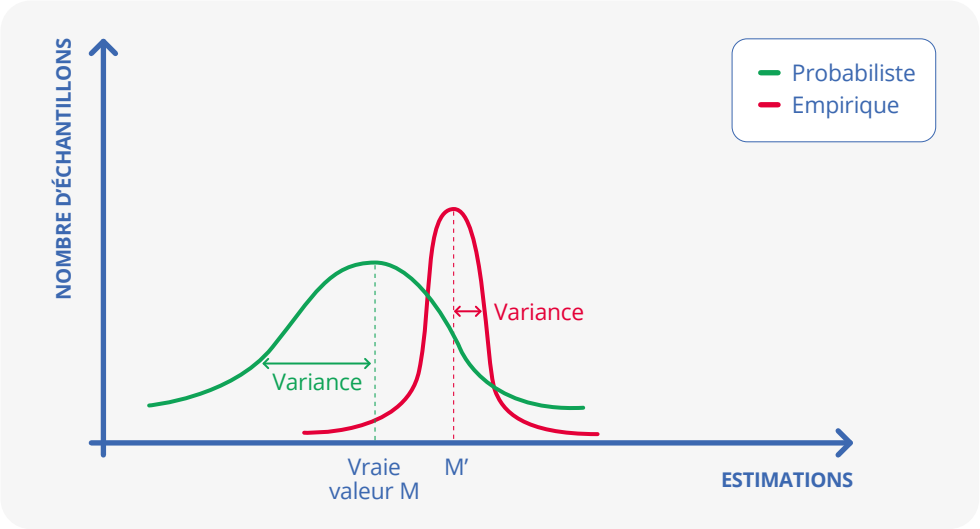
> ***Le biais empirique ne se réduit pas quand la taille de l'échantillon augmente.***

Au-delà de ces éléments logistiques et budgétaires, les considérations de qualité statistique contribuent à la prise de décision (*Mac Innis, 2018* ; *Brüggen, 2016* ; *Shirani-Mehr, 2018* ; *Forster, 2001*). Par construction, et c'est un atout des méthodes de quotas, le respect des quotas restreint la diversité des échantillons et cela se traduit par une réduction de la variance d'échantillonnage. En revanche, cette fois au désavantage des méthodes de quotas, le biais est un facteur pénalisant que n'a pas l'échantillonnage probabiliste si on fait abstraction de la non-réponse (et des défauts de couverture), et on peut vérifier, hélas, que le biais empirique ne se réduit pas quand la taille de l'échantillon augmente. On se trouve conduit à un arbitrage entre biais et variance. Si on s'intéresse à l'erreur d'échantillonnage totale, en tenant compte à la fois du biais et de la variance, il apparaît que les échantillonnages empiriques ne sont pas recommandés pour les gros échantillons. En revanche, les petits échantillons empiriques peuvent être préférables à un échantillon aléatoire équiprobable parce que l'avantage en termes de variance dépasse le handicap du biais ***(encadré 2)***. Ce principe est conforme à ce qu'on constate en pratique : les échantillons empiriques dépassent rarement 2 000 unités, et leurs tailles se situent même assez souvent aux alentours de 1 000, voire moins.

2 Un principe analogue concerne la correction de la non-réponse dans les enquêtes : un biais survient lorsqu'il subsiste une corrélation entre la variable d'intérêt et la participation à l'enquête une fois que l'on a conditionné par certaines variables explicatives (jouant un rôle équivalent aux variables de quotas).

▶ Encadré 2. Comparaison des sondages probabiliste et empirique en matière d'erreur d'échantillonnage

Considérant le seul critère de précision statistique, les deux types de sondage ici considérés ont des comportements différents, en particulier au regard de l'effet de la taille de l'échantillon. En voici les principales caractéristiques.

