QU'EST-CE QU'UNE DONNÉE ?

IMPACT DES DONNÉES EXTERNES SUR LA STATISTIQUE PUBLIQUE

*Pascal Rivière**

Le statisticien public utilise une matière première originale : les données. Mais outre celles qui sont issues d'enquêtes ou de déclarations administratives, il est amené à mobiliser des données d'autres natures, qui ne résultent pas toujours d'un processus d'observation. Comprendre ce matériau « data », c'est en explorer les principales dimensions, en s'appuyant sur le triplet <concept, domaine, valeur>.

Toute donnée se caractérise par un vaste faisceau de conventions (sémantique, nomenclatures, formats, etc.), et par l'infrastructure de connaissances dans laquelle elle s'inscrit, impliquant des choix qui n'ont rien de neutre. Une donnée se révèle aussi dépendante de l'environnement qui lui a donné naissance, et des processus productifs qui l'utilisent. On constate alors que les données ne sont pas pures et parfaites, ne vont pas de soi : paradoxalement, les données ne sont pas données.

Pour les besoins de la statistique publique, utiliser efficacement une telle matière requiert de démêler un entrelacs de conventions, et de construire une sorte d'appareil d'observation a posteriori, rigoureux sur les temporalités, et tenant compte de l'écosystème dans lequel la donnée externe s'inscrit.

Official statisticians use an original raw material, namely data: survey data, but also administrative data. They also use other management data that are not the result of an observation process. Understanding this material means exploring its main dimensions, using the definition of data as a triple <concept, domain, value>.

All data are characterized by a set of conventions, about semantics, classifications, formats, etc. Moreover, data exist within a knowledge infrastructure, and they are stored according to non-neutral choices. Data also depend on the environment in which they were born, and on the productive processes that use them. We then see that data cannot be pure and perfect: data are not given, they are side effects of operational processes.

Using efficiently such a material for the purpose of official statistics requires unravelling the implicit set of existing conventions, and building a kind of observation system a posteriori, taking into account the ecosystem in which these data were embedded.

* Chef de l'Inspection générale, Insee,
pascal.riviere@insee.fr

À l'instar de l'ébéniste, du forgeron ou du tailleur de pierre, le statisticien se confronte à un matériau brut, imparfait, traversé de nœuds et de failles. Mobilisant des outils et méthodes qui lui sont propres, il le polit, l'assemble et le met en forme. Ce matériau qu'il travaille, et qu'il contribue à créer, ce sont les données. Or celles-ci pullulent, jaillissent de toutes parts, sans cesse, et dans tous les domaines de la vie : c'est là, semble-t-il, une chance fantastique pour tous les artisans de la donnée. Il s'agit pourtant d'une matière étrange, foisonnante, incroyablement hétérogène : en apparence facilement accessible, elle nous échappe, résistant aux tentatives de définition opérationnelle. Chacun en a sa propre perception, et rétablir la neutralité de l'observateur n'est pas un vain mot. L'objet du présent article est de fournir des clés pour mieux la comprendre, et de voir en quoi les caractérisations proposées ont un impact sur l'activité du statisticien.

LES DONNÉES EN STATISTIQUE

A priori, le lien entre statistique et données va de soi : *« La statistique est la discipline qui étudie des phénomènes à travers la collecte de données, leur traitement, leur analyse, l'interprétation des résultats et leur présentation afin de rendre ces données compréhensibles par tous »*[1].

Mais il n'est pas sans intérêt de remonter plus loin dans le temps. Si l'on étudie de plus près la littérature classique sur le métier de statisticien (Volle, 1980), ou de nombreux cours de statistique mathématique, on constate que le mot « donnée », tout en étant indiscutablement présent, est moins central qu'on ne l'imaginerait. Il est souvent question de *résultats d'expériences* : la notion d'expérience, cette fois aléatoire, occupe une place centrale dans une autre discipline, celle des probabilités, dont on connaît les forts liens qu'elle entretient avec la discipline statistique. On évoque aussi fréquemment les *observations*. Dans l'esprit, les données de la statistique sont les résultats d'observations relativement à des variables que l'on a définies, et des traitements ultérieurs qu'on leur a appliqués. Implicitement, on fait appel à tout un appareil d'observation, *via* une expérience scientifique (dans le domaine de l'épidémiologie, par exemple), ou *via* une enquête, entre autres possibilités.

> *« Les données de la statistique sont les résultats d'observations relativement à des variables que l'on a définies, et des traitements ultérieurs qu'on leur a appliqués. »*

Certes, les statisticiens publics s'aventurent depuis des décennies au-delà de ce cadre, en ayant recours à des données administratives (les DADS[2], typiquement). Mais à certains égards, le processus déclaratif présente de nombreux points communs avec le processus d'enquête : en forçant un peu le concept, on peut ainsi considérer que les déclarations administratives relèvent d'une démarche d'observation, même si celle-ci s'effectue à des fins de gestion (Rivière, 2018).

Depuis le tournant des années deux-mille, et plus particulièrement depuis une dizaine d'années, on assiste à un changement de paradigme majeur : il ne s'agit plus uniquement pour le statisticien de construire son processus d'acquisition des données, car il existe dans le monde de multiples sources de données, parfaitement ouvertes (*open data*), ouvertes

1. Cf. Wikipédia, Statistique.
2. Déclarations Annuelles de Données Sociales.

sous condition (données accessibles aux chercheurs), *via* des conventions, ou bien payantes sous diverses formes. C'est donc une matière potentiellement accessible, très riche. Mais on ne sait pas comment elle a été élaborée, et en particulier rien ne garantit qu'elle résulte d'une démarche d'observation.

Or réaliser une enquête, mener une expérience scientifique, organiser un processus déclaratif, donnent une vision très particulière et à vrai dire biaisée de ce que peut être une donnée. Dans chacune de ces situations, il s'agit en quelque sorte de prendre une photo à un instant t, en interrogeant le réel, soit à travers un questionnement, soit en sollicitant le monde physique, avec l'usage de capteurs. L'observation est une façon particulière de recueillir les données qui n'est pas la seule possible. Si les données accessibles résultent d'une autre approche, le statisticien doit le savoir afin d'éviter des erreurs dans la transformation de la matière, dans l'interprétation des résultats.

Tout cela oblige à reconsidérer la manière d'appréhender le mot « donnée », dans toutes ses dimensions.

LA DONNÉE : TENTATIVE DE CARACTÉRISATION

Caractériser le concept de donnée est d'autant plus délicat que nombre d'ouvrages sur le sujet *data* éludent tout simplement la question de la définition. L'étymologie fournit un point de départ original, le verbe *donner* n'étant pas neutre ; en anglais, *datum* et son pluriel *data* sont issus du latin *dare* qui signifie… donner. Pour Howard Becker, ce choix est un accident de l'histoire (Becker, 1952) : on aurait dû pointer non pas *« ce qui a été donné »* au scientifique par la nature, mais plutôt ce qu'il a choisi de prendre, les sélections qu'il a opérées parmi l'ensemble des données potentielles. Pour évoquer le caractère partiel et sélectif inhérent aux données, il eût fallu choisir *captum* plutôt que *datum*.

