Quels formats pour quelles données ?

Alexis Dondon, Pierre Lamarche***

La donnée, pour être intelligible par ses utilisateurs et accomplir sa fonction de transport de l'information, doit être structurée. Une telle structure se formalise au travers d'un modèle de données, qui conditionne le choix du format. Les formats de données sont variés et répondent à des problématiques spécifiques, différentes selon le contexte de l'utilisation de la donnée par le métier. Dans l'idéal, les standards sous-jacents aux modèles de données et les formats associés sont partagés par le plus grand nombre d'utilisateurs. S'agissant de la donnée statistique, ces problématiques sont localisées dans les objets pour lesquels les données sont susceptibles de véhiculer de l'information, mais également dans la documentation de la donnée – la métadonnée – ou encore dans la volonté d'associer à la donnée des solutions logicielles spécifiques particulièrement adaptées à son traitement.

Sur ce dernier point, l'émergence ces dernières décennies de solutions open-source a permis de concilier les notions de sécurisation de l'import de la donnée, d'efficience de son traitement, de reproductibilité, etc. En particulier, des formats comme Parquet s'intègrent à des solutions logicielles accessibles à tous et adoptées par une communauté de plus en plus large, convaincue de ses avantages. Néanmoins, il n'existe pas de réponse définitive et unique pour le choix d'un format : des choix sont faits après une analyse précise des besoins relatifs à chaque étape du cycle de vie de la donnée. En cela, le choix d'un format est l'expression concrète d'un standard dicté par des impératifs propres à chacune de ces étapes.

In order to be intelligible to its users and to fulfil its function of conveying information, the data must be structured. This structure is then formalised through a data model, which determines the choice of format. Data formats are various and address specific problems, according to the context of use. Ideally, the standards behind the data models and the associated formats are shared as many users as possible. For statistical data, these problems are located in the objects for which the data are likely to convey information, but also in data documentation - i.e. metadata - or in the wish to link specific software solutions with the data particularly well suited to its processing.

On this last point, the emergence over the last few decades of open-source solutions has made it possible adress simultaneously different issues securing data import, efficiency of data processing, replicability, etc. In particular, formats such as Parquet are integrated into software solutions accessible to all and adopted by an increasingly large community, convinced of their advantages. Nevertheless, there is no clear-cut answer for the choice of a format: choices are made after a precise analysis of the needs relative to each step of the data's life cycle. In this way, the choice of a format is the concrete expression of a requirement driven standards specific to each of these phases.

* Data Engineer, DSI, Insee,
alexis.dondon@insee.fr

** Chef de la division Logement et Patrimoine, DSDS, Insee,
pierre.lamarche@insee.fr

La donnée, en tant qu'émanation concrète de l'information, est un élément-clé dans une société qui fonde la prise de décision sur l'information ; elle doit être transmissible et être codifiée selon des conventions établies et partagées (*Warnier, 1974*). La notion de transmissibilité s'entend au sens où le récepteur de la donnée doit être en mesure de la lire et de la comprendre. La donnée statistique se définit, quant à elle, comme une donnée à usage statistique (*Sundgren, 2010*), c'est-à-dire avec pour but l'énumération et la quantification de phénomènes sous forme codifiée. Dans un monde où la donnée est de plus en plus présente et massive, l'usage de celle-ci par des systèmes informatisés requiert l'adoption de standards et de conventions de stockage partagés, adaptés à son usage et aisément reproductibles. Ainsi, toute personne voulant utiliser, modifier ou enregistrer des données se trouve confrontée à la question des conventions, constantes dans le temps et l'espace, à adopter pour les lire ou les stocker sous un format adapté à l'usage qu'elle souhaite en faire. De la même manière qu'un langage véhicule de l'information qui se transmet si celui-ci est connu du locuteur et de l'auditeur, les conventions adoptées pour le stockage des données doivent être partagées entre utilisateurs.

▶ Modéliser l'information pour créer la donnée

Avec l'avènement de l'informatisation et de la numérisation de la donnée, la question de sa transmissibilité s'est très rapidement posée. Les réponses se trouvent dans la définition de standards et de conventions de stockage et de transmission.

> ***L'établissement de conventions partagées est essentiel pour une communication efficace et fluide entre différents acteurs d'un système d'information.***

La modélisation de la donnée répond à un besoin de définir un langage commun. Elle est étroitement liée à la nature de l'information : dans un cas fictif de données individuelles, il faut définir les attributs de chaque individu (nom, prénom, date de naissance, lien de parenté) ainsi que leur représentation. Une fois cette modélisation réalisée, un ensemble de formats pourra être mobilisé. Par ailleurs, l'établissement de conventions partagées est essentiel pour une communication efficace et fluide entre différents acteurs d'un système d'information. L'intégration d'un réseau qui partage une même convention peut, à ce titre, devenir un élément stratégique, voire vital. Plus cette convention est partagée et utilisée par un grand nombre d'utilisateurs, plus sa valeur *ipso facto* est importante. À titre d'exemple, les données de la Déclaration Sociale Nominative[1] font l'objet d'une modélisation et de l'adoption de standards par l'ensemble des parties prenantes de manière à permettre des échanges d'informations massifs[2].

Le standard constitue une réponse à un besoin de structuration de l'information dans un contexte donné. Il va définir une manière de présenter la donnée qui permet de répondre à des problématiques précises, essentielles pour les professionnels qui utilisent ces informations. Le format est la concrétisation de ce standard,

1 La Déclaration sociale nominative (DSN) est une déclaration obligatoire pour les entreprises du secteur privé qui sert à payer les cotisations sociales et à transmettre les données sur les salariés aux organismes sociaux.
2 Voir à ce sujet l'article sur les normes d'échanges dans le même numéro.

et à un standard donné peuvent correspondre plusieurs formats. Cependant, le standard ne fait pas tout et le champ des formats possibles peut s'avérer très large.

La représentation tabulaire, un modèle canonique pour la statistique

La sphère statistique a cherché très vite à définir des formats de données, avec des spécificités propres à ses activités. La donnée est évidemment centrale pour la statistique. Elle peut prendre une grande variété de formes tant les sources d'information utilisées sont diverses. Derrière cette question, se situe le sujet de la modélisation de la donnée, élément clé s'agissant des systèmes de bases de données utilisés par le statisticien, mais également pour le stockage de la donnée inerte[3].

