Module 7 : Activités post-collecte
Vidéo 1 : Introduction + encodage des réponses ouvertes et préparation des données

Sharan Sharma

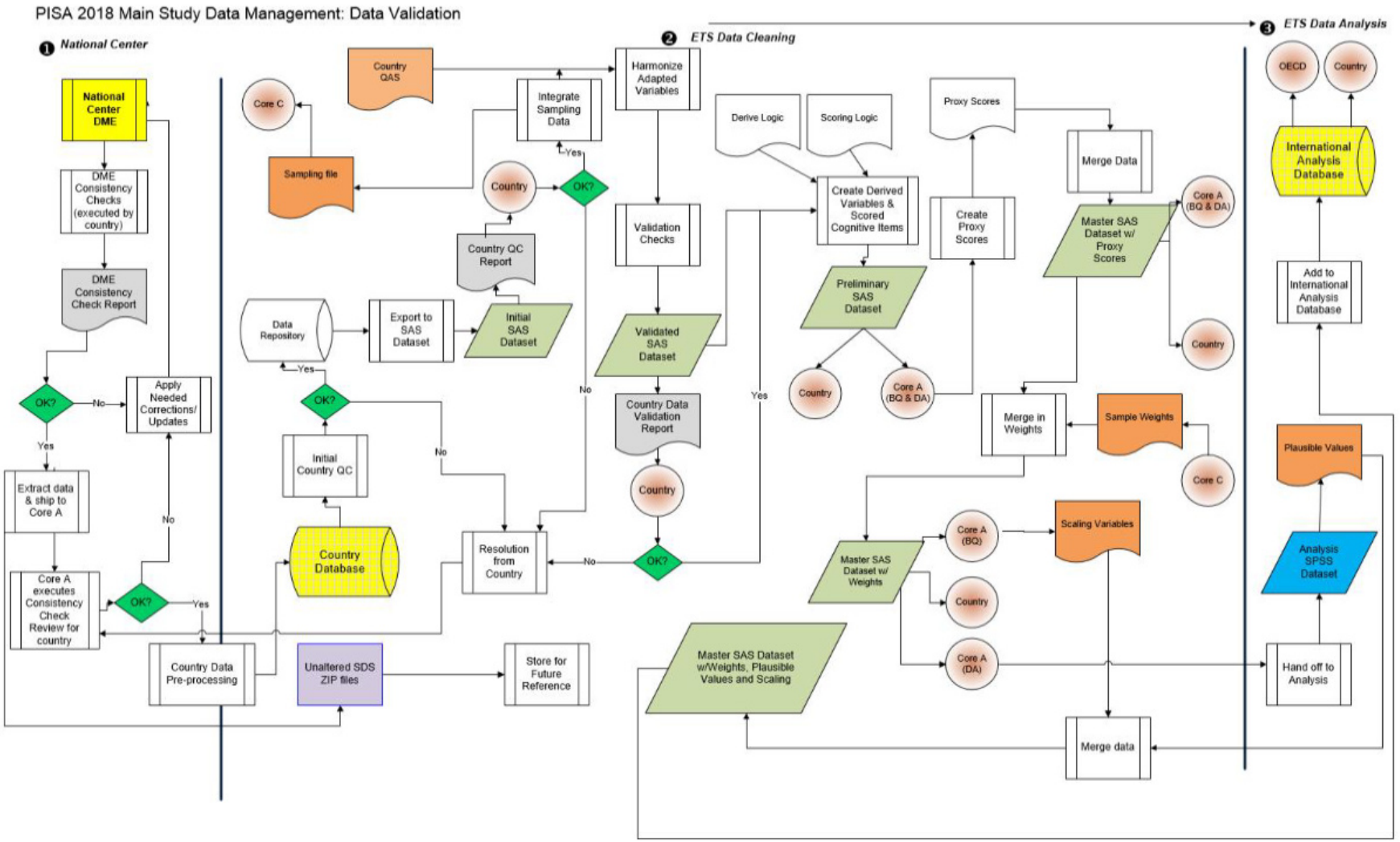
MANNHEIM
BUSINESS SCHOOL

À la fin du module, les participants devraient...

...s’être familiarisés avec les grandes étapes post-collecte

Vous avez collecté les données. Il n’y a plus qu’à les publier ?

Pas si vite !