Essayons naïvement les dictionnaires. On y trouve plusieurs explications assez disparates :

- *« Ce qui est donné, connu, déterminé dans l'énoncé d'un problème »* ;
- *« Élément qui sert de base à un raisonnement, de point de départ pour une recherche » ;*
- *« Résultats d'observations ou d'expériences faites délibérément ou à l'occasion d'autres tâches et soumis aux méthodes statistiques » ;*
- *« Représentation conventionnelle d'une information permettant d'en faire le traitement automatique ».*

Les définitions habituelles naviguent ainsi entre statut, fonction, origine et représentation de la donnée.

Dans la littérature sur les bases de données, on peut citer (Elmasri et Navathe, 2016), ouvrage de référence qui définit brièvement et de façon incidente le concept[3] : *« By data, we mean known facts that can be recorded and that have implicit meaning »*[4]. On met donc en exergue le support, l'importance de la sémantique, et on retrouve l'idée de *fait.*

3. On trouve d'autres caractérisations, qui sont celles de l'informatique, mais on les précisera dans la partie suivante.
4. *« Par données, nous entendons les faits connus qui peuvent être enregistrés et qui ont une signification implicite ».*

(Borgman, 2015) étudie plus en profondeur le sujet et aboutit à la définition suivante : « [...] *data are representations of observations, objects or other entities used as evidence of phenomena for the purposes of research or scholarship* »[5]. On retrouve l'idée de représentation du réel (utilisée pour définir l'information), et on constate avec intérêt qu'on ne se limite pas aux observations. Les mots « entité », « objet » apparaissent, l'entité étant définie ensuite par l'auteur comme « *quelque chose qui a une existence réelle* », matérielle ou digitale. Mais on se place dans le contexte d'une utilisation académique.

> *« Si un fait est faux, il cesse d'être un fait, mais si une donnée est fausse, elle reste une donnée. »*

(Kitchin, 2014) consacre tout un chapitre[6] à explorer le mot « *data* ». Il y explicite par exemple la « matière » dont sont faites les données : elles sont abstraites, discrètes, agrégeables, et ont un sens indépendamment du format, du support, du contexte. Il effectue surtout une distinction essentielle entre *fait* et *donnée* : si un fait est faux, il cesse d'être un fait, mais si une donnée est fausse, elle reste une donnée. Ainsi, les données sont ce qui existe préalablement à l'interprétation qui les transforme en faits, preuves, informations ***(encadré 1)***.

Plus généralement, on peut constituer une pyramide données > information > connaissance > sagesse ***(figure 1)***, où chaque couche précède l'autre, et se déduit de la précédente[7] par un « processus de distillation » (abstraire, organiser, analyser, interpréter, etc.), qui ajoute du sens, de l'organisation, et révèle des liens. Ce que l'on peut imager par la formulation de (Weinberger, 2012) : « *L'information est aux données ce que le vin est à la vigne* »), ou celle de (Escarpit, 1991) : « *Informer, c'est donner forme* ».

VERS UNE DÉFINITION EN TROIS DIMENSIONS...

Une des difficultés posées par la notion de donnée, c'est qu'elle embarque dans le même temps deux sujets bien distincts : d'une part la valeur (nombre, code, chaîne de caractères), que l'on va trouver sur un support quelconque, avec un certain mode de représentation, d'autre part le statut de cette valeur, sa sémantique, ce qu'elle est censée représenter. On mêle ainsi des préoccupations opérationnelles et des considérations plus abstraites.

Car si l'on assimile *donnée* à *valeur,* on bute immédiatement sur une contradiction. Lorsqu'on trouve dans un fichier le nombre 324, de quoi parle-t-on : de la hauteur de la tour Eiffel en mètres ? De la surface d'un champ, en hectares ? De la température de fusion d'un métal, en degrés Celsius ? De la vitesse maximum d'un véhicule, en km/h ? Du nombre d'habitants d'un village ?

Prise isolément, la ligne « 324 ; 3 ; 1889 » n'est qu'une succession de caractères vides de sens, au mieux une suite de trois nombres séparés par le délimiteur point-virgule, mais ce ne sont absolument pas des données. Le simple fait de préciser que ce sont des caractéristiques de la tour Eiffel, à savoir hauteur, nombre d'étages et année d'inauguration,

5. *« Les données sont des représentations d'observations, d'objets ou d'autres entités utilisées comme preuves de phénomènes à des fins de recherche ou d'érudition ».*
6. Voir le chapitre « *Conceptualising data* », pp. 2-26.
7. ... même si l'information ne requiert pas systématiquement de se fonder sur une couche de données (cf. le cri d'un animal alertant de la présence d'un prédateur).

> *« La notion de donnée est ainsi indissociable du concept auquel elle se réfère. »*

change tout et confère à cette série insignifiante de chiffres un tout autre statut : intuitivement, il s'agit bien de *données.* La notion de donnée est ainsi indissociable du concept auquel elle se réfère. Elle dépend également d'autres aspects, par exemple ici l'unité de mesure.

Tout cela nous oriente naturellement vers une définition souvent citée, notamment dans la littérature sur la qualité des données (Olson, 2003 ; Loshin, 2010 ; Berti-Équille, 2012 ; Sadiq, 2013). Dans un des premiers ouvrages majeurs sur le sujet, Redman analyse plusieurs définitions connues, et cherche la plus adaptée, en utilisant pour cela des critères : trois critères linguistiques (clarté, correspondance avec l'usage commun, absence de mention du mot « information ») et trois critères d'usage (applicabilité, possibilité d'introduire une dimension qualité, prise en compte des dimensions conceptuelle et de représentation) (Redman, 1996).

Cette démarche le conduit à une définition bien connue des informaticiens : on définit une donnée comme un triplet (entité, attribut, valeur), l'entité étant une modélisation d'objets du monde réel (physiques ou abstraits), l'attribut étant caractérisé par un ensemble de valeurs possibles, ou domaine. L'importance de ce dernier point va nous amener à reformuler très légèrement la définition de Redman, sans trahir la logique initiale, en proposant de définir une donnée par le triplet suivant[8] :

- le **concept** (par exemple hauteur d'un monument), qui se caractérise lui-même par la combinaison d'un objet (ici, monument) et d'un attribut (hauteur) ;
- le **domaine** des possibles : dans le cas d'un monument, un nombre entier positif, et la spécification de l'unité (mètres) ;
- la **valeur** : pour la tour Eiffel, 324.

Essayons maintenant de tirer le fil de chacune de ces dimensions pour mieux appréhender certaines spécificités de la notion de donnée, en particulier pour un usage statistique.

… LE CONCEPT ASSOCIÉ…

Le concept n'est rien d'autre que la signification supposée de la donnée, ce qu'elle est censée représenter, ce qui peut se matérialiser par une définition. Quelques exemples : surface d'un champ, chiffre d'affaires d'une entreprise, profession d'un salarié, mais aussi taille d'un monument, marque d'un véhicule, cours de bourse d'une action, nombre de buts marqués par un joueur, cotation du risque d'un client, diagnostic d'un patient, etc. À l'Insee, on centralise de telles définitions dans le référentiel RMéS (Bonnans, 2019).

On vérifie, à travers ces exemples, que le concept se définit toujours comme un attribut particulier d'une entité, d'un objet. En statistique publique, dans une démarche traditionnelle d'enquête ou de traitement de déclarations administratives, l'entité en question est souvent un individu (ou un ménage), une entreprise, mais ce peut être aussi un logement, un chantier,

8. La définition est équivalente, on n'ajoute ni n'enlève rien, cela permet simplement de faciliter les développements ultérieurs, en séparant clairement une dimension sémantique (concept) et une dimension plus technique (le domaine de valeurs).