> ***La donnée est évidemment centrale pour la statistique.***

La statistique a recours, dans de nombreux domaines, aux micro-données[4], que ce soient des données d'enquête ou encore des données de registres. Leur numérisation doit satisfaire les contraintes de volume et de calcul. Lorsque les capacités de mémoire des machines sont relativement limitées, les formats ouverts et peu consommateurs d'espace vont être privilégiés. La statistique s'accommode très naturellement des formats tabulaires[5], puisque cette façon de structurer la donnée correspond à l'approche théorique de la statistique. Dans la représentation tabulaire de la donnée, la ligne va correspondre de manière canonique aux observations, et les colonnes aux variables. Cette notion peut être considérée comme une application particulière d'une vision matricielle de l'information ; elle est très intimement associée dans le monde de la statistique à des outils, des modes de pensée et des méthodes de calcul largement partagés et très structurants.

Ainsi, les formats tabulaires sont privilégiés : la modélisation de la donnée va alors consister à définir des tables (ou objets) et des liens entre elles (des relations), comme pour les systèmes de gestion de base de données (ou SGBD, ***encadré 1***). Cependant, les SGBD sont particulièrement adaptés à la donnée « vivante », qui a vocation à évoluer fréquemment, et relèvent également de choix d'infrastructure informatique qui dépasse le champ de la statistique. Les notions spécifiques aux SGBD ne sont pas développées ici, contrairement aux données inertes et aux formats de stockage associés à ce type de données.

3 Qui se définit comme une donnée qui n'a pas vocation à être modifiée, mais simplement lue.
4 Au sens «données individuelles», par opposition aux données agrégées comme les statistiques.
5 Qui s'entend comme une présentation de la donnée en ligne et en colonne.

La fin définit le moyen en matière de format

L'utilisation d'un format de stockage doit assurer une qualité essentielle de la donnée statistique, son accessibilité. Le statisticien doit s'assurer que les données qu'il produit et qu'il utilise puissent être aisément réutilisées, par lui-même ou par d'autres utilisateurs, et ce en minimisant les pré-requis informatiques à la lecture et au traitement de ces données. Ce besoin est renforcé dans le contexte d'*open-data* où la donnée a de plus en plus vocation à être partagée de manière complète avec le plus grand nombre d'utilisateurs, et où la question de l'accessibilité de cette donnée est incontournable.

> ***Le format dépend avant toute chose de l'usage de la donnée et également de son volume.***

Idéalement, la donnée doit être accessible et adaptée au contexte informatique de son usage (dit autrement, on va chercher à minimiser la consommation de ressources informatiques au sens large *(Nordbotten, 1966)*). Pour résoudre cette quadrature du cercle, il n'y a pas de réponse unique ; le format dépend avant toute chose de l'usage de la donnée et également de son volume. Il convient de distinguer les micro-données, qui portent sur des observations individuelles (et qui sont souvent des données volumineuses) des données agrégées, qui résultent d'un premier exercice d'agrégation ou d'estimations.

Encadré 1 : Les bases de données, un format adapté à la donnée « vivante »

Les systèmes de gestion de base de données (SGBD) sont des solutions logicielles qui permettent de stocker, d'utiliser et de traiter des données généralement massives ; elles offrent des solutions optimisées en matière de consommation de ressources informatiques pour traiter des volumes significatifs de données. En contrepartie, elles sont constituées d'un système de gestion de fichier dont la complexité est masquée à l'utilisateur, qui a la possibilité de lancer des requêtes sur ces données à l'aide d'outils standards*. Elles vont optimiser l'utilisation de la donnée par des systèmes d'indexation des observations et de hachage des identifiants, permettant ainsi d'appréhender un large spectre de modèles de données. Le plus courant et le plus utilisé dans la statistique publique est le modèle tabulaire, incluant généralement une notion de liens entre tables (système de gestion dit relationnel)**.

Historiquement, les SGBD les plus largement utilisés sont d'abord des solutions propriétaires telles que Oracle ou MySQL. Dans la dernière décennie, les SGBD *open-source* se sont progressivement imposés dans l'univers de la donnée ; l'un des plus connus est PostgreSQL.

Les SGBD sont souvent des solutions de stockage adaptées à la donnée « vivante », c'est-à-dire à une donnée évolutive, pour laquelle ces derniers apportent des garanties en matière de préservation de l'intégrité et de la cohérence des données et de réversibilité des traitements opérés. En contrepartie, cette solution de stockage de la donnée peut être très énergivore, comparativement à d'autres. Elles doivent donc être considérées surtout pour leur usage courant, celui du stockage et du traitement de la donnée vivante. Le stockage de la donnée inerte sur un SGBD a peu d'intérêt, et peut être considéré comme un gaspillage énergétique, dans une démarche d'informatique durable***, notion appelée à devenir incontournable (Ademe, 2021).

* *Bien souvent à l'aide d'un langage standard, le SQL (Structured Query Language), qui peut varier dans ses subtilités selon le type de SGBD utilisé.*

** *Elles peuvent également gérer d'autres types de données, par exemple des données où les champs varient d'une observation à l'autre (on parle de documents flexibles).*

*** *Il existe néanmoins des SGBD adaptés au stockage de données inertes, avec une faible consommation énergétique associée.*

Naturellement, le traitement d'une information agrégée ou d'une information individuelle n'a pas les mêmes implications en matière de stockage de la donnée. Lorsqu'on veut choisir un format, il est donc essentiel de répondre aux questions suivantes :

- quel volume ?
- quels usages, pour quels utilisateurs ?
- quelle localisation de la donnée ?

Les formats textuels s'accommodent naturellement de la représentation tabulaire

Les formats textuels sont les formats les plus simples à utiliser, puisqu'ils peuvent être aussi bien interprétés par une machine que lus par un humain[6]. Les fichiers à largeur fixe ou les formats avec délimiteur (en particulier le format comma-separated values ou csv ***(figure 1b)***) permettent aisément de compacter la donnée sous un format tabulaire. Si le format à délimiteur peut se suffire à lui-même, les fichiers à largeur fixe supposent de disposer d'une information sur la position de chaque colonne ; la description des données prend ici toute son importance : sans elle, il est très difficile de retrouver la structure de la donnée, et donc de la traiter. Historiquement, le format positionnel fixe ***(figure 1a)*** a été très utilisé, car il consomme peu d'espace, en contrepartie d'une description de sa structure. Il continue à être utilisé, par exemple pour les transmissions de données entre certaines administrations et l'Insee. L'exemple de données contenant une liste d'individus et de liens de parenté entre ceux-ci nécessite cependant une description précise de la position de chaque variable.