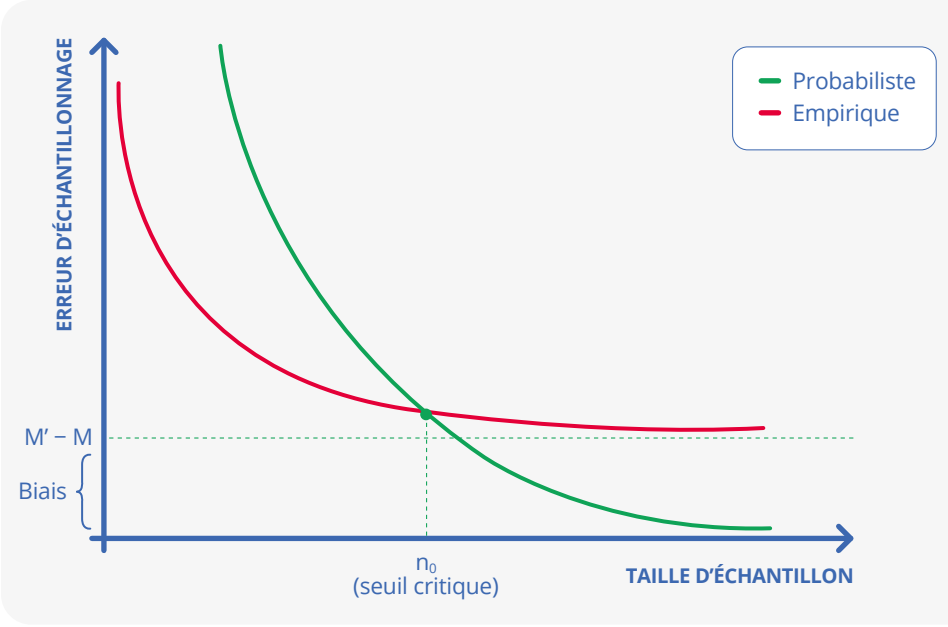


La **figure 1** compare la distribution des estimations issues respectivement d'un échantillonnage empirique (courbe rouge) et d'un échantillonnage probabiliste simple et équiprobable (courbe verte), pour une taille d'échantillon donnée et plutôt petite – par exemple quelques centaines d'unités. Ces courbes ont des allures de courbe de Gauss (« courbe en cloche »). La courbe probabiliste est centrée sur la vraie valeur M (absence de biais) et elle est plus étalée que la courbe rouge, ce qui traduit une variance d'échantillonnage plus grande. La courbe rouge est centrée autour d'une valeur M' différente de la vraie valeur M, si bien que le biais vaut $M'-M$.

▶ D'autres techniques d'échantillonnage empirique : unités-type, échantillons de volontaires, et « *Access panels* »

Les techniques empiriques d'échantillonnage s'étendent bien au-delà des enquêtes par quotas, qui n'en sont qu'une modalité. Présentons maintenant trois pratiques concurrentielles, la première (les unités-type) étant désormais peu utilisée, alors que la troisième (les « *Access panels* ») est en plein essor.

Échantillonnage ne veut pas nécessairement dire sélection aléatoire. Historiquement, ce n'est d'ailleurs pas la sélection aléatoire qui a été utilisée dans les premiers temps des enquêtes par sondage, mais plutôt des techniques où on choisissait consciencieusement – et dans l'idéal judicieusement – les individus enquêtés. On peut utiliser le terme très évocateur de « choix raisonné » pour désigner les échantillons définis sans aucune intervention du hasard. C'est une approche qui est entièrement dépendante d'un modèle et qui est donc



La **figure 2** explique comment l'erreur d'échantillonnage, mêlant biais et variance, évolue en fonction de la taille de l'échantillon. Dans le cas probabiliste (courbe verte), l'erreur est grande dans la zone des (très) petites tailles d'échantillon et la variance y contribue beaucoup. Le biais étant toujours nul et la variance tendant vers zéro quand la taille de l'échantillon augmente, la courbe verte décroît jusqu'à (presque) se confondre avec l'axe des abscisses. La courbe rouge se situe en dessous de la verte dans la zone des (très) petites tailles d'échantillon, car le tirage empirique a l'avantage par rapport au tirage probabiliste aléatoire simple et équiprobable (celui qui attribue la même probabilité de tirage à tous les échantillons de taille donnée). Comme le biais est non nul dans le cas empirique, et qu'il n'est pas (ou peu) sensible à la taille d'échantillon, la courbe rouge est décroissante mais tend vers une droite située au-dessus de l'axe de abscisses – positionnée à l'ordonnée $M'-M$, valeur du biais d'échantillonnage. Les deux courbes se croisent nécessairement « quelque part » et ce croisement définit une taille d'échantillon critique au-dessus de laquelle le sondage empirique est moins efficace que le sondage probabiliste le plus simple qui soit.

