Libérer le potentiel de vos données

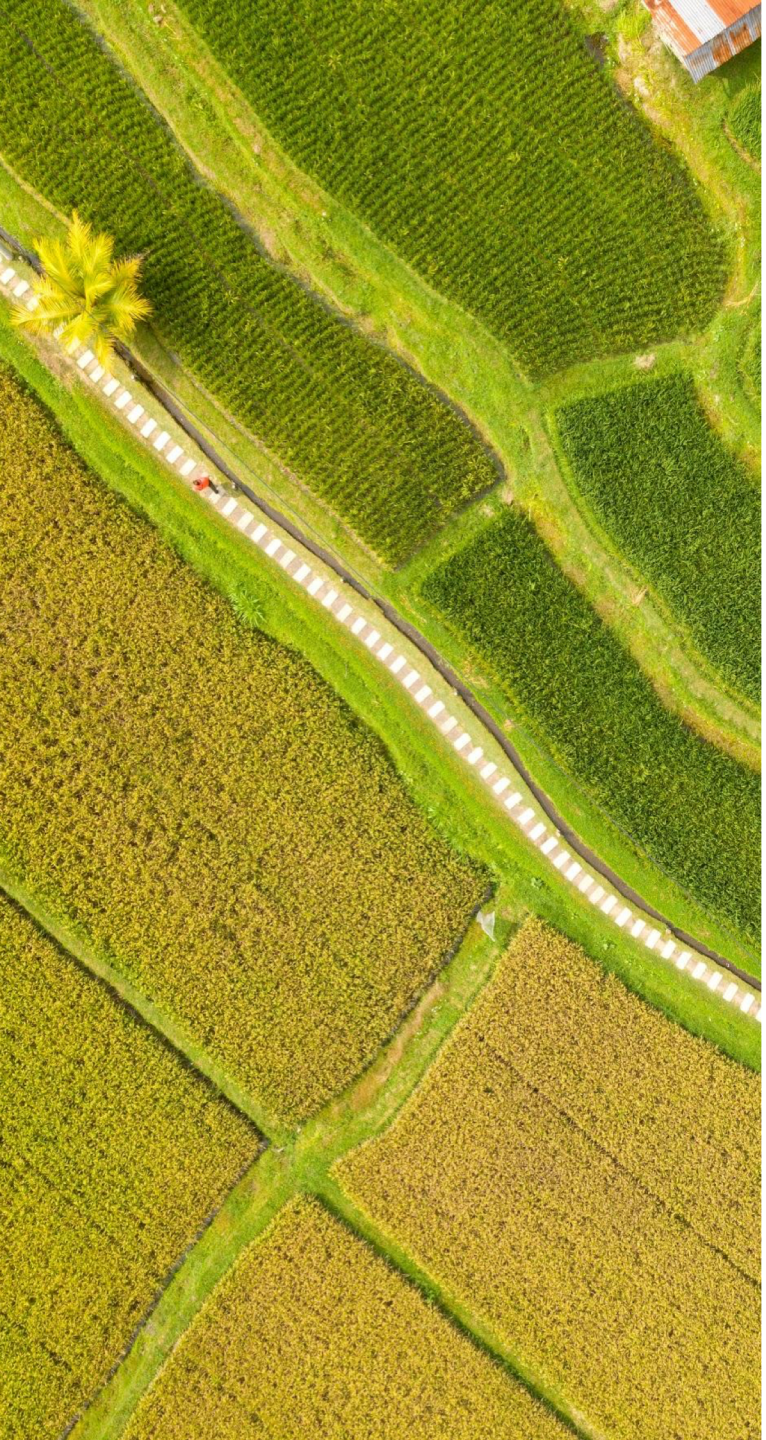
Atelier d'extraction, de traitement et de publication de données

26 Février 2019

EXPERTISE FRANCE

ANPTIC
Le label du numérique

Akvo.



