– IMPACT Initiatives –

Normes minimales et standard de vérifications pour le nettoyage des données

Date	13/01/2020
From	Unité de Recherche & de la gestion des données du siège
To	Aux équipes pays d'IMPACT
Subject	Établissement de normes minimales pour le nettoyage et le traitement des données à IMPACT

1. Introduction

Cette note de service vise à établir les normes **minimales que toutes les équipes IMPACT (sur le terrain et au siège) sont censées respecter dans la phase de nettoyage et de traitement des données du cycle de recherche. Ces normes doivent être respectées pour qu'un ensemble de données soient considérées comme validées et utilisables pour l'analyse**. Si, pour une raison quelconque, ces normes sont impossibles à respecter pour un cycle de recherche particulier, veuillez le signaler à l'Unité de Recherche & de la Gestion des Données du siège (RDD) dès que possible, afin qu'une solution soit trouvée.

Les normes minimales et standard de vérifications visent à s'aligner sur A) les SoPs relatives à la protection des données d'IMPACT et B) les prochaines directives de nettoyage des données IMPACT.

Les normes minimales et standard de vérifications comprennent (1) les actions spécifiques qui doivent être faites pendant le traitement des données en plus de (2) de la documentation clé qui doit être partagée avec toutes les données soumises pour validation du siège. Le but de la documentation clé est de permettre à l'unité RDD d'examiner le processus de nettoyage des données au complet et de s'assurer que les normes minimales ont été respectées.

2. Documentation clé nécessaire pour la validation des données par le siège

Pour permettre à l'unité RDD de vérifier que les normes minimales et standard ont été respectées pendant le nettoyage des données, **les éléments suivants de la documentation doivent toujours être partagés avec tout ensemble de données soumis à l'examen et à la validation du siège :**

1. Ensemble de données brutes
2. Base de données nettoyée
3. Fiche de nettoyage / documentation sur la procédure de nettoyage et les types de contrôles effectués ; par exemple :
    a. Une feuille de " nettoyage " supplémentaire dans la base de données Excel où les colonnes de " contrôle " sont incluses
    b. Le script/code R utilisé pour nettoyer et traiter les données
4. Journaux de nettoyage et de suppression (à l'aide du modèle de journal de nettoyage IMPACT)
5. Questionnaire KOBO/ ODK
    a. Pour toutes les évaluations utilisant KOBO/ODK pour la collecte de données, le " type de méta-question d'enregistrement d'audit " devrait être inclus lors de la conception de l'outil (plus d'informations ici) ; si possible, les données de celui-ci devraient être utilisées pour surveiller le comportement des recenseurs (voir la catégorie " Métadonnées des recenseurs " dans le tableau 1 ci-dessous)
6. Rapport de suppression des données
7. Résultats de la vérification de l'échantillonnage (voir le tableau 1 ci-dessous pour plus de détails)
8. [Pour les données qualitatives] Quelques exemples de transcriptions brutes et/ou de formulaires de débriefing utilisés pour traiter et analyser les données qualitatives.

3. Comment établir et garantir des normes minimales ?

Voir le tableau 1 ci-dessous.