scientifiquement déviante pour un statisticien qui s'obstine à vouloir éviter les modèles, car il y a nécessairement un biais d'échantillonnage et la variance d'échantillonnage perd son sens. Les individus enquêtés sont ceux qui représentent le mieux, pense-t-on, la population complète, et on peut ainsi parler d'unités-type, ou encore d'individus « représentatifs » de la population. Dans l'exemple de l'enquête sur les tâches domestiques, on pourrait ainsi choisir quelques individus coopératifs dans chaque croisement de modalités des trois variables explicatives retenues. Encore une fois, puisque sous couvert du modèle la méthode de sélection n'a pas d'importance, l'intervention du hasard n'apporte rien. Cette méthode, qui n'est utilisée que dans de rares circonstances bien adaptées, conserve tout son intérêt pour de tous petits échantillons, pour lesquels la variance serait énorme si on laissait au hasard le soin de décider de leur composition. Typiquement, on peut procéder ainsi pour définir un échantillon de quelques départements dans lesquels on effectue ensuite des tirages d'individus. Cette approche est également appliquée par l'Insee dans la sélection de certains produits et points de vente suivis dans l'indice des prix à la consommation.

L'échantillonnage de volontaires renvoie à toutes les situations pour lesquelles on laisse le soin à une partie de la population de participer à une enquête à son entière initiative. Bien sûr, les enquêtes en France sont avant tout placées sous le sceau du volontariat – même si la plupart des enquêtes de la statistique publique sont légalement obligatoires – en ce sens où un refus n'a pratiquement jamais de conséquence significativement pénalisante pour les personnes physiques comme pour les entreprises. Mais les échantillonnages probabilistes, et les échantillonnages empiriques par quotas dans une moindre mesure, sont constitués selon certaines règles qui cherchent en amont à structurer l'échantillon d'une façon efficace et la collecte s'appuie sur des principes qui tendent à préserver au maximum la composition de l'échantillon tiré. Exempt de tout cadrage, l'échantillonnage de volontaires n'a aucune de ces vertus et s'avère de fait scientifiquement critiquable. On trouve essentiellement dans cette catégorie les enquêtes d'opinion sur internet appelant les consommateurs à se prononcer sur un produit ou une prestation. Les indices de satisfaction qui en résultent sont soumis à des risques de biais considérables puisqu'on se trouve dans le cas *a priori* d'une très forte corrélation entre la probabilité de participation et la variable d'intérêt : il est par exemple assez naturel, pour un consommateur qui est mécontent d'un repas dans un restaurant, de mettre une mauvaise appréciation sur internet, et à l'inverse d'en formuler une très bonne s'il est très satisfait. Mais confronté à une prestation plus standard, fera-t-il cet effort ?

> ***Exempt de tout cadrage, l'échantillonnage de volontaires n'a aucune de ces vertus et s'avère de fait scientifiquement critiquable.***

Les échantillons tirés d'« *Access panels* » relèvent d'un cas intermédiaire enchaînant un échantillonnage de volontaires et un échantillonnage par quotas. Ce vocable désigne un ensemble de pratiques probablement assez diversifiées, mais dans bon nombre de cas il s'agit de constituer en amont un échantillon de volontaires de très grande taille, géré en continu, servant de base de sondage pour produire ensuite au fil de l'eau des échantillons beaucoup plus petits respectant des quotas de circonstance – avec les risques que l'on vient d'exposer en termes de biais. Il ne faut pas négliger le fait que les volontaires reçoivent des gratifications, sous différentes formes, en contrepartie de leur participation, et cela n'est très probablement pas sans conséquence sur la composition des échantillons, quoi qu'on en dise. Un « *Access panel* » a l'avantage de produire à moindre coût des échantillons d'individus présentant un profil ciblé, allant jusqu'à permettre de sonder des populations rares, mais c'est une source de données à géométrie variable qui peut se révéler d'une grande opacité pour les utilisateurs. De telles situations doivent éveiller la méfiance : le manque d'informations sur les méthodes est de façon générale problématique, et dans ce cas précis pose question si le processus de constitution et de gestion de l'« *Access panel* », ainsi que sa structure, ne sont pas explicites.