> ***Le format avec délimiteur est utilisé fréquemment, du fait de sa simplicité d'utilisation.***

Le format avec délimiteur est utilisé fréquemment, du fait de sa simplicité d'utilisation. Il présente cependant des inconvénients, avec en particulier la contrainte de l'adoption d'un délimiteur absent des données qu'il structure[7]. Le format avec délimiteur suppose une prise de risque relative à l'intégrité de la donnée, tout comme le format à largeur fixe et de manière générale les formats textuels impliquant une opération d'import et donc d'interprétation de la structure de la donnée. Ce risque est encore plus élevé si les données ne sont pas correctement documentées ; il est question de dictionnaire des codes, de dessin de fichiers, et plus généralement d'un ensemble de documents qui permet de traiter et d'interpréter la donnée en cohérence avec les intentions des producteurs de cette donnée. Par ailleurs, les formats textuels ne concernent pas uniquement les modèles de données tabulaires ; au contraire, les modèles de données plus « élastiques » peuvent être servis par certains formats textuels de manière plus sécurisée.

6 On parlera de format « *human readable* » dans la terminologie anglo-saxonne.
7 Par exemple, si on utilise le caractère «/» comme délimiteur, il faut s'assurer qu'aucune donnée ne puisse contenir ce caractère.

▶ Figure 1 - Quelques exemples de formats

FIGURE 1A - POSITIONNEL FIXE

```
000011MARTIN      ARTHUR                0806195575014
000012MARTIN      JEANNE                1503195607010
000021MARTIN      CLAUDE    00001100001218101984750l4
...
```

FIGURE 1B - FORMAT CSV

```
ID_IND;NOM;PRENOMS;ID_PERE:ID_MERE:DATE_NAIS;COM_NAIS
000011;MARTIN;ARTHUR;;08061955;75014
000012;MARTIN;JEANNE;;15031956;07010
000021;MARTIN;CLAUDE;000011;000012;18101984;75014
...
```

FIGURE 1C - FORMAT XML

```
<OBSERVATION>
    <ID>000021>/ID
    <NOM>MARTIN</NOM>
    <PRENOMS>CLAUDE</PRENOM>
    <DATE NAIS>18101984</DATE NAIS>
    <LIEU NAIS>75014</LIEU NAIS>
    <PERE>
            <D>000011</ID>
            <NOM>MARTIN</NOM>
            <PRENOMS>ARTHUR</PRENOMS>
            <DATE NAIS>08061955</DATE NAIS>
            <LIEU NAIS>75014</LIEU NAIS>
    </PERE>
    <MERE>
            <ID>000012</ID>
            <NOM>MARTIN</NOM>
            <PRENOMS>JEANNE</PRENOMS>
            <DATE NAIS>15031956</DATE NAIS>
            <LIEU NAIS>07010</LIEU NAIS>
</MERE>
</OBSERVATION>
...
```

FIGURE 1D - FORMAT JSON

```
[
    {
    “ID”: “000021”,
    “NOM”: “MARTIN”,
    “PRENOMS”: “CLAUDE”
    “DATE NAIS”: “18101984”
    “LIEU NAIS”: “75014”,
    “PERE”:
            {
            “ID”:”000011”
            “NOM”: “MARTIN”
            “PRENOMS”:”ARTHUR”
            “DATE NAIS”:”08061955”
            “LIEU NAIS”: “75014”
            },
    “MERE”:
            {
            “ID”:”000012”,
            “NOM”: “MARTIN”
            “PRENOMS”: “JEANNE”
            “DATE NAIS”: “15031956”
            “LIEU NAIS”: “07010”
    },
...
]
```

▶ Des formats « élastiques » pour les documents flexibles —

Les données ne se présentent pas toujours sous un format tabulaire. Ainsi, certains modèles de données peuvent introduire de la flexibilité dans les champs, ou dans la notion même d'observations et de variables. Le format JSON[8] par exemple ***(figure 1d)*** est un format minimaliste qui introduit ce type de flexibilité. Il va en particulier s'accommoder de données contenant des champs largement facultatifs, comme les données du fichier permanent des occurrences de traitement des émissions (POTE)[9] *(Lamarche et Lollivier, 2021)* relatives à la déclaration des revenus, pour laquelle un grand nombre de cases reste vide dans la plupart des observations. Format très flexible mais plus consommateur d'espace, XML ***(figure 1c)*** est particulièrement adapté au stockage de données à faible volume, c'est-à-dire essentiellement les macro-données ou données agrégées. Il permet également de stocker les métadonnées qui fournissent les informations contextuelles relatives à la donnée. Caractéristique de la technologie Web, XML fonctionne sous forme de balises : il s'agit donc là encore d'un format textuel, pour lequel certains caractères spéciaux permettent d'identifier la structure. Son avantage principal est que le format s'adapte à de nombreuses représentations de la donnée, pas nécessairement tabulaires. Il permet également de mettre dans un même réceptacle la donnée et des éléments descriptifs de celle-ci ; de fait, il donne la possibilité de coupler de manière très naturelle la donnée et la métadonnée. Les apports de ces deux formats vis-à-vis des formats tabulaires traditionnels sont importants, en particulier dans leur capacité à embarquer des relations entre les tables (par exemple la filiation entre une table de parents et une table d'enfants), de manière très naturelle[10], tout en s'accommodant des contraintes des formats tabulaires.

> ***Ces formats très flexibles et auto-suffisants sont parfaitement adaptés pour mettre à disposition des données.***

De manière générale, ces formats très flexibles et auto-suffisants sont parfaitement adaptés pour mettre à disposition des données. Ainsi, un grand nombre d'API[11] offrent la possibilité aux utilisateurs de consommer soit de la donnée, soit des services de traitement de la donnée, en adoptant ces formats. La flexibilité est un avantage précieux qui permet d'accompagner l'information d'un ensemble d'éléments contextuels, assimilable à des métadonnées, en informant l'utilisateur sur la nature de la donnée récupérée ou encore sur la qualité du traitement dont elle a bénéficié. La présentation sous forme de table est moins décisive, car le requêtage d'une API se fait de ligne à ligne : le résultat de la requête est important ainsi que sa description. En revanche, le volume de données sollicitées est nécessairement limité. Même si la finalité est d'obtenir des données sous une représentation tabulaire, le traitement élémentaire se réalise dans un cadre qui ne nécessite pas cette représentation.