Encadré 1. Quelques clés sur la notion d'information

L'irruption du concept d'information date de 1948, au confluent de plusieurs histoires, avec l'arrivée simultanée du fameux article de Shannon (Shannon, 1948) et de l'ouvrage de Wiener (Wiener, 1948). En introduisant une sorte de grammaire universelle de communication, l'un comme l'autre créent un jeu de concepts et de catégories s'appliquant à des sujets aussi divers que les télécommunications, le contrôle ou le calcul mécanique (Triclot, 2014). La notion d'information émerge ainsi dans un univers de machines, et joue un rôle unificateur essentiel, permettant de jeter des ponts entre des disciplines éloignées, en leur fournissant un vocabulaire commun. Ces travaux novateurs permettent aussi de quantifier l'information : dans la vision de Shannon, il y a en toile de fond une problématique de limites de performance pour la compression des messages et leur transmission, et le recours décisif à une représentation digitale qui permet de créer une véritable théorie du code. Shannon est conscient des limites de la chose, et ne pense pas qu'une seule conception d'information puisse rendre compte de toutes les applications possibles*.

Avec Wiener, la cybernétique fait de l'information une nouvelle dimension du monde physique : elle s'ajoute aux modalités d'explication classiques que sont la matière et l'énergie. Elle fait naître une nouvelle classe de problèmes en physique, en introduisant les processus de traitement de l'information.

Avec la montée en puissance des médias, de l'informatisation, l'usage du mot se banalise, mais sans tendre vers une définition simple ni partagée. On peut néanmoins donner quelques éléments explicatifs utiles.

De manière générale, (Buckland, 1991) identifie trois significations : information-objet (données, documents informatifs), information-processus (l'acte d'informer), et enfin information-connaissance (résultante du processus d'information).

(Floridi, 2010) la caractérise comme un bien ayant trois propriétés :

- non-rivalité : plusieurs personnes peuvent posséder la même information ;
- non-exclusivité : c'est un bien facilement partagé… et restreindre ce partage requiert un effort ;
- coût marginal nul.

Enfin, pour (Boydens, 2020), l'information :

- *« résulte de la construction (mise en forme) d'une représentation (perception du réel) ;*
- *au moyen d'un code ou d'un langage au sens large, à savoir tout système d'expression, verbal ou non* […], *susceptible de servir de moyen de communication entre objets ou êtres animés et/ou entre machines ;*
- *requiert un substrat physique pour être diffusée, qu'il s'agisse de la vibration de l'air* […], *d'une feuille de papier* […], *d'un support électronique* […];
- *doit être interprétée pour être utilisée ».*

** « Le mot « information » a été assigné à différentes significations par différents auteurs dans le champ général de la théorie de l'information. Il est probable que quelques-unes de ces significations se révéleront utiles dans certaines applications pour mériter des études supplémentaires et une reconnaissance permanente. On ne peut guère s'attendre à ce qu'une seule conception d'information rende compte de manière satisfaisante des nombreuses applications possibles de ce champ général »* (Shannon, 1953).

> *« Un objet peut avoir de nombreux attributs, mais parmi ceux-ci, certains jouent un rôle particulier : les traits d'identification. »*

un séjour hospitalier, un lycée. Et dans le libellé même du concept, on a souvent tendance à faire disparaître la référence à l'objet, car ce dernier va de soi.

Ajoutons qu'en toute rigueur, la donnée se réfère non pas à un objet en général, mais à un objet donné, à une instance[9] : ce qui fera sens, ce sera la surface d'un champ *bien précis,* la profession d'une personne *bien identifiée,* le nombre de buts marqués par un joueur *donné,* le chiffre d'affaires de *telle entreprise.*

L'attribut devra lui aussi être précisé, notamment sur le plan temporel : la profession à *telle date,* le nombre de buts marqués *telle saison,* le chiffre d'affaires *telle année,* etc. Un objet peut avoir de nombreux attributs, mais parmi ceux-ci, certains jouent un rôle particulier : les traits d'identification. Ce sont ceux qui permettent d'identifier l'objet sans ambiguïté[10], et donc de distinguer une instance d'une autre : on pense naturellement aux nom, prénom, date et lieu de naissance pour un individu ; à la raison sociale, l'adresse pour un établissement ; pour un séjour hospitalier, ce pourraient être par exemple la date de début de séjour, l'identifiant de l'établissement de santé et l'identifiant de l'individu. Les traits d'identification ont la particularité d'être soit inamovibles (date de naissance), soit rarement modifiés (adresse).

Mais revenons aux objets. Les cas de l'individu ou de l'entreprise présentent des avantages incontestables auxquels on ne pense pas toujours :

- une réelle stabilité dans le temps ;
- des traits d'identification indiscutables, permettant de les repérer, et de les distinguer entre eux, sans ambiguïté ;
- l'existence de référentiels reconnus dans lesquels on trouve ces traits, mais aussi des identifiants reconnus comme références communes (NIR, SIREN[11]), avec des principes d'immatriculation relativement transparents et partagés.

À l'inverse, les données susceptibles d'être obtenues dans d'autres sources peuvent se référer à des objets plus délicats à appréhender, car nécessitant une connaissance métier : la ligne téléphonique, le compte bancaire, le compteur électrique. Elles peuvent même concerner des objets volatils : des données comme le montant d'une transaction, la date d'un accident de circulation, renvoient à des objets (transaction, accident) qui s'apparentent à un événement, et qui n'ont donc pas de consistance temporelle.

Ainsi, pour certains types d'objet, il n'existe pas de population de référence, ayant une certaine stabilité et permettant des comparaisons macroscopiques, des contrôles ou des calages sur marges : par exemple on peut difficilement s'appuyer sur un référentiel d'accidents ou de transactions pour comparer à un total connu, avoir un cadre, une limite.

---

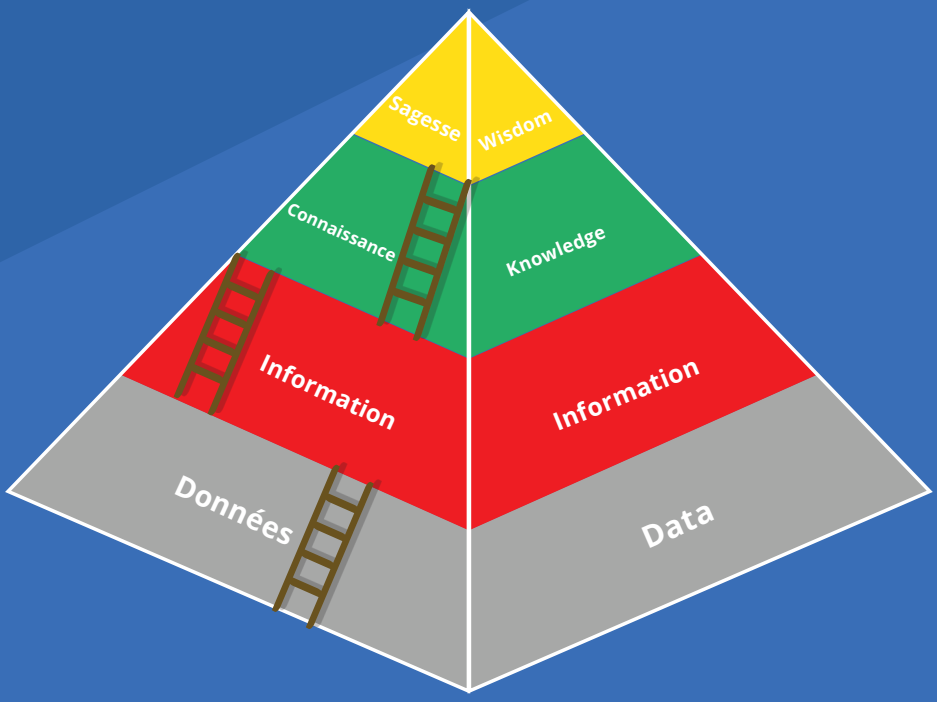
9. Ce terme est fréquent en informatique : on va parler de *classe* (pour l'entité en général) et d'*instance de classe.* Par exemple, Monument est l'entité abstraite, la classe, et la tour Eiffel, ou le Taj Mahal, en sont des instances.
10. Par exemple, (Loshin, 2010) propose plusieurs critères de qualité des données, parmi lesquels figure l'identifiabilité d'un objet (p. 142 et p. 144).
11. Numéro d'identification au répertoire (des personnes physiques) et numéro d'identification des entreprises au répertoire Sirene.