▶ Le paradoxe des *big data* : un exemple catastrophique récent…

On a coutume de penser que plus il y a de données, meilleure est la statistique. C'est faux, et même grossièrement faux : constituant le paradoxe des *big data*, il apparaît que la quantité n'est pas gage de qualité (*Meng, 2018*), et en voici deux preuves.

> ***On a coutume de penser que plus il y a de données, meilleure est la statistique. C'est faux, et même grossièrement faux.***

Le premier exemple concerne la très récente crise sanitaire. Aux États-Unis, en 2021, trois dispositifs (parmi bien d'autres) ont été conçus pour mesurer la couverture vaccinale des Américains contre le coronavirus (*Bradley, 2021*). Deux échantillons d'inspiration empirique – le *Delphi-Facebook* (DF) et le *Census Household Pulse* (CHP) – ont été sélectionnés, alors qu'un troisième, conçu par Axios-Ipsos (AI), avait les caractéristiques d'un échantillon probabiliste. L'échantillon DF, organisé en vagues hebdomadaires de 250 000 individus, a cumulé 4,5 millions de répondants entre janvier et mai 2021, pris parmi les utilisateurs actifs de Facebook. Le mode de collecte était (évidemment) un mode de collecte par Internet. L'échantillon CHP, tiré à partir d'un fichier rassemblant des adresses Internet et des numéros de téléphone, cumulait 600 000 réponses obtenues également *Online* sur la même période. Ces deux échantillons sont issus de tirages aléatoires dans des bases de sondage incomplètes (et même largement incomplète pour DF) et, surtout, sont totalement assimilables à des échantillons de volontaires compte tenu de leurs taux de réponse excessivement faibles (1 % pour DF et 6 à 8 % pour CHP). L'échantillon AI interrogeait 10 000 personnes sur la période. Il a été tiré de manière probabiliste dans une réserve de grande taille, elle-même constituée de manière probabiliste à partir d'une base de sondage d'adresses postales quasi exhaustive. Cette réserve, évoluant au fil du temps, rassemble certes des personnes volontaires pour participer à diverses enquêtes et s'apparente donc en cela à un « *Access panel* » (*Ipsos Knowledge Panel*), mais il s'agit en la circonstance d'un dispositif qui maximise les composantes probabilistes et qui est géré et contrôlé comme un échantillon probabiliste, en respectant les bonnes pratiques. Le taux de réponse final AI s'élève à 50 %. L'enquête se déroulait *Online* mais Ipsos a prêté une tablette à toutes les personnes n'ayant pas accès à Internet. La situation était très favorable pour apprécier la performance de chaque dispositif parce qu'on dispose de la vraie couverture vaccinale : en effet, l'*US Center for Disease Control and Prevention* est une administration d'État qui compile les statistiques de vaccination reflétant la réalité du terrain. Cela se fait avec un décalage temporel, mais on obtient néanmoins les 'vraies valeurs'. Tous les échantillons sont repondérés – il s'agit de redressements − afin que certaines structures socio-démographiques soient estimées de manière parfaite. Les résultats sont affligeants : le processus DF surestime en mai 2021 le vrai taux de vaccination (égal à 60 %) de 17 points, le processus CHP le surestime pour sa part de 14 points… et l'échantillon AI propose une estimation qui s'est avérée correcte ! On a vérifié, à partir des données collectées, que l'échantillon empirique hebdomadaire DF (250 000 individus) produit des estimations d'une qualité statistique équivalente à celle d'un échantillon probabiliste de… 10 individus répondants ! Une catastrophe qui s'explique en grande partie par un déséquilibre considérable des deux gros dispositifs selon différents critères, en particulier le niveau d'éducation et l'origine ethnique. En comparant avec les données du recensement, il est apparu clairement que DF et CHP sur-représentent massivement les personnes ayant un niveau d'éducation élevé (poids de l'équivalent du Bac +4 ou plus : 30 % dans la population, 36 % dans AI, 45 % dans DF et 55 % dans CHP) et sous-représentent, dans une moindre mesure, les personnes afro-américaines et, pour DF, les personnes asiatiques. Tout cela tient à la nature des bases de sondage, du mode de collecte, et d'une stratégie de gestion de la non-réponse très différente selon les enquêtes. Or, il s'avère dans les faits que les personnes à haut diplôme et blanches se font davantage vacciner que les autres catégories de la population américaine. Pressentant le piège, le dispositif CHP a redressé sur l'ethnicité et le niveau d'éducation, limitant ainsi les effets du