Source : https://www.oecd.org/pisa/data/pisa2018technicalreport/PISA2018%20TecReport-Ch-10-Data-Management.pdf

Les activités sont nombreuses...

1) Encodage des réponses ouvertes

2) Préparation des données

3) Synthèse et visualisation

4) Édition des données

5) Imputation et pondération (voir module 2)

6) Contrôle de la divulgation

7) Traitement final, documentation et diffusion

Encodage des questions ouvertes

- Au préalable : sauvegarde des données brutes originales.

- À ce stade, ciblage sur les réponses aux questions du type : « Quel type de travail faites-vous ? » : les chaînes de réponse doivent être encodées en code de profession.

- Deux méthodes d'encodage :
  - Entièrement manuelle
  - Automatisée

Encodage manuel

- Les encodeurs comparent la réponse à une grille d'encodage et la classent selon une catégorie prédéfinie.
  - Pour les questions nouvelles ou non standard, il peut être nécessaire d'adapter les catégories : c'est le problème du « trop de réponses "Autre" ».

- Il est conseillé de faire un encodage de vérification pour une partie des cas afin d'avoir une estimation de la fiabilité.

Encodage manuel...

- La fiabilité inter-encodeurs mesurée par
  - Une simple comparaison de pourcentages
  - Des mesures plus complexes, comme le kappa de Cohen (k) sont largement utilisées: elles tiennent compte de l'accord aléatoire.
  - κ inférieur à un certain seuil, 70 % par ex., mérite un réexamen

- Le fait que les réponses longues sont toujours encodées de manière plus fiable relève du mythe [Belloni et al (2016), Conrad et al (2016)].
  - Plus de texte est parfois source de confusion.

Encodage manuel...

- Une pratique rare mais très utile : discuter avec les encodeurs des cas de désaccord. Les encodeurs ont souvent recours à des règles informelles (Conrad et al., 2016) et ces discussions peuvent permettre d'améliorer la formalisation.

- Incitez les encodeurs a être plus explicites sur les incertitudes (prévoyez une possibilité de code secondaire, par ex.)
  - La nomenclature peut être utile, comme la Classification internationale type des professions (ISCO-88) qui présente une structure hiérarchisée avec 10 grands groupes, 28 sous-grands groupes, 116 sous-groupes et 390 groupes de base.
  - En cas d'informations lacunaires aux niveaux inférieurs, des codes de niveau supérieur sont employés.

I. Grands groupes

1 Managers

2 Professionals

3 Technicians and Associate Professionals

4 Clerical Support Workers

5 Services and Sales Workers

6 Skilled Agricultural, Forestry and Fishery Workers

7 Craft and Related Trades Workers

8 Plant and Machine Operators and Assemblers

9 Elementary Occupations

0 Armed Forces Occupations

« Directeurs et gérants, hôtellerie » = 1411

« Directeurs, cadres de direction et gérants » = 1000

II. Grands groupes et sous-grands groupes

1 Managers
- 11 Chief Executives, Senior Officials and Legislators
- 12 Administrative and Commercial Managers
- 13 Production and Specialized Services Managers
- 14 Hospitality, Retail and Other Services Managers

III. Grands groupes, sous-grands groupes et sous-groupes

14 Hospitality, Retail and Other Services Managers
- 141 Hotel and Restaurant Managers
- 142 Retail and Wholesale Trade Managers
- 143 Other Services Managers

IV. Grands groupes, sous-grands groupes, sous-groupes et groupes de base

14 Hospitality, Retail and Other Services Managers
- 141 Hotel and Restaurant Managers
  - 1411 Hotel Managers
  - 1412 Restaurant Managers
- 142 Retail and Wholesale Trade Managers
  - 1420 Retail and Wholesale Trade Managers
- 143 Other Services Managers
  - 1431 Sports, Recreation and Cultural Centre Managers
  - 1439 Services Managers Not Elsewhere Classified

Source : https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_172572.pdf

Méthodes automatisées

- Comme le montre l'exemple ISCO, l'encodage peut être compliqué, et c'est une activité coûteuse et chronophage (en particulier pour les enquêtes de grande ampleur).

- Les méthodes automatisées sont de plus en plus populaires, comme le système de codage informatisé du NIOSH (NIOCCS) https://wwwn.cdc.gov/nioccs3/

- Il peut s'agir d'une simple table à consulter basée sur des données historiques, ou de méthodes statistiques plus sophistiquées.