Pour les lignes téléphoniques ou comptes bancaires, les opérateurs ont cette vision mais rien ne les oblige à la partager. De plus, un utilisateur statisticien ou *data scientist* doit pouvoir se ramener à des entités ayant un sens pour l'analyse que l'on veut effectuer : par exemple se ramener au niveau individu, ménage, ou entreprise. Or d'une part il n'y a pas de bijection entre les deux (une personne peut avoir plusieurs comptes bancaires, plusieurs lignes téléphoniques), d'autre part la mise en évidence de ce lien ne va pas de soi (sans parler des problèmes juridiques soulevés).

… LE DOMAINE…

Alors que le *concept* auquel on se réfère reste abstrait, le *domaine* oblige à aborder des considérations plus opérationnelles. (Olson, 2003) le décrit comme l'ensemble des « *valeurs acceptables pour encoder un fait spécifique* ». Il précise que le domaine est générique, indépendant de la manière dont il sera physiquement implémenté. Le définir revient à expliciter les règles que devront respecter les *valeurs*, indépendamment des applications qui vont les utiliser. Si elles sont définies proprement, *via* un référentiel de métadonnées[12], cela facilite considérablement le travail des développeurs des applications utilisatrices[13].

Figure 1. Chaîne de valeur de la connaissance



12. À l'image du référentiel RMéS, décrit dans (Bonnans, 2019).
13. Olson souligne (p. 145) que, dans les faits, peu d'entreprises le mettent en pratique, en prenant l'exemple des dates, des codes postaux et des unités de mesure.

Le domaine dépend de la nature de la valeur, de son type[14]. Si c'est une quantité par exemple, elle sera souvent représentée par un intervalle ; ainsi, on accepte des nombres négatifs pour une température en degré Celsius, mais pas pour la hauteur d'un monument. Elle pourra se référer à une unité de mesure (degré Celsius, mètre), qui fera partie intégrante de la définition du domaine, mais ce n'est pas une obligation, notamment pour les quantités entières (nombre d'enfants).

Dans le cas d'une date, il faudra préciser les formats attendus, de type *jj-mm-aa,* ou *aa-mm-jj,* et on définira souvent l'ensemble des possibles par un intervalle, qui n'a pas tout à fait la même signification qu'un intervalle numérique, étant donné les spécificités des dates (jour ≤ 31, mois ≤ 12).

S'il s'agit d'un code alphanumérique[15], il arrive qu'on définisse le domaine **en intension**[16], par exemple pour les identifiants à structure complexe au sein de grands répertoires (NIR ou numéro de sécurité sociale pour les individus, SIRET pour les établissements d'une entreprise), les numéros de téléphone, et même les noms (Elmasri et Navathe, 2015). Caractériser le domaine par la liste **en extension** aurait en effet deux inconvénients, d'une part la liste serait trop longue à expliciter (il y en a des millions), d'autre part et surtout, elle ne cesse d'évoluer. Et l'on a donc plutôt intérêt à spécifier le domaine sur la base de règles de restrictions de l'ensemble des possibles : par exemple le NIR a une structure très précise, le 1er caractère étant le code sexe, les deux suivants représentant l'année de naissance, puis le mois, etc.[17]

Exprimer l'ensemble en extension est plus fréquent : c'est ce que l'on fait pour un questionnaire en associant codes et réponses (A – Tout à fait d'accord, B – plutôt d'accord, C – plutôt pas d'accord, D – Pas du tout d'accord), ou en discrétisant une variable quantitative (par exemple les tranches d'âge).

> *« Spécifier le domaine revient à le nommer et à définir un certain nombre de propriétés à respecter. »*

Cette liste de codes peut aussi dériver d'une nomenclature… sans en être une. En effet, les nomenclatures comme celles des professions, des activités économiques, des maladies se présentent sous forme d'arborescences à différents niveaux (Amossé, 2020) ; par exemple, dans la nomenclature d'activités française (NAF), les niveaux section, division, groupe, classe, sous-classe, etc. Ainsi, si une donnée a pour concept l'activité économique, le domaine ne pourra être la nomenclature dans son ensemble : ce sera nécessairement un niveau de cette nomenclature, ou plus généralement tout découpage jugé pertinent fondé sur celle-ci, conduisant *in fine* à une liste « à plat ».

On pourrait poursuivre à l'envi avec d'autres types de données, mais dans tous les cas, spécifier le domaine revient à le nommer et à définir un certain nombre de propriétés à respecter : le typage de la donnée, mais aussi une liste de valeurs acceptables, des éléments tels que les règles de restriction de longueur (ou tout autre règle de restriction du champ

14. En toute rigueur, nature et type sont deux propriétés distinctes : par exemple, la CSP est un code à deux chiffres, c'est sa nature. Ainsi, 21 > artisans. On peut représenter ce code par le nombre 21, ou par la chaîne de caractères « 21 » : ce sont deux *data types* différents.
15. Symbole formé d'une succession de caractères qui sont soit des chiffres, soit des lettres.
16. *i.e.* qu'on le définisse par ses propriétés et non par la liste de ses éléments.
17. De telles structures lexicales peuvent être définies mathématiquement par un langage dédié, celui des expressions régulières : voir l'exemple du cahier technique de la DSN (Cnav, 2020), pp. 71-73.

des possibles), les règles relatives à la représentation des valeurs manquantes, voire des points plus techniques comme le jeu de caractères utilisables[18].

Sur le fond, spécifier un domaine revient aussi à effectuer des choix de granularité[19], décisifs pour les usages ultérieurs : la finesse de description dans une nomenclature (cf. l'exemple de la NAF), l'unité de mesure d'une distance (m, km, années-lumière), la précision d'une date (année, jour, minute, seconde, nanoseconde). Cela consiste également à associer aux valeurs admissibles des consignes, des commentaires, indispensables pour créer la donnée (ex. *supra* : si l'on veut dire « plutôt d'accord », on saisira B) et pour l'interpréter (lorsque quelqu'un lira B, il en comprendra le sens).