déséquilibre de l'échantillon interrogé, mais pas DF. Bien que cela n'ait pas été prouvé, on soupçonne également des déséquilibres dommageables portant sur l'opinion politique et sur le partage entre résidence urbaine et résidence rurale. Ce soupçon est fondé, car l'échantillon AI a été redressé en fonction de l'opinion politique (*partisanship*) et en fonction de la catégorie de commune (*metropolitan status*), et ce n'est pas le cas des deux autres dispositifs, alors même que l'on sait pertinemment que ces deux variables ont un effet significatif sur la propension à se faire vacciner (par exemple on se fait moins vacciner en milieu rural).

▶ … un précédent tout aussi révélateur

Le second exemple est emprunté à l'histoire (*Antoine, 2005* ; *Lusinchi, 2012*). Dans les années 1930, aux États-Unis, les médias avaient coutume de procéder à des opérations d'enquête dites « vote de paille », consistant à poser des questions par voie postale à des personnes figurant sur des fichiers nominatifs accessibles de diverses natures tels que des abonnés à un magazine, des abonnés au téléphone, des propriétaires de véhicule, ou des listes électorales. Le célèbre *Literary Diges*t a utilisé cette technique en 1936 pour prédire le vainqueur de l'élection présidentielle, qui opposait le démocrate Franklin Roosevelt et le républicain Alfred Landon. Au même moment, trois précurseurs – George Gallup, Elmo Roper et Archibal Crossley – ont utilisé, ce qui était assez nouveau et audacieux, des échantillonnages respectant certains quotas : sans avoir la rigueur des sondages probabilistes, ils s'efforçaient néanmoins de diversifier autant que possible les profils des répondants, et leurs structures selon plusieurs variables étaient « contrôlées ». Chaque sondeur avait conçu son enquête, et les trois échantillons comprenaient chacun quelques milliers ou dizaines de milliers d'individus. En face, le *Literary Digest* se glorifiait d'avoir collecté deux millions de réponses auprès de volontaires (sans qu'on ne sache précisément la taille exploitée en réalité – mais elle était très importante), qui lui permettaient de prédire une victoire très nette de A. Landon avec 57,4 % des suffrages. Les trois sondeurs annonçaient au contraire la victoire de F. Roosevelt. Le verdict a été sans appel : Roosevelt l'a emporté haut la main avec 61 % des votes. Que s'est-il passé ? Les individus recevant les courriers du *Literary Digest* étaient plus éduqués et plus fortunés que l'Américain « moyen » : il fallait au moins savoir lire et écrire pour pouvoir répondre, et l'abonnement à des journaux ou la possession de certains biens – téléphone, véhicule entre autres − témoignaient d'une éducation certaine, d'une aisance financière, etc. Ces personnes étaient majoritairement en faveur du parti républicain.