8 JSON : JavaScript Object Notation.
9 Le POTE est le « fichier permanent des occurrences de traitement des émissions », élaboré par les services de la DGFiP à partir des émissions des avis d'imposition sur les revenus.
10 Sous forme de données emboîtées ou encapsulées.
11 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.

Ces formats sont très adaptés pour des jeux de données de faible volume, mais dont la structure peut être évolutive, ou tout du moins difficile à établir *a priori*. Pour traiter de gros volumes, il est nécessaire d'adopter des modèles de données plus contraignants et le modèle tabulaire est bien souvent adapté ; il faut alors chercher l'efficacité dans la capacité à interpréter rapidement les caractéristiques des données.

La place cruciale des métadonnées dans la diffusion rend ces formats incontournables et les statisticiens s'y trouvent naturellement confrontés dans les processus de production.

La métadonnée, un élément de plus en plus incontournable

Les métadonnées peuvent contribuer fortement à donner plus de sens aux données qu'elles décrivent, car elles créent des liens entre les différentes sources de données. Nombre d'initiatives au niveau international visent à définir des standards et des formats associés partagés par le plus grand nombre d'acteurs ***(encadré 2)***.

De manière plus pragmatique, le code de bonnes pratiques de la statistique européenne mentionne les métadonnées comme un élément de la qualité des statistiques mises à disposition de l'usager : fournir aux utilisateurs de l'information sur la donnée elle-même est un critère très important pour juger de son accessibilité et de sa clarté. Si la communauté statistique européenne donne à ce terme de métadonnée un sens précis, celui-ci recouvre en réalité un large spectre d'informations relatives à la donnée. Le premier élément est la description de la donnée en termes informatiques :

Encadré 2 : le standard DDI et le format SDMX

Les formats incorporant des métadonnées sont très souvent des formats « élastiques », qui permettent de relier la métadonnée avec la donnée de manière naturelle. Ils sont spécifiques, car ils comportent une couche d'abstraction supplémentaire et embarquent la sémantique propre à la métadonnée. Acteur important de définition d'une sémantique partagée pour la métadonnée, la Data Documentation Initiative (DDI) est un consortium international d'instituts de recherche et de producteurs de statistiques qui vise à définir des standards pour la documentation des données statistiques, avec un focus particulier sur les données d'enquêtes, des méthodes de collecte et des référentiels (nomenclatures, codifications, etc.) utilisés pour la collecte. Ce consortium existe depuis 1995 : il travaille sur la définition d'un format de stockage de ces métadonnées visant à créer un éco-système d'outils, à la fois en amont de la collecte et en aval pour la diffusion des données produites. Cette vision intégrée et l'adoption des outils produits par cet éco-système doivent bénéficier à la qualité des données produites et à l'utilisateur final tout autant qu'au producteur. La pertinence de ces outils, leur simplicité et leur exhaustivité sont déterminants pour qu'un format s'impose.

SDMX* est une initiative impliquant des instituts de statistiques et des organismes internationaux, visant en particulier à renforcer le caractère transmissible de la donnée. Le format SDMX s'appuie sur la métadonnée, car la description standardisée de la donnée facilite sa mobilisation. Là aussi, un éco-système d'outils accompagne l'adoption du standard (Salou et Sosnovsky, 2010) : les institutions internationales comme Eurostat**, la BCE***, la BIS**** ou encore l'OCDE***** poussent pour l'adoption de ces standards, car leur adoption rend possible la résolution de nombreux sujets relatifs à l'agrégation de données et aux questions d'entrepôts de données, sujets très importants pour ces instituts. Par ailleurs, l'adoption la plus massive possible de standards communs permet de consolider les connaissances et d'accroître les potentialités de relier les données entre elles.

** SDMX : Statistical Data and Metadata eXchange.*

*** Eurostat est l'institut statistique communautaire, associé à la Commission européenne.*

**** La BCE est la Banque centrale européenne.*

***** La BIS est la Bank for International Settlements ou Banque des règlements internationaux.*

****** L'OCDE est l'Organisation de coopération et de développement économiques.*

> ***Il faut distinguer la métadonnée qualifiée de technique de la métadonnée documentaire.***

quel type des variables présentes dans les tables, quelles unités de mesure, quelle largeur de colonnes pour les variables en format caractère, etc. Mais le champ couvert par les métadonnées va beaucoup plus loin, et doit permettre de documenter le processus de génération des données de manière générale. Il faut distinguer la métadonnée qualifiée de technique, qui donne de l'information sur le type des variables par exemple, de la métadonnée documentaire, qui informe l'utilisateur sur le contexte de production de la donnée, les choix méthodologiques, les nomenclatures adoptées, etc. Cas particulier de cette dernière, la notion de ligne de données permet de suivre le cycle de vie de la donnée, en traçant les sources à l'origine des données et les transformations subies par celles-ci pour aboutir au résultat. L'ensemble de ces métadonnées constitue un enjeu essentiel dans un contexte où les statistiques produites sont de plus en plus utilisées dans le cadre de comparaisons internationales, et pour lesquelles il devient crucial pour l'utilisateur de comprendre les raisons de divergences potentielles liées à des modes de collecte radicalement différents. Ainsi, le sujet de la métadonnée est porté par les institutions internationales traitant de la production statistique, au premier plan desquelles Eurostat.

S'agissant des données statistiques, ce mouvement d'internationalisation des conventions s'inscrit dans un contexte où les données, de par leur nature numérique, ont de plus en plus vocation à s'appuyer sur des technologies *web*. La flexibilité de ces technologies pour le stockage des données, et en particulier leur caractère adaptatif à la structure que celles-ci peuvent prendre, favorise leur adoption dans les standards ; en particulier, le consortium *Data Documentation Initiative* (DDI) repose sur le format XML pour la définition de ses conventions de stockage.

Tabulaires ou flexibles, les formats textuels ont pour grand avantage leur lisibilité et une forme de stabilité dans le temps. Sauf modification très profonde du système informatique, les données stockées sous ces formats seront lisibles et mobilisables longtemps après, et ce avec d'autant plus de facilité si les données contiennent des métadonnées. Cette disponibilité dans le futur a un coût dans le présent : le stockage de données toujours plus nombreuses s'accommode assez mal des contraintes qu'un format textuel implique. Il est donc essentiel d'introduire des formats qui, s'ils apportent moins de garantie de stabilité, permettent en revanche de répondre aux problématiques présentes.

