



NIVEAU 3 SUR 3

Approfondi

Croiser, corriger, concevoir, argumenter : le chapitre sans aide.

Sans aucune aide · 60 minutes · sur 20

Données — Structuration et traitement

Corrigé du contrôle **niveau 3 · Approfondi**, avec le barème au point près.

❏ RAPPEL

La règle du cours que l'exercice met à l'épreuve, en tête de chaque exercice.

✓ RÉPONSE

Dans le cadre bordé : ce qui rapporte les points.

→ POURQUOI

Sous chaque réponse : la raison, pour comprendre et non recopier.

1

Croiser deux fichiers ☆ EXERCICE-PHARE

⌚ Croiser des sources · Repérer une erreur et la corriger

⌚ 16 min

6 pts

RAPPEL DE COURS

Le croisement suffit à réidentifier : quelques informations anodines, une date de naissance, une commune, le sexe, désignent souvent une seule personne, et une seconde source lève le dernier doute. Retirer le nom rend un fichier pseudonyme, pas anonyme, et la réglementation continue de s'y appliquer. La minimisation limite d'avance ce qui peut être croisé.

a Quelles informations les deux fichiers partagent-ils, directement ou presque ?**1 pt**

✓ **Le jour de naissance (« naissance » et « anniversaire »), la commune, et le sexe, qui se déduit de l'équipe féminine ou masculine ; le sport est commun aussi : le document 2 ne parle que de handball.**

→ **Pourquoi.** Un croisement s'appuie sur les colonnes communes, même quand elles ne portent pas le même nom : il faut lire les en-têtes et les légendes. 0,5 pour deux informations, 1 pour trois.

b Réalise le croisement. Quelles personnes du document 1 peux-tu nommer, et qu'apprends-tu sur elles ?**1,5 pt**

✓ **A3 est Clara Martin (née le 03/02, femme, Montbrac, handball) : elle a eu une fracture. A5 est Yanis Petit (né le 17/09, homme, Montbrac, handball) : il a eu une entorse.**

→ **Pourquoi.** Chaque combinaison jour de naissance, sexe, commune et sport ne désigne qu'une ligne du document 1 et qu'un joueur du document 2. 0,75 par personne réidentifiée avec sa blessure.

c A1 et A3 sont nées le même jour. Que montre cet exemple sur la manière dont on réidentifie ?**1 pt**

✓ **Que la date seule ne suffit pas : A1 et A3 sont deux handballeuses nées le 03/02, et c'est la commune qui les sépare. C'est la combinaison de plusieurs colonnes anodines qui rend une ligne unique.**

→ **Pourquoi.** Aucune colonne, isolée, ne désigne personne ; leur combinaison désigne une seule personne. C'est pourquoi on raisonne sur les croisements possibles, pas colonne par colonne.

d L'association soutient que le règlement européen ne la concerne pas, puisqu'elle a effacé l'identité de ses adhérents. Discute cet argument à la lumière de ta réponse précédente.**1 pt**

✓ **Elle est fausse : le fichier est pseudonyme, pas anonyme, puisque l'exercice vient de réidentifier deux personnes. Ces données se rapportent à des personnes identifiables par recoupement : ce sont des données personnelles, et le RGPD s'y applique.**

→ **Pourquoi.** Une donnée personnelle se définit par l'identifiabilité, pas par la présence d'un nom ; retirer le nom ne fait pas sortir un fichier de la réglementation.

- e On propose de supprimer seulement la colonne « naissance » du document 1. Le croisement devient-il impossible ? Justifie, puis propose une mesure plus sûre.

1,5 pt

✓ **Non : A3 reste la seule femme de Montbrac qui joue au handball, et Clara Martin la seule joueuse de Montbrac du document 2 ; A5 reste même le seul handballeur du fichier. Une mesure plus sûre : ne publier que des effectifs, par exemple le nombre de blessures de chaque type par sport, sans aucune ligne individuelle, ou ne collecter que les colonnes nécessaires à l'étude.**

→ **Pourquoi.** Supprimer une colonne ne suffit pas tant qu'une combinaison des autres reste unique ; publier des effectifs, ou appliquer la minimisation dès la collecte, supprime les lignes que l'on pourrait croiser. 0,75 pour la justification, 0,75 pour la mesure.

⚠ CE QUI COÛTE DES POINTS

Clara Martin associée à la ligne A1 : le croisement sur la seule date ne vaut aucun point à la question b.

Barème : 6 points — 1 + 1,5 + 1 + 1 + 1,5.

🔍 ERREURS FRÉQUENTES

- Croiser sur la seule date de naissance et attribuer à Clara Martin la ligne A1.
- Croire qu'il suffit de supprimer une colonne pour que plus personne ne soit reconnaissable.
- Répondre que le RGPD ne s'applique pas puisque le fichier ne contient aucun nom.