Ces deux exemples illustrent les effets pernicieux d'un échantillonnage insuffisamment contrôlé et insuffisamment corrigé. Ainsi, les *big data* du *Literary digest*, de DF et de CHP n'ont paradoxalement pas pesé lourd face aux dispositifs beaucoup moins volumineux mais bien mieux réfléchis de Gallup et d'Axios-Ipsos. Sur le fond, c'est inquiétant parce qu'on peut toujours craindre qu'une variable explicative du phénomène – souvent complexe et protéiforme − que l'on veut mesurer ne soit pas prise en compte ni dans le processus d'échantillonnage, ni lors de l'estimation au travers des redressements. Ce peut être par ignorance, par manque de connaissance des vraies structures, ou pour des raisons de nature culturelle ou juridique. Par exemple, dans les enquêtes menées en France, l'ethnicité et la sensibilité politique − dont on peut penser qu'elles sont corrélées à un certain nombre de comportements – ne sont *a priori* que rarement prises en considération dans l'établissement des quotas.

> ***Il est sécurisant de produire des estimations issues de sondages dans un cadre mathématique maîtrisé, offrant un minimum de garanties ainsi que des mesures de qualité des estimations produites.***

L'affaire du *Literary Digest* a certainement joué un rôle catalyseur dans le développement de la théorie formalisée des sondages probabilistes, qui date de cette époque. Elle a montré en quoi il était sécurisant de produire des estimations issues de sondages dans un cadre mathématique maîtrisé, offrant un minimum de garanties ainsi que des mesures de qualité des estimations produites.

En guise de conclusion

Un échantillon est un objet multidimensionnel complexe qui possède de nombreuses facettes. Il peut être très harmonieux sous certains angles et fort disgracieux sous d'autres si bien que pour l'apprécier pleinement, il faudrait pouvoir l'examiner sous toutes ses faces. Si la taille de l'échantillon répondant est une donnée essentielle pour la qualité d'une enquête, une autre clé du problème réside dans le rôle que l'on confie au hasard. Lorsque la taille de l'échantillon est suffisante, le hasard généré par un algorithme a l'avantage de réduire considérablement le risque de déformation de l'échantillon dans toutes ses dimensions, tandis que celui que l'on attribue à l'homme tout au long du processus de sélection peut s'avérer dévastateur. Confier les échantillonnages empiriques à des structures professionnelles expérimentées permet de réduire les risques de biais. Mais pour éteindre les polémiques touchant à l'échantillon, on peut aussi être tenté de changer la nature du hasard : le hasard des modèles de comportement, traduisant une hypothèse simplificatrice de la réalité, permet de changer de paradigme, offrant un cadre séduisant mais compliqué à comprendre et reportant finalement les risques sur les erreurs de spécification du dit-modèle. Éviter de devoir conditionner les estimations diffusées à cet acte de foi est un argument fort de la Statistique publique pour limiter autant que possible l'utilisation de méthodes empiriques.

Par ailleurs, avec l'expérience des échantillons empiriques on apprend qu'il faut éviter de se fier à la quantité d'information, qui ne protège pas contre le risque de désastre statistique : c'est le paradoxe des *big data*… dont les journalistes du *Literary Digest* ont été parmi les premières victimes de l'histoire !

Finalement, si on ne tient pas compte du cas spécifique des très petits échantillons, pour lesquels la technique d'unités-type est la mieux adaptée, en termes d'efficacité statistique et à taille d'échantillon donnée, la méthode probabiliste est toujours préférable à la méthode empirique. Car on sait tirer des échantillons probabilistes respectant des quotas - avant qu'ils ne soient perturbés par la non-réponse : cette méthode magique s'appelle le tirage équilibré (*Deville, 2004*). Reste aux sondages empiriques l'avantage indéniable de la rapidité de mise en œuvre et de l'économie de moyens.

▶ Annexe. L'origine du biais dans les sondages empiriques

Le biais de sélection résultant d'un échantillonnage dépend de la relation entre la variable d'intérêt et la probabilité de sélection. Il est formalisé de la façon suivante.