Avec le *domaine,* on établit ainsi une convention sémantique et technique, commune aux concepteurs et utilisateurs des données, qui est, comme le *concept*, consubstantielle à la donnée. Muni du concept et du domaine, on dispose d'une sorte de réceptacle, au moins théorique, qu'il s'agit maintenant de nourrir d'une *valeur* (ou pas, d'ailleurs).

… LA VALEUR

Avec le troisième élément du triptyque, la *valeur*, on passe à une réalité plus tangible. Il s'agit là du nombre, du code, de la date, de la chaîne de caractères, que l'on associe donc à un domaine et à un concept. Ce dernier est cette fois instancié : la valeur sera relative à une entité et à un attribut bien définis, par exemple la profession et catégorie socio-professionnelle de tel salarié de tel établissement (selon la nomenclature PCS-ESE[20] et au 1er novembre 2020), la masse (en mégatonnes) du soleil, l'altitude (en m) du Mont-Blanc, etc.

Une précision s'impose ici : la valeur est *associée* au concept instancié, elle s'y *réfère*… mais cela ne veut pas dire qu'elle le représente fidèlement, cela n'assure à aucun moment qu'elle soit exacte : on peut trouver des valeurs différentes de l'altitude du Mont-Blanc selon les sources ; un code profession transmis peut être erroné, car il n'est plus à jour, ou parce que la personne qui a renseigné le code s'est trompée ; etc. Ainsi, la question de la fiabilité de la donnée n'est aucunement assurée par le triptyque concept-domaine-valeur. Celui-ci permet de donner une assise à la valeur, de lui conférer un statut autre qu'un simple nombre ou qu'un code vide de sens, mais sa véracité renvoie à d'autres sujets, en particulier à la manière dont la donnée a été construite.

À défaut de certitudes sur la qualité, la référence au domaine peut nous donner une garantie de conformité, une fiabilité de nature syntaxique (Batini et Scannapieco, 2016)[21] : au minimum, la valeur devrait appartenir à celui-ci (par exemple, le fait qu'une donnée représentant une date respecte bien les propriétés que doit avoir une date). Une telle conformité est fréquente… mais non obligatoire : cette fois, tout dépend des contrôles qui auront été effectués automatiquement sur la valeur. Très souvent, les outils de saisie vont être conçus pour que le choix s'effectue parmi des valeurs admissibles présentées à l'écran

18. Voir (Olson, 2003), p. 149, et aussi le cahier technique DSN (Cnav, 2020) évoqués *supra.*
19. On fait aussi un choix de granularité dans le concept, par exemple le fait qu'on se place à un niveau géographique plus ou moins fin.
20. Nomenclatures des professions et catégories socioprofessionnelles des emplois salariés des employeurs privés et publics.
21. Les auteurs distinguent *syntactic accuracy* et *semantic accuracy.* La première est l'adéquation au domaine, indépendamment de la véracité, la seconde est la proximité à la supposée valeur vraie. (Volle, 1980) effectue la même distinction, voir p. 60.

(liste de codes prête), ou bien pour rejeter la valeur si elle n'appartient pas au domaine prévu : c'est ce qu'on trouve par exemple dans la collecte de données par enquêteur. Lorsqu'on a affaire à des échanges de données informatisés et normalisés, le protocole d'échange permet d'assurer des propriétés qui vont même au-delà de l'appartenance à un domaine : c'est le cas des déclarations sociales (Renne, 2018). Plus généralement, les systèmes de gestion de bases de données intègrent implicitement des contrôles de typage. Dans toutes ces situations, par construction, la valeur est donc *conforme,* elle appartient à l'ensemble des valeurs admissibles. Cependant, il peut arriver que les données soient renseignées ou calculées automatiquement sans qu'aucun contrôle ne soit effectué, et soient stockées ensuite dans un fichier, sans garantie de conformité.

Enfin, sur le plan pratique, la valeur se trouve sur un support (physique, logique), qui n'a aucune raison d'être le même que le concept et le domaine. Et elle se réfère à un certain nombre de standards d'encodage qui sont, la plupart du temps parfaitement inconnus des utilisateurs finaux[22].

Comme pour le concept et le domaine, la valeur emporte donc avec elle ses propres règles, qui sont ici de nature plus technique, et des choix techniques auront ainsi été faits : cela vaut pour le type de donnée (*data type*), mais aussi, par exemple, pour la manière de prendre en compte les valeurs manquantes. Ainsi, comme pour les deux autres éléments du triptyque, la valeur rend nécessaire l'existence de conventions. Avec cette troisième et dernière dimension, on constate cependant que raisonner de façon individuelle, donnée par donnée, est largement artificiel.

LA DONNÉE, LES DONNÉES…

> *« La question de la fiabilité de la donnée n'est aucunement assurée par le triptyque concept-domaine-valeur. »*

En effet, le matériau « donnée » a ceci de particulier qu'il ne peut être envisagé seul, tel un grain de donnée séparé du reste. (Borgman, 2015) explique que « […] *data have no value or meaning in isolation; they exist within a knowledge infrastructure – an ecology of people, practices, technologies, institutions, material objects, and relationships* »[23]. Pour leur attribuer un sens, pour en extraire de l'information, il faut les mettre en regard de leurs congénères proches afin d'effectuer des comparaisons, de disposer de l'environnement nécessaire à l'interprétation.

Par ailleurs les données se présentent toujours de façon groupée, elles fonctionnent en meute, en quelque sorte. La plupart du temps, il s'agit de plusieurs instances du même concept, pour un même attribut (par exemple n individus, et pour chaque individu, le code profession), ou de plusieurs attributs de la même instance d'objet (pour un certain individu, toutes les données de déclaration fiscale), ou de plusieurs spécifications temporelles de la même instance, pour un même attribut (par exemple, l'effectif d'une certaine entreprise, année par année).

22. Pour les caractères, les standards ASCII et EBCDIC, par exemple.
23. *« Les données n'ont ni valeur ni sens isolément, elles existent à travers une infrastructure de connaissances – une écologie de personnes, pratiques, technologies, institutions, objets matériels, et relations » .*

> *« Cet enregistrement collectif et non individualisé des données oblige à se donner des règles documentées permettant de retrouver une donnée bien précise au sein d'un vaste ensemble. »*

Cet enregistrement collectif et non individualisé des données oblige à se donner des règles documentées permettant de retrouver une donnée bien précise au sein d'un vaste ensemble de données, et tout ceci dépend de la manière dont les données ont été stockées. Il faut bien les placer quelque part, mais il faut penser en même temps aux moyens de les récupérer.

Il existe pour cela de nombreuses possibilités. Historiquement les données se présentaient dans un fichier structuré, et une technique classique consistait à caractériser chacune d'elles par une plage de colonnes du fichier, en « positionnel fixe »[24], ou par le numéro d'ordre de la donnée dans une liste, en fixe délimité[25]. De telles approches nécessitent toute une documentation associée, en général assez lourde, et complexe à mettre à jour.

On peut procéder autrement, en adoptant un langage de balisage type XML[26], où l'on retrouve en partie l'idée que la valeur doit être enveloppée d'un concept et d'un domaine : chaque valeur figure entre deux balises, ces balises étant elles-mêmes, dans des schémas XML, associées à des règles formelles à vérifier (logique de domaine), et par un nom décrivant le concept. La caractérisation des données est ainsi autoporteuse et ne dépend plus d'une documentation. On peut également citer le format JSON[27], qui est un format plus léger que XML, moins verbeux, efficace, mais en même temps moins riche.