2

Trouve l'erreur

🕒 Repérer une erreur et la corriger

🕒 10 min

4 pts

RAPPEL DE COURS



Le CSV écrit une table plate avec un séparateur ; le JSON range des couples nom-valeur. Un nombre stocké comme texte fausse tout tri. On filtre ou l'on corrige en gardant le fichier d'origine intact. Une donnée seule ne dit rien : ce sont les métadonnées qui indiquent ce qu'elle mesure, et donc ce qu'on peut en conclure.

LA COPIE ANNOTÉE

- 1 Le fichier « lectures » est en JSON : une ligne par enregistrement, un point-virgule entre deux valeurs.
- 2 J'ai d'abord supprimé, dans le fichier d'origine, les lignes qui me semblaient fausses.
- 3 Après un tri décroissant de la colonne « livres », le plus grand lecteur en a lu 9, devant 15 et 12.
- 4 La colonne « temps » vaut en moyenne 37 : les élèves lisent donc 37 minutes par jour.

⚠ erreur 1

⚠ erreur 2

⚠ erreur 3

⚠ erreur 4

CRITÈRE	CE QUI EST ATTENDU	POINTS
Le format annoncé au début de la copie.	Ce qui est décrit, une ligne par enregistrement et des valeurs séparées par des points-virgules, c'est un fichier CSV. Un fichier JSON range des couples nom-valeur, éventuellement imbriqués. → La description est juste, mais le nom du format est faux : elle confond une table plate en texte brut et une structure arborescente.	1 pt
Le traitement des lignes jugées fausses.	On ne modifie jamais le fichier d'origine : on travaille sur une copie, ou l'on filtre, en gardant l'original intact. Et l'on ne supprime pas une ligne parce qu'elle « semble » fausse sans avoir dit pourquoi. → Garder le fichier d'origine intact permet de revenir aux données et de vérifier chaque choix ; une suppression non justifiée fausse les résultats.	1 pt
Le classement des lecteurs après le tri.	La colonne « livres » est rangée comme du texte : le tri compare caractère par caractère, et « 9 » passe devant « 15 » et « 12 ». Il faut convertir la colonne en nombres : le plus grand lecteur a lu 15 livres, devant 12 et 9. → Un nombre stocké comme texte fausse tout tri et tout calcul : c'est ce qu'on vérifie avant de traiter un fichier.	1 pt
La conclusion tirée de la moyenne de la colonne « temps ».	Rien n'indique l'unité ni la période de cette colonne : 37 peut désigner des minutes par semaine, des pages ou autre chose. Il faut lire les métadonnées avant de conclure, et conclure seulement sur ce que la colonne mesure. → Le nombre 37 ne dit rien sans indication sur ce qu'il mesure, quand et comment : ce sont les métadonnées qui autorisent une conclusion.	1 pt

⚠ CE QUI COÛTE DES POINTS

Une erreur repérée mais non corrigée : 0,5 sur 1.

Barème : 4 points — 1 par erreur : 0,5 pour la repérer, 0,5 pour la corriger.

🔍 ERREURS FRÉQUENTES

- Voir dans la ligne 3 une simple faute de calcul, sans reconnaître un tri sur du texte.
- Accepter la conclusion de la ligne 4 parce que 37 minutes par jour paraît vraisemblable.

3 Concevoir un fichier

🕒 Concevoir une structure

🕒 12 min

4 pts

RAPPEL DE COURS



Le CSV convient aux tables plates, une valeur par case ; le JSON accepte des structures imbriquées, une liste dans un champ, un objet dans un objet. Les métadonnées disent ce que contient le fichier, d'où il vient, quand et comment il a été établi.

a Quel format choisis-tu pour enregistrer ces menus tels quels ? Justifie à partir de la forme des données.

1 pt

- ✓ **Le JSON, parce que le nombre de desserts change d'un jour à l'autre et que chaque mets porte sa propre liste d'allergènes : des listes rangées dans des objets, que ce format accepte telles quelles.**

→ **Pourquoi.** Le choix du format découle de la forme des données : imbriquée, elle appelle le JSON ; plate, le CSV.

b Écris l'enregistrement d'un jour dans le format choisi, avec au moins deux desserts.

1,5 pt

- ✓ **Par exemple : `{"jour": "lundi", "entree": {"nom": "salade", "allergenes": []}, "plat": {"nom": "lasagnes", "allergenes": ["gluten", "lait"]}, "desserts": [{"nom": "yaourt", "allergenes": ["lait"]}, {"nom": "pomme", "allergenes": []}]}`**

→ **Pourquoi.** On attend des couples nom-valeur, un objet par mets et des listes pour les desserts et les allergènes. 0,5 pour les couples nom-valeur, 0,5 pour la liste des desserts, 0,5 pour les allergènes imbriqués.