Considérons une enquête par quotas dans une population de taille N impliquant deux variables de quotas, dont les modalités sont repérées respectivement par les indices i et j. Soit Y_k la valeur de la variable d'intérêt pour l'individu k et $\bar{Y}_{i,j}$ la vraie moyenne de cette variable dans la cellule (i,j). On peut toujours définir ϵ_k tel que : $Y_k = \bar{Y}_{i,j} + \epsilon_k$ pour tout $k \in (i,j)$. On note P_k la valeur de la probabilité de sélection de l'individu k. Partant de cette formalisation, on peut montrer que le biais d'échantillonnage de la moyenne simple calculée dans l'échantillon vaut :

$$\frac{1}{n} \times \sum_{i,j} N_{i,j} . \, Cov_{i,j}(P,Y)$$

où n désigne la taille de l'échantillon, $N_{i,j}$ la taille de la population dans la cellule (i,j) et $Cov_{i,j}(P,Y)$ la covariance entre la variable P et la variable Y dans la population constituant la cellule (i,j), soit :

$$Cov_{i,j}(P,Y) = \frac{1}{N_{i,j}} \sum_{k \in (i,j)} (Y_k - \bar{Y}_{i,j}) . \, (P_k - \bar{P}_{i,j})$$

$\bar{P}_{i,j}$ désigne la vraie moyenne des P_k dans la cellule (i,j).

La covariance est positive quand les variables Y_k et P_k varient dans le même sens. Elle est négative si ces variables varient dans le sens contraire. Elle vaut 0 si les deux variables sont indépendantes.

Ce dernier cas est le seul qui annule le biais.

Contrairement à ce que l'allure de la formule de biais peut laisser penser, ce dernier n'est pas sensible à la taille d'échantillon : cela résulte du fait que la probabilité de sélection P_k est toujours du même ordre de grandeur que le taux de sondage n/N.

▶ Bibliographie

- ANTOINE, Jacques, 2004. *Histoire des sondages*. Éditions Odile Jacob, 20 février 2004. EAN13 : 9782738115874.

- ARDILLY, Pascal et LAVALLÉE Pierre, 2017. *Les sondages pas à pas*. Éditions TECHNIP. ISBN 9782710811794.

- ARDILLY, Pascal, 2006. *Les techniques de sondage*. Éditions TECHNIP. ISBN 978-2-7108-0847-3

- ARDILLY, Pascal, CASTELL, Laura et SILLARD Patrick, 2022. Il y a sondage et sondage. In : *Blog Insee*. [en ligne]. 25 juillet 2022. [Consulté le 10 septembre 2023]. Disponible à l'adresse : *https://blog.insee.fr/il-y-a-sondage-et-sondage/.*

- BRADLEY, Valerie C., KURIWAKI, Shiro, ISAKOV, Michael, SEJDINOVIC, Dino, MENG, Xiao-Li et FLAXMAN, Seth, 2021. U*nrepresentative big surveys significantly overestimated US vaccine uptake*. Décembre 2021. In : *Nature*. Volume 600. Disponible à l'adresse : *https://www.nature.com/articles/s41586-021-04198-4*.

- BRÜGGEN, Elisabeth, VAN DEN BRAKEL, Jan A. et KROSNICK, Jon, 2016. *Establishing the accuracy of online panels for survey research*. In : *site de CBS Statistics Netherland*. Discussion Pape. [en ligne]. 11 avril 2016. [Consulté le 10 septembre 2023]. Disponible à l'adresse : *https://www.cbs.nl/en-gb/background/2016/15/establishing-the-accuracy-of-online-panels-for-survey-research.*

- DEVILLE, Jean-Claude et TILLÉ, Yves, 2004. *Efficient balanced sampling: The Cube method*. In : *Biometrika*. Décembre 2004. Volume 91, N°4, pp. 893-912.

- DEVILLE, Jean-Claude, 1991. Une théorie des enquêtes par quotas. In : *Techniques d'enquête*. [en ligne]. Décembre 1991. Statistiques Canada. Volume 17, N° N2, pp. 177-195. [Consulté le 23 octobre 2023]. Disponible à l'adresse : *https://www150.statcan.gc.ca/n1/pub/12-001-x/1991002/article/14504-fra.pdf*.