… DANS DES BASES, ENTREPÔTS, LACS, FLUX

Mais la possibilité la plus répandue consiste évidemment à stocker les données dans une *base de données*, gérée par un SGBD (système de gestion de base de données) : ceci induit une forte normalisation, fournit des garanties d'intégrité des données et un langage d'accès aux données (SQL) ; les **bases de données relationnelles** offrent la possibilité d'effectuer de façon fréquente de nombreuses modifications, et s'avèrent donc très adaptées à des processus de gestion (Codd, 1970). Une autre logique, celles des **entrepôts de données**, est en revanche conçue pour faciliter les travaux d'analyse[28], utiles dans le domaine de l'aide à la décision : ce sont alors des données figées, qui requièrent de respecter un cadre contraignant, avec des axes d'analyse communs et donc en particulier des nomenclatures communes alors que les sources d'information sont multiples. La technique des **lacs de données**, là aussi sur données figées, est bien moins contraignante en conception que les entrepôts de données, mais tout le travail de normalisation doit s'effectuer au moment de l'accès aux données.

24. Positionnel fixe : la position d'une donnée est indiquée par les colonnes de début et de fin entre lesquelles elle se trouve, par exemple un prénom entre les colonnes 31 à 50.
25. Fixe délimité : les valeurs sont séparées par des délimiteurs, et on saura par exemple que la donnée que l'on cherche est la troisième dans l'ordre.
26. *Extensible Markup Language*, ou « *langage de balisage extensible* » en français.
27. *JavaScript Object Notation.*
28. On parle de *On-Line Analytical Processing* (OLAP), les premiers papiers sur ce sujet datent de 1993 et impliquent de nouveau Codd, le concepteur des bases de données relationnelles.

Il existe même des techniques de structuration des données conçues pour des situations où elles n'ont pas le temps d'être stockées (systèmes temps réel, notamment) : ce sont les systèmes de gestion de **flux de données**, ou DSMS (*Data Stream Management Systems*) (Garofalakis, Gehrke et Rastogi, 2016).

Au total il existe de nombreuses méthodes pour enregistrer des données de façon organisée, les différences de méthodes ne devant pas être perçues comme étant de nature technique, mais plutôt comme fonction de l'usage envisagé des données (gestion, décisionnel, temps réel) et des éventuelles contraintes sur celles-ci (de volume, en particulier). Dans chaque cas on retrouve d'une manière ou d'une autre le concept et le domaine, soit en tant que (méta)donnée, soit dans une documentation liée. Soulignons enfin que le fait de raisonner sur un ensemble de données et non plus sur une donnée fait émerger d'autres notions : tout d'abord, il apparaît une exigence de cohérence d'ensemble, sur le plan structurel et sur le plan sémantique. Le périmètre que représente cet ensemble de données devient un sujet d'intérêt, à analyser en tant que tel et qui concerne pleinement les statisticiens. Enfin, la connaissance de la source d'information associée à cet ensemble de données est un critère de qualité (Loshin, 2010)[29]. Ainsi, on ne peut aborder les données sans questionner l'écosystème dans lequel elles se trouvent.

ENVIRONNEMENT DE LA DONNÉE

Le triplet (concept, domaine, valeur) fournit un cadre utile mais ne suffit pas à épuiser tout ce que transporte avec elle la matière « donnée ».

Les données ne sont pas données, elles n'existent pas dans la nature de façon immanente : leur existence résulte d'un besoin, elles sont imbriquées dans un environnement. Le fait qu'elles soient définies très rigoureusement ne les rend pas pour autant pures et parfaites, car elles sont *intrinsèquement liées à un usage.* Elles ne sont là que pour concourir à un objectif et non pour mettre à disposition du public une information de référence[30].

> *« Les données ne sont pas données, elles n'existent pas dans la nature de façon immanente. »*

(Denis, 2018) cherche ainsi à « *détricoter les fils de la donnée* »[31]. Il présente le cas tout simple d'un décès : ici, concept et domaine sont aisés à définir. Mais il met en évidence les multiples acteurs qui sont concernés par cette information, et les usages variés en fonction de leurs activités : compagnie d'assurances, administration, etc. Par exemple, les assurances ont besoin de justificatifs : le fait qu'une personne soit considérée comme décédée dans leur système d'information signifie qu'il existe une preuve. Mais à l'inverse une personne décédée pourra être enregistrée comme vivante dans ce même système… tout simplement parce que la preuve n'a pas été transmise : dès lors, la donnée ne reflète pas l'information vraie, elle prépare un usage dans le cadre d'un processus de gestion. De manière plus générale, les données souhaitées peuvent être fonctions d'événements dont rien ne garantit qu'ils soient matérialisés, déclarés de façon simultanée (exemple : l'arrêt de l'activité d'une entreprise).

29. Voir cohérence structurelle et sémantique (pp. 137-139) et la source (*lineage*), comme critère de qualité (pp. 135-136).
30. Sauf dans certains cas particuliers … notamment les données statistiques.
31. Voir pp. 18-19.

Autre exemple : lorsqu'une personne part à la retraite, le montant de cette retraite dépend de ses données de carrière. Une bonne partie de celles-ci sont transmises automatiquement[32], mais ce n'est pas la totalité, et il arrive parfois que le futur retraité transmette des éléments de carrière (feuilles de paie notamment) au dernier moment. Donc dans les données de carrière d'un individu qui n'est pas proche de la retraite, il peut rester des « trous » dus au fait que certaines périodes d'activité n'ont pas encore été déclarées : la carrière telle qu'elle figure dans le système de gestion ne correspondra donc pas à la carrière réelle.

Ainsi les données disponibles ne sont-elles pas nécessairement égales à la valorisation du concept, à la « donnée vraie » telle qu'on l'imagine. Les processus de gestion ne cherchent pas à produire « la vérité » ce qui ne veut pas dire que la donnée soit « mauvaise ». Comprendre la donnée, c'est comprendre l'objectif, les règles du jeu des processus qui la font naître.

> *« Les processus de gestion ne cherchent pas à produire « la vérité » ce qui ne veut pas dire que la donnée soit « mauvaise ». »*

Plus généralement, lorsque les données dérivent d'un cadre juridique (ex : sécurité sociale), il existe d'inévitables écarts entre les données telles qu'elles devraient être selon la loi, les données telles qu'elles seraient si l'on avait une observation parfaite du réel… et les données telles qu'elles sont dans les bases de données (Boydens, 2000). Il existe ainsi un écart entre données théoriques et données réellement récupérées, lié à la réactivité par rapport aux événements. Ceci permet d'aborder différemment la « qualité » des données : la qualité, c'est l'aptitude à l'usage[33], et non la justesse.

Il faut donc comprendre d'où elles proviennent, comment elles naissent (capteur, interface de saisie, calcul automatique, etc.) et connaître l'événement déclencheur de la naissance d'une donnée, mais aussi les processus ultérieurs qui s'en servent (exemple : déclenchement d'une prestation, d'un remboursement). On voit alors moins la donnée comme une vérité absolue, mais comme un maillon nécessaire dans une chaîne complexe. Cela peut permettre d'identifier les raisons pour lesquelles une donnée est absente (exemple de la carrière pour un régime de retraite), ou inexacte sans que le processus opérationnel soit déficient (exemple du décès vu par une compagnie d'assurance).