La nouvelle donne de la donnée : volume, vélocité, variété, virtualisation, transversalité

Les statisticiens sont concernés par plusieurs sujets, qui ont directement un impact sur les choix de formats. L'adoption par le plus grand nombre de conventions et de standards est une source d'efficience qui bénéficie directement à tous les acteurs qui font le choix de les adopter. De ce point de vue, il est nécessaire d'intégrer les problématiques qui traversent l'ensemble du monde de la donnée et les choix qui en découlent de la part de la plupart des acteurs.

Tout d'abord, l'expansion continue de la donnée impose de développer des solutions logicielles permettant de traiter des volumes toujours plus importants de la manière la plus efficace possible. Le *Big data*, ou donnée massive, a mis sur le devant de la scène les préoccupations de stockage et de performance auxquelles des logiciels propriétaires tels que SAS® ou Oracle avaient apporté de premières réponses. En particulier, le fait que cette donnée massive s'accompagne de plus en plus d'une structuration flexible de l'information impose des formats alternatifs au format tabulaire historique.

Cette expansion s'accompagne d'un recours toujours plus important à la virtualisation des traitements et du stockage, avec l'avènement du *cloud computing*. Ce principe conditionne les solutions techniques et les formats associés. Une solution de stockage de type S3[12] est ainsi souvent associée à des outils de requêtage optimisés pour certaines catégories de format. Son adoption va donc de pair avec ces outils.

Par ailleurs, l'efficience budgétaire est un élément important qui préside souvent au choix de solutions *open source* et de formats ouverts que celles-ci impliquent, considérées comme étant moins coûteuses pour les processus de production. Toutefois, si l'*open source* est un vecteur très intéressant de mutualisation des investissements, il n'est pas une ressource disponible gratuitement, pour deux raisons essentiellement. D'une part, il faut disposer des compétences permettant de maîtriser ces outils ; d'autre part, le développement de ces outils a un coût, et la sécurisation des processus fondés sur ces outils nécessite une identification de ces développements et une maîtrise de l'assurance que ceux-ci se poursuivent dans le temps.

> ***Le renforcement nécessaire de la confiance dans la statistique officielle passe nécessairement par une attention plus forte aux notions de transparence et de reproductibilité.***

Plus important encore, dans une société concernée par les sujets de *fact-checking* et de *fake news*, le renforcement nécessaire de la confiance dans la statistique officielle passe nécessairement par une attention plus forte aux notions de transparence et de reproductibilité. Ces notions sont le moteur d'un recours toujours plus fort à la donnée ouverte ou *open data*, et à la possibilité donnée à des utilisateurs de reproduire de la manière la plus complète possible les processus aboutissant à la production de la donnée sous son état final. L'*open data* implique par essence le recours à des formats ouverts, et la reproductibilité à des solutions logicielles *open source*, par définition transparentes vis-à-vis des algorithmes qu'elles mettent en œuvre.

Enfin, l'approche historique, avec un unique logiciel statistique comme « couteau suisse » permettant de réaliser l'intégralité du traitement de la donnée, est progressivement remplacée par une vision plus fragmentée ; chaque étape du processus (que l'on peut modéliser dans le cadre du GSBPM[13]) est mise en œuvre grâce à une solution logicielle dédiée et spécifique, pour laquelle l'outil le plus approprié sera choisi. De ce point de vue,

12 Une solution de stockage développée par Amazon qui tend à s'imposer comme un standard dans le monde de la donnée.

13 Le modèle générique de description des processus de production statistique (GSBPM pour Generic Statistical Business Process Model) décrit les différentes étapes à suivre pour produire des statistiques publiques.

> ***Chaque étape du processus est mise en œuvre grâce à une solution logicielle dédiée et spécifique, pour laquelle l'outil le plus approprié sera choisi.***

la complémentarité entre les langages de programmation R et Python est souvent mise en avant ; elle suppose l'adoption d'un format de données transversal, qui permet de basculer aisément d'un outil à l'autre sans les phases d'import-export de données fastidieuses et périlleuses. Ce type de format est un élément de sécurisation du processus dans son ensemble. Les formats de stockage ont ainsi évolué, nécessairement vers des solutions fondamentalement nouvelles.

▶ Le format binaire, une efficience sous contrainte

L'import de données à partir de fichiers texte est une tâche risquée, compliquée ou coûteuse en calculs : la sécurisation de ce processus d'import passe par sa minimisation (on le réalise le moins de fois possible). À ces formats ouverts, lisibles par un œil humain, on oppose les formats de stockage de données binaires, qui contiennent la donnée sous un format autre que textuel. Cet impératif de stocker la donnée sous un format illisible pour un simple éditeur de texte permet de répondre à un autre besoin, celui de présenter la donnée de la manière la plus rapidement interprétable par la machine. La conversion de l'information entraîne la transformation des données en une séquence non directement déchiffrable de morceaux élémentaires de donnée informatique, les bits. Ces bits ne prennent que deux valeurs possibles, 0 ou 1 : on parlera alors de "format binaire". On distingue ainsi deux manières de stocker de l'information en informatique, selon qu'on puisse le déchiffrer par un éditeur de texte (format ouvert) ou pas (format binaire).

Lorsque le format binaire est un format propriétaire, il ne peut généralement être déchiffré que par les outils associés à ce format. Les formats propriétaires présentent des avantages considérables en matière d'utilisation de la donnée, propres au format binaire : accessibilité quasi immédiate à celle-ci, et bien souvent à un grand nombre de métadonnées stockées de manière native, permettant de décrire la donnée.

Les formats propriétaires (au premier rang desquels SAS®) ont rapidement émergé dans les activités où la donnée était un élément central. Les solutions logicielles associées à ces formats se sont imposées par leur capacité à traiter des volumes importants de données de manière optimisée, en fournissant à l'utilisateur une sorte de « couteau suisse » permettant à la fois de traiter la donnée et de lui appliquer des procédures statistiques proches de l'état de l'art. On a ainsi une forte cohérence entre le format de stockage et la solution logicielle adoptée par la statistique publique pour la production et l'analyse de la donnée.