CASCOT : Computer Assisted Structured Coding Tool

- Warwick Institute for Employment Research [https://warwick.ac.uk/fac/soc/ier/software/cascot/details/]

- Attribue une note de certitude. En général, l'encodage automatisé est accepté si la note est supérieure à un certain seuil, 70 par ex. (Belloni et al, 2016)

- Le logiciel propose aussi d'autres codes, qui aident les encodeurs manuels à prendre une décision pour les notes inférieures au seuil.

- CASCOT a été comparé aux données encodées manuellement de haute qualité : 80 % des données saisies ont obtenu une note > 40, et parmi ces dernières, 80 % correspondaient aux données encodées manuellement.

Les activités sont nombreuses...

1) Encodage des réponses ouvertes

2) Préparation des données

3) Synthèse et visualisation

4) Édition des données

5) Imputation et pondération

6) Contrôle de la divulgation

7) Traitement final, documentation et diffusion

Préparation des données

- Convertissez les formats larges (horizontaux) en formats longs (hiérarchisés).
  - Permet de réduire les colonnes vides ; plus compact
  - Facilite le travail des analystes

household ID	ID member 1	ID member 2	ID member 3	age member 1	age member 2	age member 3	relationship with hh member 1	relationship with hh member 2	relationship with hh member 3
1	1	2	3	30	28	10	Head of household	Spouse of head	Unarried child
2	1	2	3	28	62	68	Head of household	Father	Mother
3	1	2	3	40	25	23	Head of household	Spouse of head	Married child
4	1	2	3	39	80	82	Head of household	Grandfather	Grandmother
5	1	2	3	55	31	5	Head of household	Daughter- in- law	Grandchild

Format large (horizontal)

household ID	ID members	age	relationship with head of household
1	1	30	Head of household
1	2	28	Spouse of head
1	3	10	Unarried child
2	1	28	Head of household
2	2	62	Father
2	3	68	Mother
3	1	40	Head of household
3	2	25	Spouse of head
3	3	23	Married child
4	1	39	Head of household
4	2	80	Grandfather
4	3	82	Grandmother
5	1	55	Head of household
5	2	31	Daughter- in- law
5	3	5	Grandchild

Format long (hiérarchisé)

- Chaque saisie (ligne) se trouve désormais à un niveau individuel unique.
- Les colonnes sont les variables.

Source : https://guide-for-data-archivists.readthedocs.io/en/latest/prepData.html

Variable ID

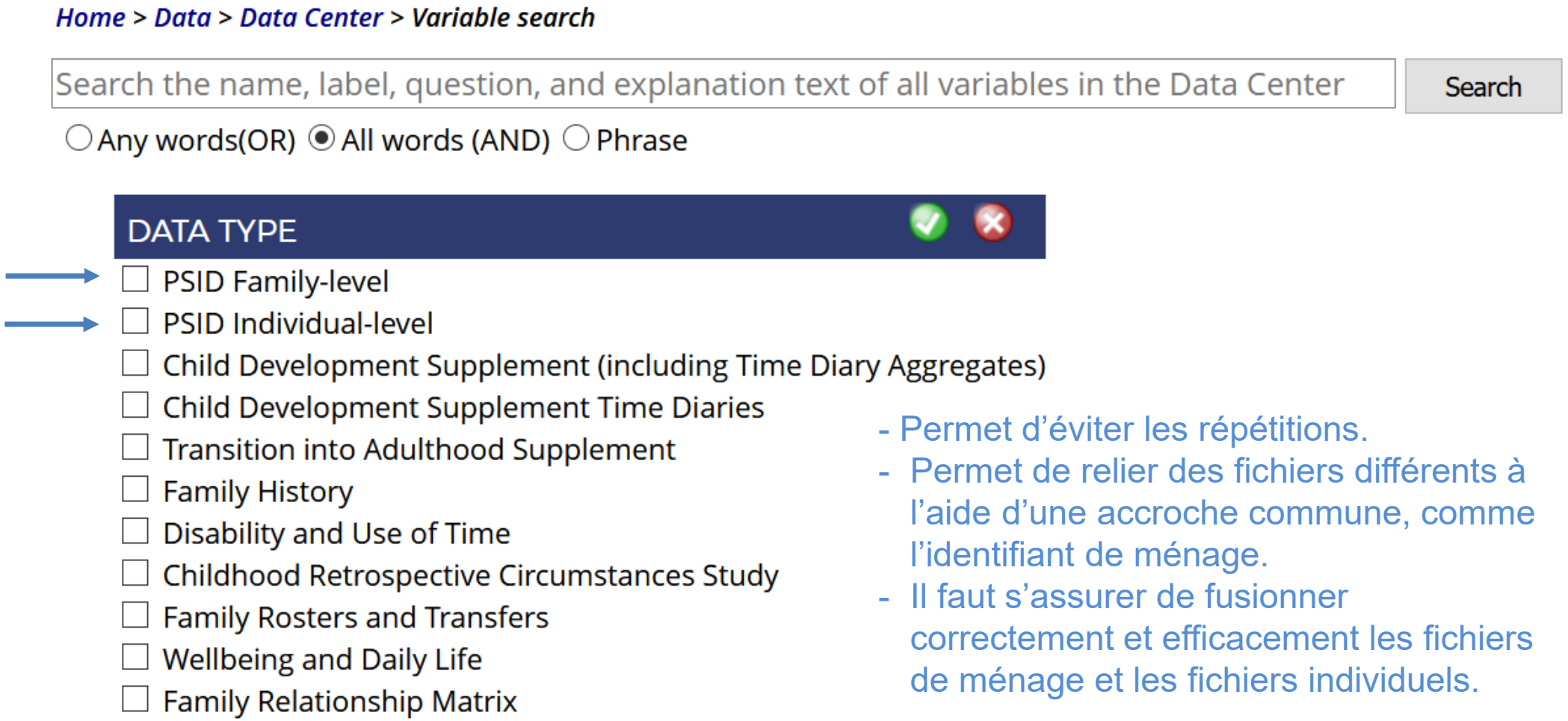
- Définit une ligne dans une base de données. Doit absolument être unique. Attention à éviter les données manquantes.