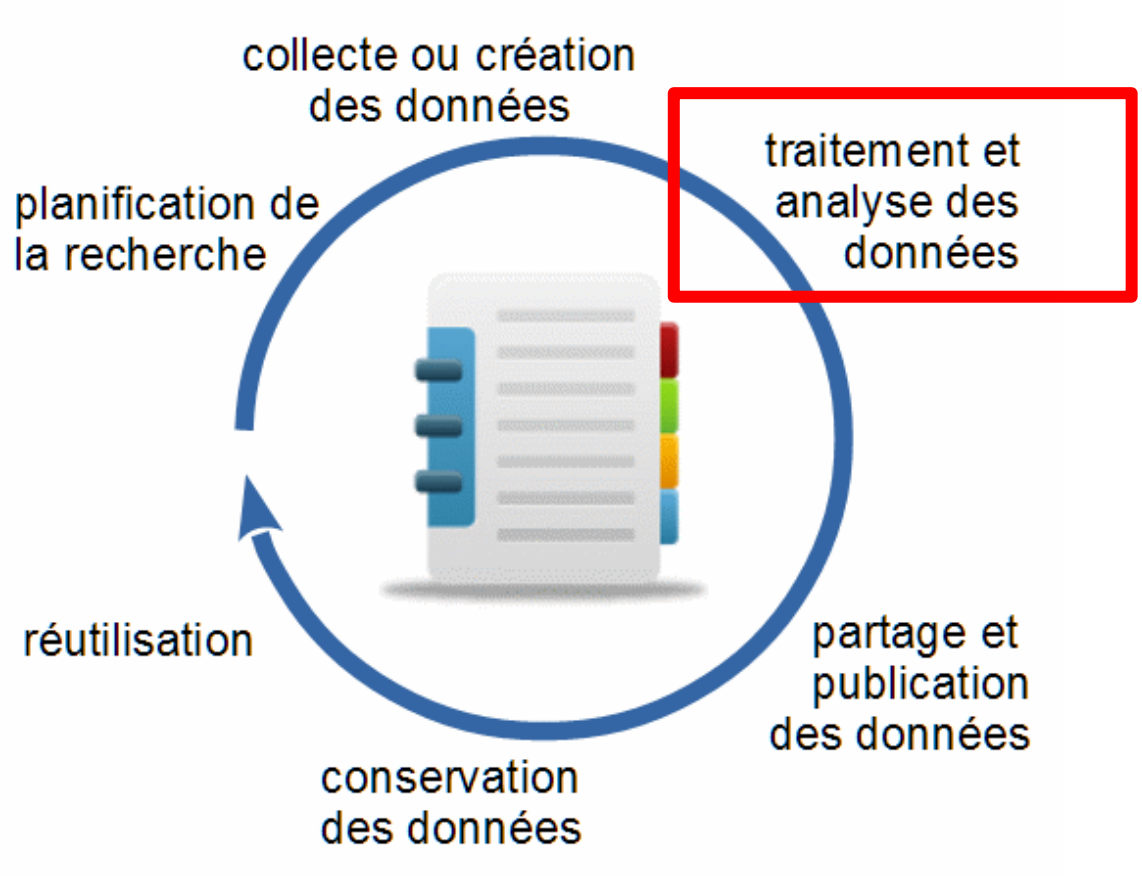
Sommaire

1. Production de jeux de données
2. Traitement et analyse des données
3. Publication des données
4. Valorisation des données (réutilisation des données)