- c La diététicienne veut compter, sur un mois, les plats qui contiennent du gluten. Propose l'en-tête d'une table plate qui le permet, et nomme les opérations qu'elle appliquera.

1 pt

✓ **Par exemple « jour;type;mets;gluten », une ligne par mets et par jour, la colonne gluten valant oui ou non. Elle filtrera sur « type = plat » et « gluten = oui », puis calculera l'effectif.**

→ **Pourquoi.** Une table plate se conçoit pour la question posée : une ligne par mets rend le filtre et le comptage immédiats, ce que la structure imbriquée ne permet pas aussi simplement. 0,5 pour l'en-tête, 0,5 pour les opérations.

- d Cite deux métadonnées à joindre au fichier pour qu'il soit exploitable.

0,5 pt

✓ **Par exemple : la période couverte et la date de mise à jour ; qui a établi les menus ; la liste des allergènes pris en compte (deux au choix).**

→ **Pourquoi.** Sans métadonnées, on ne sait ni de quand datent les menus, ni d'où ils viennent, ni ce que signifie une liste d'allergènes vide. 0,5 pour deux métadonnées pertinentes ; 0 pour une valeur du fichier citée comme métadonnée.

⚠ CE QUI COÛTE DES POINTS

Un format choisi sans regarder la forme des données : 0,5 sur 1.

Barème : 4 points — 1 + 1,5 + 1 + 0,5.

🔍 ERREURS FRÉQUENTES

- Choisir le CSV puis écrire les desserts et les allergènes dans une seule case.
- Citer comme métadonnée une valeur du fichier, un nom de plat par exemple.

RAPPEL DE COURS

Une donnée seule ne dit rien : publier un fichier utile suppose des métadonnées. Mais un fichier sans nom reste personnel tant que le croisement permet de réidentifier ; le RGPD impose alors finalité, minimisation, durée limitée et droits des personnes. Avec beaucoup de données et de sources, les croisements se multiplient, et les corrélations trouvées ne sont pas des causes.

CRITÈRE	CE QUI EST ATTENDU	POINTS
Introduction : l'enjeu posé, l'intérêt d'une étude face à la protection des élèves, et le plan annoncé.	Exemple : « Publier les notes d'un lycée pourrait éclairer la réussite scolaire, mais ces données concernent des élèves. Nous verrons ce qui rendrait ce fichier utile, puis ce qui le rend dangereux, avant de dire à quelles conditions le publier. » → L'introduction pose les deux exigences contradictoires que l'argumentation devra concilier, et annonce le chemin suivi.	1 pt
Premier paragraphe : ce qui rendrait le fichier utile, une structure claire et des métadonnées.	Une table plate au format CSV, une ligne par élève et par matière, un en-tête explicite ; des métadonnées qui disent ce que mesure chaque note, sur combien, à quelle période. Sans elles, les nombres ne permettent aucune conclusion, et une corrélation trouvée ne prouverait aucune cause. → L'utilité d'un jeu de données tient à sa structure et à ses métadonnées : un tableau qui n'en a pas n'est qu'une suite de nombres.	1,5 pt
Second paragraphe : le risque de réidentification et ce qu'impose le RGPD.	Sans nom, mais avec la classe, l'option, le sexe ou la date de naissance, un élève est souvent seul de sa combinaison ; un trombinoscope, une liste d'admis ou un réseau social lèvent le dernier doute, et plus les sources sont nombreuses, plus ces croisements sont faciles. Le fichier est pseudonyme : le RGPD s'applique, avec la finalité, la minimisation, la durée limitée et les droits des élèves. → Un fichier sans nom reste personnel tant que le croisement permet de réidentifier ; la variété des sources disponibles rend ce croisement plus probable.	1,5 pt
Conclusion : une réponse nette, avec des conditions précises.	Exemple : « Ce fichier ne doit pas être publié ligne par ligne. On peut publier des effectifs et des moyennes par niveau, avec leurs métadonnées, pour une finalité annoncée, sans aucune colonne qui permette de croiser. » → La conclusion répond à la question posée et transforme les arguments en conditions applicables.	1 pt
Expression : le vocabulaire du chapitre employé juste, des connecteurs logiques, des paragraphes reliés.	Par exemple : métadonnées, pseudonyme, croisement, minimisation, finalité ; cependant, c'est pourquoi, donc. → Le vocabulaire précis et les connecteurs rendent visible le raisonnement : ce qui s'oppose, ce qui découle de quoi.	1 pt

UN EXEMPLE D'ARGUMENTATION

Publier en ligne les notes d'un lycée pourrait aider à comprendre la réussite scolaire : c'est l'intérêt des données ouvertes. Mais ces notes concernent des élèves, et un fichier publié échappe à son auteur. Faut-il le publier, et à quelles conditions ? Nous verrons d'abord ce qui rendrait ce fichier utile, puis ce qui le rend dangereux.