- Numérique en général, il peut s'agit d'un simple numéro de série. Il est aussi possible de concaténer des segments séparés.
  - Par ex. ID ménage = ID état + ID district+ ID PSU (PSU = unité d'échantillonnage primaire) + numéro de série au sein du PSU *(ici, « + » ne représente pas une addition mais une concaténation)*
  - Facilite l'identification au besoin mais génère aussi un risque de divulgation, à utiliser avec prudence.

Préparation des données

- Convertissez les formats larges (horizontaux) en formats longs (hiérarchisés).

- Répartissez les bases de données en fonction des différentes unités d'analyse.



- Permet d'éviter les répétitions.
- Permet de relier des fichiers différents à l'aide d'une accroche commune, comme l'identifiant de ménage.
- Il faut s'assurer de fusionner correctement et efficacement les fichiers de ménage et les fichiers individuels.

Source : https://simba.isr.umich.edu/VS/s.aspx

Préparation des données

- Convertissez les formats larges (horizontaux) en formats longs (hiérarchisés).
- Répartissez les bases de données en fonction des différentes unités d'analyse.
- Veillez à éviter les doublons dans la base de données ; vérifiez le nombre d'enregistrements.
- Vérifiez la classe de variable (nombre entier, chaîne de caractères, par ex.).
  - Les codes « DK » (Ne sait pas)/« RF » (Refus) peuvent forcer le passage d'une classe numérique à une classe chaîne.
- Vérifiez l'exactitude de la labélisation des variables et des valeurs.

Préparation des données

- Faut-il réencoder des noms de variable ? S'il y a besoin de mettre des variables dans un certain ordre, par ex.

- Faut-il réencoder des valeurs de variable ?
  - Attention à ne pas assigner de codes DK/RF par erreur à des réponses correspondant en réalité à des codes de catégories normales.

- À cette étape, enregistrez les données comme une nouvelle version. D'une manière générale, il est recommandé d'enregistrer les données à une étape donnée du traitement si des changements importants ont eu lieu/sont prévus avant de passer à l'étape suivante.

- Ressources de l'International Household Survey Network (IHSN) : *https://guide-for-data-archivists.readthedocs.io/en/latest/*

Les activités sont nombreuses...

1) Encodage des réponses ouvertes

2) Préparation des données

3) Synthèse et visualisation

4) Édition des données

5) Imputation et pondération

6) Contrôle de la divulgation

7) Traitement final, documentation et diffusion

FIN DE LA VIDÉO 1