Catégorie	Type de contrôle et point(s) d'action pertinent(s)	Résultats à soumettre à siège	Quand cette vérification doit-elle être faite	
			Pendant la collecte des données	Après la collecte des données
Métadonnée du questionnaire	Tous les enregistrements ont des **IDs uniques** ou des UUID • Action requise : Tout doublon devrait être supprimé et enregistré dans le journal de nettoyage des données.	Journal de suppression selon le modèle de journal de nettoyage IMPACT.		
Protection des données	Toutes les **informations qui peuvent être utilisés pour identifier des personnes ou ménages doivent être retirées** de l'ensemble de données. Exemple d'information : Coordonnées GPS ; Noms ; Numéros de téléphone ; profession du répondant / organisation ; Information sur les recenseurs / les informateurs clés ; le sexe, l'âge et le lieu du répondant ; etc. • Action requise : Supprimez ou extrayez soigneusement toutes les informations personnelles identifiables. Toutes les mesures prises à l'égard des renseignements personnels identifiables doivent être conformes à : (1) la politique de Protection des données IMPACT SoPs pour les renseignements personnels identifiables et (2) Indicateur de risque matrice complétée dans les TdR (Annexe du Plan de Gestion des Données).	Rapport de suppression des données selon le modèle d'IMPACT SoPs sur la protection des données pour les informations personnellement identifiables.		
Métadonnée du questionnaire	Les données finales sont **conformes à la stratégie d'échantillonnage prévue**, c.-à-d : (1) les lieux/points d'entrevue et les lieux/points d'échantillonnage prévus sont cohérents, à moins qu'il n'y ait une justification claire (et que les limites de ce dernier soient bien comprises) ; (2) le nombre d'enregistrements par strate correspond aux cibles prévues par strate (moins le tampon si celui-ci a été ajouté à la cible pour atténuer les non-réponses) ; (3) pour l'échantillonnage en grappes stratifié, des entrevues avec des groupes de population (strates) pour une grappe qui a été échantillonnée en fonction du PPS au sein des groupes assignés ;	Un résultat clair (carte, tableau ou résumé écrit) décrivant les résultats de ces vérifications.		

	(4) il y a des variables dont les valeurs correspondent exactement aux noms des strates dans la base de sondage (le cas échéant) ; (5) il y a une variable dont les valeurs correspondent exactement aux noms des grappes dans la base de sondage (s'il y a lieu). Action requise si l'un des problèmes mentionné ci-dessus est identifié : • Tout écart observé doit être vérifié et compris. • Un suivi régulier pendant la collecte des données devrait être effectué pour comparer l'échantillon prélevé à l'échantillon prévu à l'origine. Cela pourrait être fait soit par : (1) en préparant une carte générale superposant les lieux d'échantillonnage prévus avec les lieux où les données ont été récoltées et (2) la tenue d'une feuille de calcul de suivi comparant des cibles par emplacement et par strate.			
Métadonnées des Enumérateurs	Vitesse **d'entrevue de l'énumérateur** (c.-à-d. le temps pris pour l'entrevue ou l'enquête) est raisonnable. Pour la plupart des évaluations mises en œuvre par IMPACT, un temps <10 minutes devrait être une indication raisonnable. Pour un examen plus approfondi, (p. ex. MSNA), le point de référence devrait être plus élevé (<20 minutes). En fin de compte, le repère devrait être fondé sur quel est le temps minimum nécessaire pour compléter ce questionnaire spécifique, et il devrait être établi par l'équipe d'évaluation au cours de la phase de Journal de suppression selon le modèle de journal de nettoyage IMPACT Fiche de nettoyage / documentation de la procédure de nettoyage et les types de contrôles effectués (par exemple, une feuille supplémentaire dans la Base de données Excel où les colonnes "check" sont incluses ou le script/code R utilisé à des fins de nettoyage 4) la conception et la mise à l'essai, y compris la prise en compte de plusieurs types d'enquêtes (p. ex. grand/petit ménage, avec et sans MUAC, logique de remplissage des réponses différente).	Journal de suppression selon le modèle du journal de nettoyage IMPACT Fiche de nettoyage / documentation de la procédure de nettoyage et les types de contrôles effectués (par exemple, une feuille supplémentaire dans la Base de données Excel où les colonnes "check" sont incluses ou le script/code R utilisé pour le nettoyage).		