- FORSTER, Jonathan, 2001. *Sample Surveys: Nonprobability Sampling*. In : *International Encyclopedia of the Social & Behavioral Sciences*. Oxford, UK. Elsevier Ltd.. Pp. 13467-13470.

- LOHR, Sharon L., 2021. *Sampling: Designs and Analysis*. In : *Texts in Statistical Science*. 30 novembre 2021. Éditions Chapman & Hall, vol 3. ISBN 978-0367279509.

- LUSINCHI, Dominic, 2012. *"President" Landon and the 1936 Literary Digest Poll: Were Automobile and Telephone Owners to Blame?* In : *Social Science History*. Volume 36, N°1, pp. 23-54.

- MacINNIS, Bo, KROSNICK, Jon A., HO, Annabell S. et CHO, Mu-Jung, 2018. *The accuracy of measurements with probability and nonprobability survey samples*. In : *Public Opinion Quarterly*. [en ligne]. 31 octobre 2018. Volume 82, N° N4, pp. 707-744. [Consulté le 10 septembre 2023]. Disponible à l'adresse : *https://academic.oup.com/poq/article/82/4/707/5151369?login=true.*

- MENG, Xiao-Li, 2018. *Statistical paradises and paradoxes in big data: law of large populations, big data paradox, and the 2016 US presidential election*. In : *The Annals of Applied Statistics*. [en ligne]. Juin 2018. Volume 12, N°2, pp. 685-726. [Consulté le 23 octobre 2023]. Disponible à l'adresse : *https://statistics.fas.harvard.edu/files/statistics-2/files/statistical_paradises_and_paradoxes.pdf*.

- SHIRANI-MEHR, Houshmand, ROTHSCHILD, David, GOEL, Sharad et GELMAN, Andrew, 2018. *Disentangling Bias and Variance in Election Polls*. In : *Journal of American Statistical Association*. [en ligne]. 25 juillet 2018. Volume 13, n°522, pp. 685-726. [Consulté le 23 octobre 2023]. Disponible à l'adresse : *https://www.tandfonline.com/doi/full/10.1080/01621459.2018.1448823*.

- SMITH, Terence Michael Frederick, 1983. *On the validity of inferences from non-random samples*. In : *Journal of the Royal Statistical Society*. [en ligne]. Juillet 1983. Volume 146, n°4, pp. 394-403. [Consulté le 23 octobre 2023]. Disponible à l'adresse : *https://www.jstor.org/stable/2981454*.

PRÉSENTATION DU NUMÉRO N10

Avec le numéro 10, le Courrier des statistiques fête ses cinq années de publication nouvelle formule et poursuit l'exploration des problématiques et des méthodes de la statistique publique.

La revue débute par un sujet désormais incontournable pour les statisticiens : la visualisation des données ou datavisualisation. Entre diffusion et communication, la dataviz cherche à simplifier les messages pour faciliter la compréhension des lecteurs et leur donner envie de lire.

Le second article, sur les statistiques de la défense, aborde un domaine où les données, souvent sensibles, sont à la fois très confidentielles et ouvertes aux chercheurs dans des conditions très sécurisées.

Quelles données administratives, quelles enquêtes, quels choix pour les statistiques sur le sport ? C'est tout l'enjeu du troisième article.

Dans ce numéro, deux articles sur des répertoires font écho à ceux déjà publiés sur ce sujet dans le numéro 8. FINESS est le répertoire des établissements sanitaires et sociaux et joue un rôle fondamental dans l'écosystème des systèmes d'information de santé. Les usages de Ramsese, le répertoire académique et ministériel sur les établissements du système éducatif sont très variés : pilotage, gestion, interopérabilité et besoins statistiques. Ces deux répertoires partagent, dans leur domaine respectif, centralité et fortes exigences de qualité.

Enfin, le dernier papier évoque, de façon pédagogique et en s'appuyant sur des exemples marquants, les différences entre sondages aléatoire et empirique.

ISSN 2107-0903
ISBN 978-2-11-162412-2

Insee
Mesurer pour comprendre

STATISTIQUE PUBLIQUE