Dans un contexte d'augmentation du volume de données, le choix d'un outil comme SAS®, qui les traite comme un système de fichiers, a permis d'apporter une première réponse au défi de calcul que de telles données ont pu poser aux machines. Cette vision a longtemps présenté des avantages dépassant largement les inconvénients que peuvent constituer l'adoption d'un format par essence fermé, imposant à l'ensemble des utilisateurs l'usage de la solution logicielle associée, ou encore la modélisation purement tabulaire,

et donc par essence frustre, de la donnée. Elle a néanmoins été battue en brèche ces dernières décennies par une demande d'ouverture de la donnée, par une exigence de reproductibilité, et par l'apparition dans le paysage de véritables écosystèmes de logiciels libres, associant ouverture et performance.

Combiner les avantages : les formats binaires et *open-source*

Si les formats propriétaires ont permis d'associer un format à une solution logicielle, il existe des formats binaires développés en *open-source*, qui ont pour but de répondre au seul besoin d'optimiser leur utilisation par la machine, tout en relâchant la contrainte de la solution logicielle.

De tels formats sont souvent inter-opérables par différentes solutions logicielles, souvent elles aussi *open-source*. C'est la première contrainte à laquelle doit s'adapter le format de stockage convenant au mode actuel d'utilisation des données : sa capacité à s'abstraire d'une solution logicielle définie *a priori* (le « couteau suisse » mentionné précédemment), pour que l'utilisateur puisse choisir l'outil le plus adapté au traitement qu'il veut réaliser, dans un contexte où celui-ci s'insère dans une chaîne incorporant des traitements de natures et de finalités variées. En résumé, l'utilisateur doit pouvoir recourir à une palette d'outils divers sans pour autant modifier à chaque changement d'outil la nature de l'objet qu'il traite.

> ***Dans le monde de la statistique, la donnée est très souvent une donnée que l'utilisateur va lire plus fréquemment qu'il ne la modifie.***

Par ailleurs, une statistique est une donnée qui sera bien plus fréquemment lue que modifiée. De ce point de vue, le format de stockage peut s'apparenter au principe WORM (*Write Once, Read Many*) : le format doit chercher à optimiser le processus de lecture qu'en fait l'utilisateur, moins que son processus de transformation et d'écriture.

Ce principe est d'autant plus pertinent dans un contexte où l'infrastructure correspond à l'état de l'art, avec une virtualisation du traitement de la donnée et de son stockage, accompagnée d'un découplage de ces deux notions : le stockage et le calcul se déroulent sur des infrastructures différentes, tout en maintenant une grande capacité à mobiliser très rapidement les données pour les traiter ensuite.

Enfin, s'agissant des données massives, dans un contexte contraint où il devient impératif de consommer l'énergie avec parcimonie et discernement, la donnée stockée sous forme de bases de données énergivores doit l'être pour de bonnes raisons : il s'agit de pouvoir transformer cette donnée de manière sécurisée et à grande vitesse. En règle générale, la donnée inerte suffit à couvrir de manière très satisfaisante la plupart des usages analytiques ou de production des statisticiens.

Néanmoins, d'autres problématiques ont vu le jour dans le paysage de la statistique : la nécessité, en matière d'efficience cette fois, d'adopter des standards et des conventions les plus partagées possibles et des solutions de calcul toujours plus complexes, en raison de l'expansion considérable des données disponibles.

▶ Parquet, un format compact et décomposable

Les formats répondant à ces impératifs ont émergé ces dernières années avec l'avènement de solutions intégrées de traitement de la donnée massive, telles Hadoop[14]. En particulier, le format Parquet ***(figure 2)*** permet de solliciter de manière très naturelle la donnée de façon parallélisée, c'est-à-dire en la scindant, en la distribuant très rapidement à plusieurs unités de traitement et en la traitant de cette manière en parallèle. Le nom de ce format résume à lui seul ses propriétés : schématiquement, la donnée va être stockée sous forme de « lames » denses, plus ou moins fortement compressées et mobilisables chacune de manière indépendante. Ce format permet de stocker la donnée grâce à différents algorithmes de compression, qui réduisent de manière très significative la taille des données sans dégrader la vitesse à laquelle cette donnée peut être mobilisée *(Uber, 2022)*.

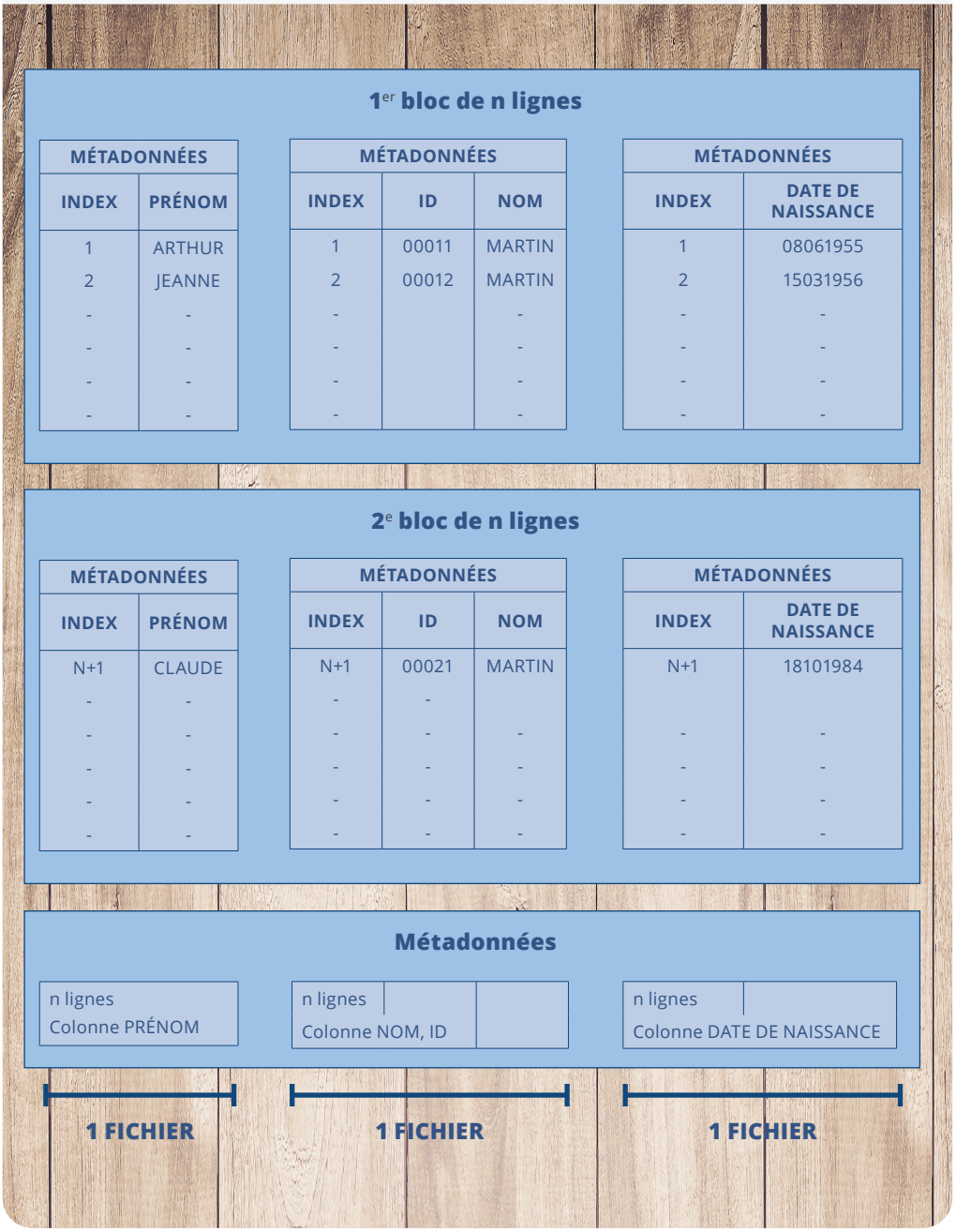
> ***Le format Parquet permet de solliciter de manière très naturelle la donnée de façon parallélisée.***