	• Action nécessaire : Si le délai est plus court que prévu, un suivi devrait être effectué pour confirmer si cela est possible.			
Métadonnées des Enumérateurs	Aucun des énumérateurs ne suit systématiquement l**e plus court ou exactement le même cheminement,** c'est-à-dire fournir la même réponse pour plusieurs enregistrements d'informations. Par exemple, nous avons remarqué qu'il y a un énumérateur (identifié par variable d'identification des énumérateurs) qui a tendance à entrer exactement les mêmes réponses entre plusieurs informateurs clés. Cela semble un peu suspect et pourrait être une indication de falsification de données. L'agent d'évaluation du siège, ou l'équipe d'examen n'a pas toujours les connaissances contextuelles nécessaires pour juger qu'il s'agisse effectivement de questions ou qu'il soit logique que tous les règlements dans la zone de dénombrement ont exactement la même situation. Il est donc important de faire un suivi auprès des énumérateurs pour clarifier la situation. • Action requise : Une justification claire devrait être identifiée pour permettre de démontrer que les entrevues et les données ne sont pas falsifiées	Journal de nettoyage selon le modèle du journal de nettoyage IMPACT Fiche de nettoyage / documentation de la procédure de nettoyage et les types de contrôles effectués (par exemple, une feuille supplémentaire dans la Base de données Excel où les colonnes "check" sont incluses ou le script/code R est utilisé pour le nettoyage).		
Vérifications logiques	Il n'y a pas **d'observations aberrantes inexplicables ou impossibles**, c'est-à-dire une observation / un point de données spécifique qui se trouve à une distance anormale des autres valeurs dans le jeu de données. Par exemple, si nous connaissons le revenu moyen dans une zone spécifique qui est d'environ 500 USD/mois, si un ménage déclare un revenu de 100 000 USD, cela pourrait être le résultat d'une erreur de saisie. • Action requise : Toutes les valeurs sortant de la norme devraient être identifiées, étudiées et corrigées comme il convient. • Action requise : Il est également important que les valeurs aberrantes identifiées ne soient pas automatiquement supposées être incorrectes et supprimées sans suivi. Dans l'exemple fourni ci-dessus, des niveaux de revenu aussi élevés pourraient être possible par exemple si la taille du ménage est supérieure à la moyenne de la zone. En d'autres termes, ce que nous considérons parfois comme une valeur aberrante n'en est pas nécessairement une.	Journal de nettoyage selon le modèle du journal de nettoyage IMPACT Fiche de nettoyage / documentation de la procédure de nettoyage et les types de contrôles effectués (par exemple, une feuille supplémentaire dans la Base de données Excel où les colonnes "check" sont incluses ou le script/code R est utilisé pour le nettoyage).		

Vérifications logiques	Il y a une **cohérence logique entre les différentes réponses** au sein d'un dossier. Lors du nettoyage quotidien des données au niveau pays, les types d'incohérences à surveiller et les mesures à prendre si une telle incohérence est identifiée devrait être claire pour tous ceux qui travaillent sur la procédure de nettoyage. Pendant l'examen du siège, il n'est pas possible d'identifier et de vérifier toutes les incohérences logiques sur chaque ensemble de données. Cependant, les incohérences potentielles les plus évidentes sont vérifiées et signalées. Par exemple : Tous les IC disent que (1) plusieurs personnes n'ont pas pu avoir accès à la nourriture au cours du dernier mois et (2) que plusieurs personnes n'avaient pas accès à leurs activités de subsistance habituelles. Cependant, ils disent aussi (3) " La sensation de faim est faible, les stratégies sont disponibles pour y faire face " et (4) lorsqu'il n'y avait pas assez de nourriture, la stratégie que les gens utilisaient était d'"emprunter de la nourriture aux autres". Le lien entre ces réponses d'un même IC pour un même village peut être illogique et donc doit donc être vérifié. • Action requise : Les incohérences entre les questions devraient être Identifiées, enquêtées et corrigées, le cas échéant. • Action requise : Les questions de suivi devraient être vérifiées à deux reprises pour la cohérence avec les questions de haut niveau (p. ex. les niveaux d'accès signalés à la nourriture et l'utilisation de stratégies pour faire face au manque de nourriture). • Action requise : Vérifier que dans chaque variable, toutes les données ont la même unité (par exemple, le nombre de jours ou la devise en dollars américains) dans toutes les rangées.	Journal de nettoyage selon le modèle du journal de nettoyage IMPACT. Fiche de nettoyage / documentation de la procédure de nettoyage et les types de contrôles effectués (par exemple, une feuille supplémentaire dans la Base de données Excel où les colonnes "check" sont incluses ou le script/code R est utilisé pour le nettoyage).		
Journal de nettoyage	Un **registre de nettoyage clair et complet est tenu conformément au** modèle du journal de nettoyage IMPACT. Tous les différents types de contrôles de données effectués et la ou les mesures de suivi prises doivent apparaître clairement à l'examen du journal. • Action requise : Ajouter exactement une ligne pour chaque entrée de données individuelle qui a été signalée lors des contrôles quotidiens des données.	Journal de nettoyage selon le modèle du journal de nettoyage IMPACT. Dans le cas peu probable où aucun problème n'a été identifié au cours du processus de nettoyage		