En comparaison des exemples précédents, les modes d'obtention traditionnels des données en statistique, à savoir les enquêtes, apparaissent comme une étrangeté : la donnée n'est pas directement liée à un usage, par un processus de gestion en aval, elle ne fait qu'alimenter le calcul de données agrégées, les *statistiques*, qui constituent justement un but en soi.

LE STATISTICIEN FACE AUX DONNÉES EXTERNES

Dans un monde numérisé où les *data* accompagnent en permanence nos vies et celles de nos organisations, on pourrait avoir tendance à penser que la multiplication des sources de données sur tous sujets rend le travail du statisticien plus facile. Pour reprendre une expression familière, il n'aurait « qu'à se baisser » pour les ramasser.

32. Le tout récent RGCU (Répertoire général de carrières unique) devrait améliorer sensiblement les choses.
33. *« Quality is fitness for use »* (Juran, 1951).

Cette vision est largement erronée, car elle recèle une confusion majeure sur le sens du mot « donnée » : on peut effectivement accéder, plus ou moins aisément d'ailleurs, à des quantités considérables de *valeurs* (nombres, codes, libellés, dates, etc.)… mais pas à l'équivalent en *données*. Car ces valeurs demeurent lettre morte tant qu'on ne dispose pas de clés solides permettant de les interpréter : concept, domaine, mais aussi périmètre suffisamment explicités. On ne peut envisager d'utiliser ces valeurs à des fins d'analyse sur la base de conventions faibles ou inexistantes, de caractérisations vagues, mal assurées, implicites.

Dès lors, la première responsabilité du statisticien face à des jeux de données externes est de révéler et démêler un entrelacs de conventions sous-jacentes à celles-ci (Martin, 2020)[34] : conventions sur les définitions, les objets, temporalités, nomenclatures, formats, valeurs manquantes, etc., tout ce qui va lui permettre de reconstituer le triptyque concept-domaine-valeur, et de caractériser la population couverte. Les conventions existantes sont à rechercher activement, à valoriser, car ce sont des alliées : on pense par exemple aux normes comptables, très utiles aux statisticiens d'entreprise (elles offrent au passage des garanties de qualité), mais aussi aux fondements juridiques des concepts, qui peuvent conduire à asseoir la qualité de la donnée ou au contraire à devoir pallier ses dérives.

Mais ce n'est pas suffisant : il devra passer de données vivantes, évolutives, liées à des processus opérationnels, à des *observations*. Celles-ci devront être figées dans le temps au même instant t (par exemple la situation des entreprises au 31/12/2019) ou sur une même période (par exemple l'activité des entreprises pour l'année civile 2019), et se référer à des objets distincts les uns des autres, *homogènes* entre eux, de même que leurs attributs. Ce travail d'homogénéisation est central dans l'activité statistique, mais à vrai dire il a commencé bien avant la statistique publique, avec le système métrique et l'unification des poids et mesures (Desrosières, 1993).

En d'autres termes, il s'agit de reconstituer *a posteriori* un pseudo-appareil d'observation, à partir de données qui n'ont pas été conçues à cette fin. Pour ce faire, le statisticien public utilise un puissant socle de conventions, largement partagé : répertoires, nomenclatures, définitions de concepts, unités statistiques, conventions de notation, ce qui se traduit par un vaste ensemble de métadonnées. Les nomenclatures font l'objet d'une vaste coordination (celles de l'Insee ou de l'ATIH[35] par exemple), et les conventions relatives aux populations de référence lui permettent de caler, contrôler et comparer les statistiques obtenues.

> *« Il s'agit de reconstituer* a posteriori *un pseudo-appareil d'observation, à partir de données qui n'ont pas été conçues à cette fin. »*

Avec des données externes non maîtrisées, le statisticien va aussi être confronté à une multitude de défauts de celles-ci, qui peuvent le conduire à les remettre en cause, à questionner leur qualité. Et ce, même si elles sont parfaites pour l'usage auquel elles sont assignées. Pour éviter cet écueil classique, il devra comprendre l'environnement opérationnel des données d'origine pour déterminer ce qu'il peut en faire. Cela affectera les étapes de

34. Sur l'importance des conventions, voir pp. 186-191.
35. Agence technique de l'information sur l'hospitalisation.

contrôle automatique et manuel, d'imputation des valeurs douteuses ou manquantes[36]. L'appariement, et donc l'identification d'objets, l'obligera à questionner le sens, la stabilité de ceux-ci, et leurs liens avec les unités statistiques envisagées (Christen, 2012). La validation finale des agrégats et leur interprétation renverra à la signification de la population de référence couverte.

Ces données externes ne sont qu'un intrant du travail du statisticien, qui s'en servira pour élaborer… de nouvelles données, en l'occurrence des données agrégées nommées « statistiques ». En tant que données, celles-ci auront leurs propres caractérisations, avec une exigence de cohérence particulièrement élevée, ce qui met en exergue le rôle essentiel des métadonnées.

EN GUISE DE CONCLUSION

Il n'existe pas de donnée dans la nature. Pas la moindre. Pour l'exprimer en d'autres termes, les données ne sont pas données, il faut les construire, les prendre (*captum vs datum*). Elles requièrent en amont un travail de modélisation, d'abstraction, de spécification des concepts, puis des domaines, avant d'imaginer produire des valeurs. Elles sont dépendantes de choix eux-mêmes liés à des usages. Mais elles existent dans le vaste monde numérique, et il semble logique que la statistique publique s'interroge sur la manière de les utiliser efficacement pour ses propres besoins, quitte à en inventer d'autres qui soient aptes à nourrir le débat public[37].

C'est là un changement profond pour les statisticiens publics, même si ceux-ci ont une longue expérience d'utilisation de données administratives. Jusqu'au XIX^{e} siècle et même au début du XX^{e}, le recensement était la forme majeure de la statistique publique. Dans la deuxième moitié du XX^{e}, c'est l'enquête qui prend une place prépondérante, avec une maîtrise de bout en bout mobilisant tout un arsenal mathématique et technique. Le XXI^{e} siècle, sans renier les autres formes, serait celui des *data* : le statisticien est aussi informaticien, explorateur des contrées numériques, et, de fait, *data scientist*. Mais en relisant (Volle, 1980), on voit que des fondations demeurent : les répertoires, les nomenclatures, les codes, les définitions, les unités statistiques, tout ce qu'on regroupe aujourd'hui sous le vocable de *métadonnées*. L'évolution majeure de l'activité, outre l'importance prise par l'informatique, réside plutôt dans la nécessité de s'immerger dans les métiers d'origine de la donnée pour mieux s'en servir. C'est ce besoin d'ouverture aux processus externes (et pas seulement aux besoins des utilisateurs), cette polyvalence, cette agilité, cette curiosité teintée de grande rigueur, qui vont constituer la marque d'une nouvelle génération de statisticiens.

36. Exemple : dans le cas du statut vital, précédemment cité (Denis, 2018), on aura intérêt à laisser la donnée telle quelle en cas de décès, et peut-être à effectuer une imputation pour une personne présumée vivante au-delà d'un certain âge.
37. Par exemple les statistiques de mobilité pendant la crise sanitaire de 2020, qu'il aurait été extrêmement difficile et coûteux de bâtir avec des enquêtes classiques (cf. l'article de Jean-Luc Tavernier dans ce même numéro).