Parquet se base sur un principe algorithmique de stockage décrit et mis en œuvre par les équipes de Google dans leur processus de requêtage *(Melnik et al, 2010)*. Les données sont représentées sous format tabulaire, et les « lames » vont regrouper plusieurs colonnes et un groupe d'observations. Selon l'usage que l'on veut faire des données, il s'agira donc de définir des regroupements de colonnes ainsi qu'une stratification des observations la plus efficiente possible. Chaque « lame » contient également un ensemble de métadonnées décrivant les colonnes présentes, ainsi que la partie d'observations contenues dans cette « lame ». Les métadonnées vont être lues de manière indépendante des données *via* les « connecteurs » des solutions logicielles, de manière à permettre une navigation optimisée dans la table. Une des spécificités de cette conception de format est de permettre également de décomposer les données en plusieurs fichiers, de manière à rendre plus concrète et visible à l'utilisateur la notion de « lame » ; les connecteurs vont alors récupérer l'information associée aux métadonnées pour l'ensemble des fichiers Parquet contenus dans un même dossier, en considérant ces fichiers comme décrivant une seule et unique table.

Le format Parquet est donc particulièrement adapté pour gérer de la donnée volumineuse et distante[15], sans la dupliquer *(Mauvière, 2022)*. En s'imposant comme un standard *open source*, il donne la possibilité à l'utilisateur de travailler avec différents outils logiciels dont la complémentarité est précieuse. Il ne saurait néanmoins être vu comme une réponse unique à l'ensemble des problématiques du monde moderne de la donnée.

14 Système de calcul distribué développé en *open-source* et qui s'est imposé comme un outil largement utilisé pour traiter des données volumineuses.
15 C'est-à-dire stockée dans un endroit différent de celui où se réalise le calcul.

▶ Figure 2 - Le format Parquet



▶ Un format pour chaque étape du cycle de vie de la donnée

Quelles sont les questions à se poser pour choisir un format ? Tout d'abord, quel usage ? Le stockage des données et le formatage qu'il implique doit répondre à trois objectifs distincts : le traitement de la donnée, son analyse et son stockage pour un usage ultérieur. À chacune de ces étapes correspondent des usages très différents, et donc des besoins très spécifiques.

Ensuite, le volume est naturellement déterminant, mais les caractéristiques attendues du formatage vont également dépendre des utilisateurs et plus précisément de leur degré de maîtrise des outils de traitement de données. La micro-donnée, plus volumineuse et plus exigeante, sera plutôt réservée à un public averti, quand le grand public et de manière générale l'utilisateur *lambda* se référeront plus souvent à des tableaux agrégés.

> ***Le stockage des données et le formatage qu'il implique doit répondre à trois objectifs distincts : le traitement de la donnée, son analyse et son stockage pour un usage ultérieur.***

Enfin la localisation des données est un critère à prendre en compte pour la détermination du format le plus adapté : ce choix est en effet très dépendant de l'infrastructure informatique et des solutions logicielles disponibles pour traiter la donnée. En particulier, l'avènement de la virtualisation[16] du traitement est déterminant pour la définition des formats et de leurs aspects purement techniques. La virtualisation de la donnée amène à privilégier des formats offrant de bonnes performances en matière de lecture [17], comme le format Parquet, et incite à bien distinguer la donnée qui a vocation à être transformée de celle qui va être uniquement lue. À ces questions sont donc associés des critères relatifs au caractère temporaire ou durable de la donnée, ou dit autrement, à son degré de maturité.

Les différents formats sont déclinés selon l'usage des données et leur degré de maturité, en se plaçant dans différents contextes informatiques, comme décrit précédemment, et où l'utilisateur a la possibilité de virtualiser la chaîne de traitement qu'il applique aux données ***(voir Tableau)***. En particulier, dans ce contexte, la mise en place d'un SGBD peut se faire de manière transitoire, celui-ci étant construit de manière à être à « usage unique » pour procéder au traitement des données, puis être immédiatement supprimé une fois le traitement accompli, et la donnée en sortie stockée de manière plus permanente.

16 La notion de virtualisation décrit un principe selon lequel les données sont stockées sur des espaces physiques, bien souvent distants, pour lesquels l'utilisateur n'a pas besoin de connaître le détail de l'infrastructure de stockage, et peut les utiliser de façon transparente.
17 On retrouve ici la notion de WORM de la donnée.