	• Action requise : Effectuer un contrôle final pour vérifier que le nombre d'entrées du journal de nettoyages (= le nombre de contrôles) est raisonnable compte tenu du type du questionnaire et du contexte de la collecte des données. ○ Pour la gestion directe de la collecte des données, il est suggéré une référence de 5 % du total des documents pour lesquels les émissions ont été identifiées, suivies ○ Pour la gestion à distance de la collecte des données, il est suggéré une référence de 10 % du total des documents pour lesquels les émissions ont été identifiés, suivies	des données, une note doit encore être laissée dans le journal de nettoyage expliquant les contrôles effectués et leur raison d'être, ainsi que le fait que des problèmes n'aient pas été identifiés malgré ces contrôles.		
Journal de nettoyage	Un **journal de suppression clair et complet est tenu à jour** conformément au modèle du journal de nettoyage IMPACT. • Action requise : Ajouter exactement une ligne pour chaque enregistrement d'enquête supprimé ; la justification de la suppression doit apparaître clairement dans le journal, sur la base de normes minimales établies dès le départ de l'évaluation pour déterminer quand les documents doivent être supprimées dans leur intégralité. • Action requise : À la fin de la collecte des données, si un pourcentage élevé des enquêtes (>10%) doit être supprimé en raison de la qualité des données, de l'absence de consentement ou d'erreurs de l'enquêteur, tout biais introduit en conséquence doit être clairement signalé lors de la présentation des résultats.	Journal de suppression selon le modèle de journal de nettoyage IMPACT.		
Formatage des données	L'ensemble de données est présenté dans un **format propre, ordonné et utilisable à des fins d'analyse.** • Action requise : Les "autres" réponses ont été recodées dans les catégories existantes ou nouvelles catégories selon les cas. • Action requise : Les champs de données manquants sont laissés vides ou remplacés par NA là où c'est nécessaire.	Ensemble de données final et nettoyé à utiliser pour l'analyse.		

	• Action requise : Dans chaque variable, il est vérifié que toutes les données ont la même unité (par exemple, le nombre de jours ou la devise en dollars US) dans toutes les lignes. • Action requise : Pour les variables numériques, si pour la collecte de données, d'autres ont été introduits (c'est-à-dire .999 - non recommandé), elles sont remplacées par un blanc ou un NA dans l'ensemble de données nettoyées finales. • Action requise : Le cas échéant, la variable dans l'ensemble de données qui devrait être utilisée pour le calcul et l'application des pondérations devrait être clairement indiquée.			
Formatage des données	L'ensemble de données est présenté dans un format **propre**, **ordonné et utilisable** par toute personne ne connaissant pas la recherche, avec une feuille LISEZ-MOI claire. • Action requise : Toutes les étapes décrites dans les [à venir] Lignes Directrices sur le Nettoyage des Données pour la Publication des Ensembles de Données doivent être prises en compte.	Publication d'un ensemble de données final et nettoyé.		