

Données — Structuration et traitement

2^{de}

Programme de seconde générale et technologique — BO spécial n°1 du 22 janvier 2019

Le nombre 37 ne dit rien. Une température, un âge, un nombre d'élèves, un prix ? Sans indication sur ce qu'il mesure, quand et comment, il n'autorise aucune conclusion. Tout ce chapitre consiste à ajouter à des nombres ce qui les rend utilisables — et à voir ce qu'ils continuent de révéler quand on croit les avoir anonymisés.

1 Structurer pour exploiter

DÉFINITION

Des **données** sont dites **structurées** lorsqu'elles sont organisées selon un schéma connu, qui permet à un programme de les lire sans interpréter. Le modèle le plus courant est la **table** : une ligne par **enregistrement**, une colonne par **descripteur**, et la même nature d'information dans toute une colonne. C'est cette régularité, et elle seule, qui rend un traitement automatique possible.

PROPRIÉTÉ

Deux formats de fichiers à connaître. Le **CSV** — *comma-separated values* — écrit une table en texte brut : une ligne par enregistrement, des valeurs séparées par un caractère, et une première ligne qui nomme les colonnes. Le **JSON** organise les **données** en couples nom-valeur, et accepte des structures **imbriquées** — une liste dans un champ, un objet dans un objet. Le **CSV** convient aux tables plates, le **JSON** aux données arborescentes.

CSV : une table plate · JSON : une structure imbriquée

DÉFINITION

Les **métadonnées** sont les **données sur les données** : ce que mesure une colonne, dans quelle unité, à quelle date, par quel instrument, avec quelle précision. Elles ne figurent pas dans les valeurs, et pourtant sans elles aucune valeur n'est interprétable. Un tableau sans **métadonnées** n'est pas un jeu de **données** : c'est une suite de nombres.

LE PIÈGE QUI COÛTE LE PLUS DE POINTS

Croire qu'un fichier partagé sans nom ni identifiant ne concerne plus personne. — Le croisement suffit à **réidentifier**. Un fichier sans nom mais portant la date de naissance, le code postal et le sexe désigne souvent une seule personne dans une commune donnée, et l'ajout d'une seconde source lève le dernier doute. Supprimer le nom ne rend donc pas des **données** anonymes : cela les rend seulement **moins immédiatement** nominatives.

Le réflexe : Se demander : avec **quelle autre source** ce fichier pourrait-il être croisé ?

2 Données personnelles et droit

DÉFINITION

Une **donnée** personnelle est une information se rapportant à une personne **identifiée ou identifiable**, directement ou **par recoupement**. Le nom en est une, mais aussi une adresse **IP**, un identifiant de téléphone, une photographie, un historique d'achats. Cette définition large est délibérée : elle vise l'**identifiabilité**, pas le seul état civil. Protéger la vie privée suppose donc de raisonner sur tout ce qui permet de remonter à une personne, et non sur le seul nom.

PROPRIÉTÉ

Le **RGPD** — règlement général sur la protection des données — s'applique en Europe depuis **2018**. Il pose quelques principes qu'un élève doit savoir nommer : la **finalité** — on ne collecte que pour un usage annoncé —, la **minimisation** — on ne collecte que le nécessaire —, la **durée limitée** de conservation, le **consentement** libre et éclairé, et des droits pour la personne concernée : accès, rectification, effacement, opposition. Le **RGPD** ne protège pas les **données** : il encadre ce que l'on a le droit d'en faire.

On parle de **big data** lorsque le volume, la vitesse d'arrivée et la variété des **données** dépassent ce qu'un traitement classique absorbe. Le changement n'est pas seulement d'échelle : avec beaucoup de **données**, on cherche des **corrélations** sans hypothèse préalable. D'où une vigilance nécessaire — une corrélation n'est pas une cause, et un très grand nombre de croisements en produit mécaniquement qui ne signifient rien.

3 Traiter

PROPRIÉTÉ

Quatre opérations élémentaires suffisent à exploiter une table : **rechercher** — retrouver les enregistrements vérifiant une condition ; **filtrer** — n'en garder qu'une partie ; **trier** — les ordonner selon une colonne ; **calculer** — produire une somme, une moyenne, un effectif. Elles se combinent, et l'ordre dans lequel on les applique change le coût du traitement, jamais le résultat.

EXPLOITER UN FICHIER DE DONNÉES STRUCTURÉES

- ① Lire d'abord la **ligne d'en-tête** : que contient chaque colonne, et dans quelle **unité** ?
- ② Chercher les **métadonnées** : origine, date, méthode de collecte, précision annoncée.
- ③ Repérer les valeurs **manquantes** ou **aberrantes** avant tout calcul.
Une moyenne calculée sur des cases vides comptées comme des zéros est fausse.
- ④ Appliquer les opérations : filtrer, trier, puis calculer sur ce qui reste.
- ⑤ Formuler la conclusion en rappelant ce que les **données** mesurent, et ce qu'elles ne mesurent pas.

À VÉRIFIER

- Ai-je vérifié ce que mesure chaque colonne, et dans quelle **unité** ?
- Ai-je cherché les **métadonnées** avant d'interpréter un chiffre ?
- Ai-je traité les valeurs manquantes autrement qu'en les comptant pour zéro ?
- Ai-je choisi le format adapté : **CSV** pour une table plate, **JSON** pour une structure imbriquée ?
- Mon fichier « anonyme » résiste-t-il à un **croisement** avec une autre source ?
- Ai-je distingué une **corrélation** d'une cause ?

À RETENIR

- Une **donnée** structurée suit un schéma connu : une ligne par enregistrement, une colonne par descripteur.
- **CSV** : table plate en texte brut. **JSON** : couples nom-valeur, structures imbriquées.
- Les **métadonnées** disent ce que mesure une donnée, quand et comment.
- Sans **métadonnées**, un tableau n'est qu'une suite de nombres.
- Est personnelle toute donnée se rapportant à une personne **identifiable**, y compris par recoupement.
- Supprimer le nom ne rend pas un fichier anonyme : le **croisement** réidentifie.
- **RGPD**, en vigueur depuis **2018** : finalité, minimisation, durée limitée, consentement, droits.
- Traiter, c'est rechercher, filtrer, trier, calculer.
- Traiter les cases vides comme des zéros fausse une moyenne.
- **big data** : volume, vitesse, variété — et une **corrélation** n'est pas une cause.
- La vie privée se protège en raisonnant sur l'**identifiabilité**, pas sur le seul nom.