Pour être utile, le fichier doit être structuré : une table plate au format CSV, une ligne par élève et par matière, un en-tête qui nomme chaque colonne. Il lui faut surtout des métadonnées. Une note de 12 ne dit rien si l'on ignore sur combien elle est, à quelle période elle a été donnée, et s'il s'agit d'un devoir ou d'une moyenne. Même ainsi, une étude qui trouverait un lien entre deux colonnes n'aurait trouvé qu'une corrélation, pas une cause.

Cependant, retirer les noms ne suffit pas. Avec la classe, l'option, le sexe et la date de naissance, un élève est souvent le seul de sa combinaison ; un trombinoscope ou un profil public lève alors le dernier doute, et plus les sources disponibles sont nombreuses, plus ce croisement est facile. Le fichier est donc pseudonyme, et non anonyme : ce sont des données personnelles, et le RGPD s'applique. Il impose une finalité annoncée, la minimisation des colonnes, une durée limitée et le respect des droits des élèves, dont celui de s'opposer.

C'est pourquoi le fichier détaillé ne doit pas être publié ligne par ligne. Le lycée peut publier des effectifs et des moyennes par niveau, accompagnés de leurs métadonnées, pour une étude dont la finalité est annoncée, sans aucune colonne qui permette de croiser. L'étude garde ainsi son intérêt, et aucun élève ne peut être reconnu.

⚠ CE QUI COÛTE DES POINTS

Une réponse sans condition précise : 0,5 sur 1 pour la conclusion.

Barème : 6 points — 1 + 1,5 + 1,5 + 1 + 1.

Q ERREURS FRÉQUENTES

- Répondre oui ou non sans peser l'intérêt de l'étude et le risque pour les élèves.
- Affirmer qu'un fichier sans nom ne pose aucun problème.
- Énumérer des principes du RGPD sans les appliquer au fichier des notes.



Barème de notation

Ce que le correcteur regarde, exercice par exercice.

20 pts

CE QUI EST ATTENDU	EXERCICE	POINTS	CE QUI FAIT PERDRE LE POINT
Le croisement réalisé et discuté	1	6 points — 1 + 1,5 + 1 + 1 + 1,5	Clara Martin associée à la ligne A1 : le croisement sur la seule date ne vaut aucun point à la question b.
Les erreurs repérées et corrigées	2	4 points — 1 par erreur : 0,5 pour la repérer, 0,5 pour la corriger	Une erreur repérée mais non corrigée : 0,5 sur 1.
La structure conçue et justifiée	3	4 points — 1 + 1,5 + 1 + 0,5	Un format choisi sans regarder la forme des données : 0,5 sur 1.
L'argumentation construite	4	6 points — 1 + 1,5 + 1,5 + 1 + 1	Une réponse sans condition précise : 0,5 sur 1 pour la conclusion.
Total		20 points	



Que faire de ta note ?

Le contrôle ne s'arrête pas au chiffre : voici la suite selon ton score.

 Niveau 1 · Essentiel	 Niveau 2 · Standard	 Niveau 3 · Approfondi — tu es ici
---------------------------------	--------------------------------	--



16 à 20

NIVEAU MAÎTRISÉ

C'est le niveau le plus élevé, et tu le maîtrises. Pour aller plus loin, applique le réflexe du chapitre à un fichier réel que tu utilises : avec quelle autre source pourrait-il être croisé ?



10 à 15

REFAIRE CE NIVEAU

Relis la correction de l'exercice où tu as perdu le plus de points. Refais le croisement de l'exercice 1 sans regarder le corrigé, réécris ton argumentation en suivant son plan, puis refais le niveau 3 dans une semaine.



moins de 10

NIVEAU INFÉRIEUR

Repasse par le niveau 2 — Standard : il entraîne la vérification d'un fichier, les requêtes combinées et l'analyse d'un formulaire, dont ce niveau a besoin.

✓ COMPARE AVEC TA PRÉVISION

- Reprends la grille « Avant de rendre ta copie » : là où tu avais coché « acquis » et perdu des points, relis le rappel de cours de l'exercice.
- Refaire un contrôle quelques jours plus tard, sans regarder le corrigé, mesure ce que tu as retenu, et non ce dont tu te souviens.
- La fiche de cours du chapitre reprend chaque notion : ProgresSchool · SNT Seconde · Données — Structuration et traitement.