Tableau : Propositions de formats associés à chaque usage de la statistique publique

	TRAITEMENT	DIFFUSION ET ANALYSE	ARCHIVAGE
DONNÉES AGRÉGÉES		XML, JSON	CSV, XML, JSON
MICRO-DONNÉES PEU VOLUMINEUSES	Format ouvert type CSV ou binaire type Parquet	Binaire type Parquet	CSV, XML, JSON
MICRO-DONNÉES VOLUMINEUSES, STOCKÉES LOCALEMENT	SGBD, binaire ad hoc lié à la solution logicielle ou Parquet	Binaire ad hoc lié à la solution logicielle ou Parquet	CSV
MICRO-DONNÉES VOLUMINEUSES ET VIRTUALISÉES	SGBD type Postgres ou binaire type Parquet	Binaire type Parquet	CSV

Dans son cycle de vie, la donnée doit initialement présenter une grande flexibilité car elle est sujette à de nombreuses modifications. Le format utilisé pour son stockage doit tenir compte de ces caractéristiques. Mais, plus la donnée gagne en maturité, plus il est souhaitable de recourir à des formats adaptés à des données figées, permettant une plus grande facilité de lecture de la donnée et optimisant son utilisation.

En guise de conclusion

Le choix du format est fonction de la finalité recherchée, du volume manipulé, de la localisation de la donnée. S'agissant de la production statistique, il est lié à la question de l'infrastructure et des solutions logicielles adoptées, qui appellent à tirer parti du potentiel offert par différents types d'infrastructure *(Comte et al., 2022)*. Les choix de stockage effectués par les producteurs ne sauraient ignorer ces nouveautés, visant à stocker la donnée de la manière la plus disponible possible, tout en donnant à l'utilisateur le maximum de liberté quant aux outils à mettre en œuvre pour l'utilisation souhaitée.

La notion de format adapté est appelée à évoluer dans le temps avec l'avènement de nouvelles solutions techniques et de nouveaux standards. Se pose également la question de la gestion du patrimoine des instituts statistiques. Le modèle de données a peu évolué au cours du temps, et la plupart de l'information produite et analysée par les statisticiens se présente sous format tabulaire. La conversion purement technique d'un format vers un autre est par essence relativement triviale. En revanche, les standards de description de la donnée et la notion de métadonnée posent avec une acuité nouvelle la question de la gestion de ce patrimoine. La façon d'y répondre — et en particulier avec une documentation la plus proche des standards actuels — aura un impact sur la capacité des chargés d'études à utiliser ce patrimoine de données ; cela constitue un fort enjeu pour la statistique publique lorsqu'il s'agit de décrire des phénomènes sur longue période.

Par ailleurs, les travaux récents relatifs à la question des *linked open-data (Zimmermann et al. 2011)* a renforcé l'importance des métadonnées, en particulier sur la question critique du référencement de la donnée et de son accessibilité. Ces travaux, expérimentaux à ce stade, vont nécessairement conditionner les standards, en particulier du point de vue de la diffusion, et de ce fait, avoir un impact sur les choix de format associés.

Choisir efficacement le format appelle aussi à réviser les principes de gouvernance de la donnée au sein du service statistique public, en décrivant de manière plus formalisée son cycle de vie et en cherchant à caractériser de façon normalisée et systématique la position de la donnée traitée au sein de celui-ci. De ce point de vue, les enjeux de parcimonie de stockage, de jeux de données de référence ou encore le principe de minimisation issu du Règlement Général à la Protection des Données conditionnent pour beaucoup la manière dont la donnée doit être stockée, traitée et gérée. Les formats les plus adaptés de ce point de vue sont ceux qui s'accommodent le mieux des solutions techniques propres à ces principes de gouvernance.

▶ Bibliographie

- ADEME, 2021. Prospective - Transitions 2050 - Rapport : Agir maintenant pour le climat. In : *Rapport de l'Ademe*. [en ligne]. [Consulté le 18 avril 2023]. Disponible à l'adresse : *https://librairie.ademe.fr/recherche-et-innovation/5072-prospective-transitions-2050-rapport.html*.

- 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. 68-85. [Consulté le 18 avril 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6035940?sommaire=6035950*.

- LAMARCHE, Pierre et LOLLIVIER, Stéfan, 2021. Fidéli, l'intégration des sources fiscales dans les données sociales. In : *Courrier des Statistiques*. [en ligne]. 8 juillet 2021. Insee, N° N6, pp. 28-46. [Consulté le 18 avril 2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/5398683?sommaire=5398695*.

- MAUVIÈRE, Éric, 2022. Parquet devrait remplacer le format CSV. Post de blog. [en ligne]. [Consulté le 18 avril 2023]. Disponible à l'adresse : *https://www.icem7.fr/cartographie/parquet-devrait-remplacer-le-format-csv/*.

- MELNIK, Sergey *et al.*, 2010. *Dremel : interactive analysis of web-scale datasets*. In : *Proceedings of the VLDB Endowment*, 3 (1-2), pp. 330-339. [en ligne]. Septembre 2010. [Consulté le 18 avril 2023]. Disponible à l'adresse : *https://dl.acm.org/doi/abs/10.14778/1920841.1920886*.

- NORDBOTTEN, Svein, 1996. *A statistical file system*. In : *site researchgate.net*. [en ligne]. Septembre 2010. [Consulté le 18 avril 2023]. Disponible à l'adresse : *https://www.researchgate.net/publication/283934373_A_Statistical_File_System.*

- SALOU, Gérard et SOSNOVSKY, Xavier, 2010. *SDMX as the logical foundation of the data and metadata model at the ECB*. [en ligne]. [Consulté le 18 avril 2023]. Disponible à l'adresse : *https://ideas.repec.org/h/bis/bisifc/33-09.html*.

- SUNDGREN, Bo, 2010. *Statistical databases – an introduction*. [en ligne]. 20 octobre 2010. [Consulté le 18 avril 2023]. Disponible à l'adresse : *https://www.diva-portal.org/smash/get/diva2:386476/FULLTEXT01.pdf*.

- UBER, Blog, 2022. *Cost Efficiency At Scale in Big Data File Format*. Post de blog. [en ligne]. 25 janvier 2022. [Consulté le 18 avril 2023]. Disponible à l'adresse : *https://www.uber.com/en-TZ/blog/cost-efficiency-big-data/*.

- WARNIER, Jean-Dominique, 1974. L'organisation des données d'un système, In : *Précis de logique informatique*. Édité par Honeywell Bull, CII Honeywell Bull. ISBN 978-2-7081-0229-3.

- ZIMMERMANN, Antoine., LOPES, Nuno, POLLERES, Axel, et STRACCIA, Umberto. (2011) *A general framework for representing, reasoning and querying with annotated semantic web data.* In : *Journal of Web Semantics, 11, 72-95*. [en ligne]. 7 mars 2011. [Consulté le 22 juin 2023]. Disponible à l'adresse : *https://arxiv.org/pdf/1103.1255*.