BIBLIOGRAPHIE

AMOSSÉ, Thomas, 2020. La nomenclature socioprofessionnelle 2020 – Continuité et innovation, pour des usages renforcés. In : *Courrier des statistiques.* [en ligne]. 27 juin 2019. N° N4, pp. 62-80. [Consulté le 1er décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4497076/courstat-4-5.pdf*.

BATINI, Carlo et SCANNAPIECO, Monica, 2016. *Data and Information Quality – Dimensions, Principles and Techniques.* Springer. ISBN 978-3-319-24104-3.

BECKER, Howard, 1952. Science, Culture, and Society. In : *Philosophy of Science.* Octobre 1952. The Williams & Wilkins Co. Volume 19, n° 4, pp. 273–287.

BERTI-ÉQUILLE, Laure, 2012. *La qualité et la gouvernance des données au service de la performance des entreprises.* Lavoisier-Hermes Science, Cachan. ISBN 978-2-7462-2510-7.

BONNANS, Dominique, 2019. RMéS, le référentiel de métadonnées statistiques de l'Insee. In : *Courrier des statistiques.* [en ligne]. 27 juin 2019. N° N2. [Consulté le 1er décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4168396/courstat-2-6.pdf.*

BORGMAN, Christine L., 2015. *Big Data, Little Data, No Data – Scholarship in the Networked World.* The MIT Press. ISBN 978-0-262-02856-1.

BOYDENS, Isabelle, 1999. *Informatique, normes et temps – Évaluer et améliorer la qualité de l'information : les enseignements d'une approche herméneutique appliquée à la base de données «LATG» de l'O.N.S.S.* Éditions E. Bruylant. ISBN 2-8027-1268-3.

BOYDENS, Isabelle, 2020. *Documentologie.* Presses Universitaires de Bruxelles, Syllabus de cours. ISBN 978-2-500009967.

BUCKLAND, Michael K., 1991. Information as Thing. In : *Journal of the American Society for Information Science.* [en ligne]. Juin 1991. 42:5. pp.351-360. [Consulté le 1er décembre 2020]. Disponible à l'adresse : *http://skat.ihmc.us/rid=1KR7VC4CQ-SLX5RG-5T39/BUCKLAND(1991)-informationasthing.pdf*.

CHRISTEN, Peter, 2012. *Data Matching – Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection.* Springer. ISBN 978-3-642-31164-2.

CNAV, 2020. *Cahier technique de la DSN 2021-1.* [en ligne]. 14 janvier 2020. [Consulté le 1er décembre 2020]. Disponible à l'adresse : *https://www.net-entreprises.fr/media/documentation/dsn-cahier-technique-2021.1.pdf*.

CODD, Edgar Franck, 1970. A Relational Model of Data for Large Shared Data Banks. In : *Communications of the ACM.* [en ligne]. Juin 1970. Volume 13, n° 6, pp. 377-387. [Consulté le 1er décembre 2020]. Disponible à l'adresse : *https://www.seas.upenn.edu/~zives/03f/cis550/codd.pdf*.

DENIS, Jérôme, 2018. *Le travail invisible des données – Éléments pour une sociologie des infrastructures scripturales.* [en ligne]. Août 2018. Presses des Mines, Collection Sciences Sociales. [Consulté le 1er décembre 2020]. Disponible à l'adresse : *https://books.openedition.org/pressesmines/3934?lang=fr*.

DESROSIÈRES, Alain, 1993. *La politique des grands nombres – Histoire de la raison statistique.* Réédité le 19 août 2010. Éditions La Découverte, collection Poche / Sciences humaines et sociales n°99. ISBN 978-2-707-16504-6.

ELMASRI, Ramez et NAVATHE, Shamkant B., 2015. *Fundamentals of Database Systems.* 8 juin 2015. Pearson, 7e édition. ISBN 978-0-13397077-7.

ESCARPIT, Robert, 1991. *L'information et la communication – Théorie générale.* 23 janvier 1991. Hachette Université Communication. ISBN 978-2-010168192.

FLORIDI, Luciano, 2010. *Information – A very short introduction.* Février 2010. Oxford University Press. ISBN 978-0-199551378.

GAROFALAKIS, Minos, GEHRKE, Johannes Gehrke et RASTOGI, Rajeev, 2016. *Data Stream Management – Processing High-Speed Data Streams.* Springer. ISBN 978-3-540-28607-3.

JURAN, Joseph M., 1951. *Quality-control handbook*. McGraw-Hill industrial organization and management series.

KITCHIN, Rob, 2014. *The Data Revolution – Big Data, Open Data, Data Infrastructures and Their Consequences*. SAGE Publications. ISBN 978-1-4462-8747-7.

LOSHIN, David, 2010. *The Practitioner's Guide to Data Quality Improvement.* 15 octobre 2010. Morgan Kaufmann. ISBN 978-0-080920344.

MARTIN, Olivier, 2020. *L'empire des chiffres.* 16 septembre 2020. Éditions Armand Colin. ISBN 978-2-20062571-9.

OLSON, Jack E., 2003. *Data Quality – The Accuracy Dimension*. [en ligne]. Janvier 2003. Morgan Kaufmann. ISBN 1-55860-891-5.

REDMAN, Thomas C., 1997. *Data Quality for the Information Age*. Janvier 1997. Artech House Computer Science Library. pp. 227-232. ISBN 978-0-89006-883-0.

RENNE, Catherine, 2018. Bien comprendre la déclaration sociale nominative pour mieux mesurer. In : *Courrier des statistiques.* [en ligne]. 6 décembre 2018. N° N1, pp. 35-44. [Consulté le 1er décembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3647029/courstat-1-7.pdf*.

RIVIÈRE, Pascal, 2018. Utiliser les déclarations administratives à des fins statistiques. In : *Courrier des statistiques.* [en ligne]. 6 décembre 2018. N° N1, pp. 14-24. [Consulté le 1erdécembre 2020]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3647013/courstat-1-5.pdf*.

SADIQ, Shazia, 2013. *Handbook of Data Quality : Research and Practice*. Springer. ISBN 978-3-642-36257-6.

SHANNON, Claude Elwood, 1948. A Mathematical Theory of Communication. In : *The Bell System Technical Journal*. Juillet 1948. Volume 27, N° 3, pp. 379-423.

SHANNON, Claude Elwood, 1953. The lattice theory of information. In : *Transactions of the IRE Professional Group on Information Theory.* Février 1953. Volume 1, n° 1, pp.105-107.

TRICLOT, Mathieu, 2014. *Le moment cybernétique – La constitution de la notion d'information.* Champ Vallon. ISBN 978-2-876736955.

VOLLE, Michel, 1980. *Le métier de statisticien.* [en ligne]. Éditions Hachette Littérature. ISBN 978-2-010045295. [Consulté le 1er décembre 2020]. Disponible à l'adresse : *http://www.volle.com/ouvrages/metier/tabmetier.htm*.

WEINBERGER, David, 2012. *Too Big to Know.* 1er janvier 2012. Éditions Basic books, New York, p.2. EAN 978-0-465021420.

WIENER, Norbert, 1948. *Cybernetics – Or Control and Communication in the Animal and the Machine*. 1961, 2e édition. The MIT Press, Cambridge, Massachusetts. ISBN 978-0-262-73009-9.