Module II: Traitement et analyse des données

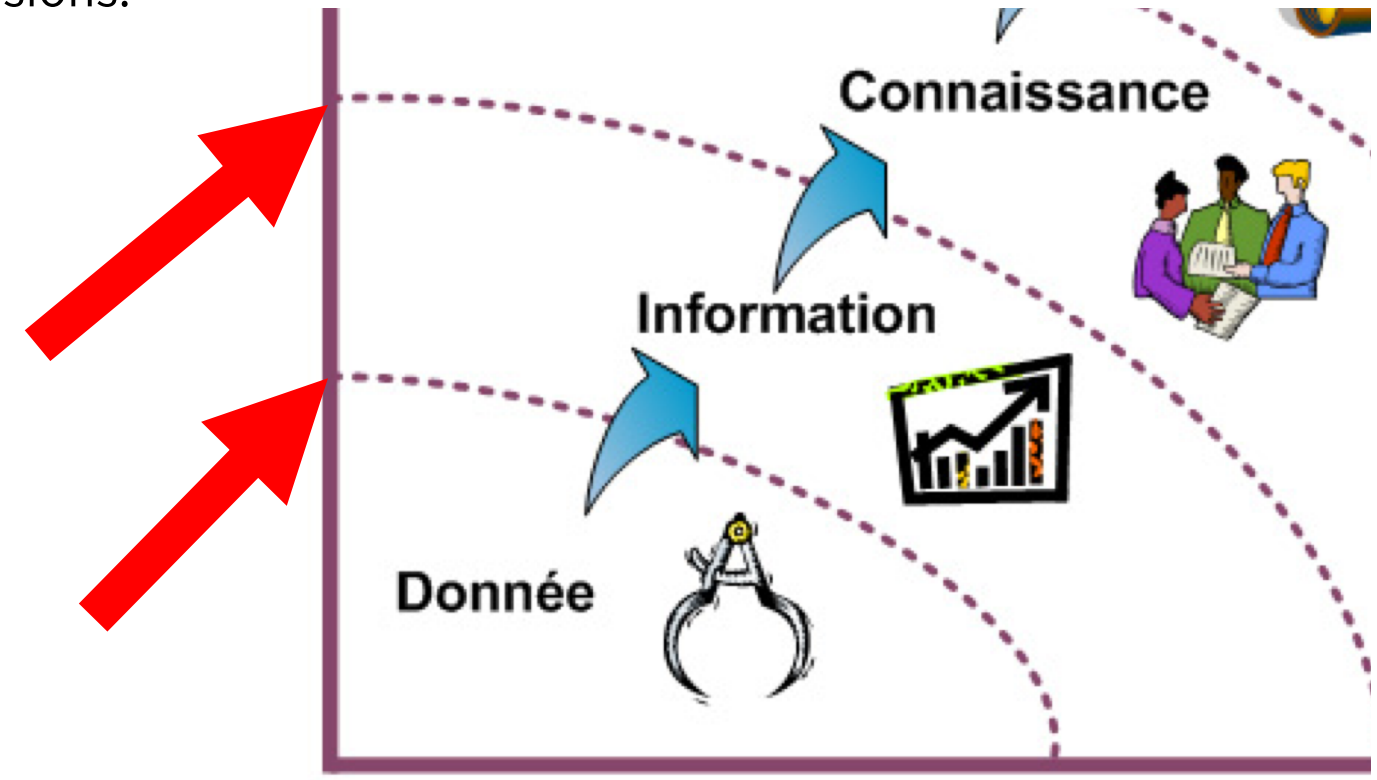
Traitement et analyse des données

- Troisième phase du cycle de vie de données.



Traitement et analyse des données

- L’analyse de données est un processus d'**inspection**, de **nettoyage**, de **transformation,** de **visualisation** et de **modélisation** des données
- dans le but de découvrir des informations utiles et de prendre des décisions.



Le nettoyage des données

Pourquoi nettoyer les données

- La qualité de l'information qui sera produite dépend fortement de la qualité des données analysées. **Les données de bonnes qualités permettent de prendre de bonnes décisions**

- Les données de bonnes qualité **améliore l'exploitabilité**

- Lors de la publication sur un portail (Open Data Burkina Faso), aucune mesure pour vérifier la qualité des données (on indique juste le format de données)

- Il faut donc nettoyer les données avant de passer à l'analyse/la publication.

Le nettoyage des données

- Nettoyer les données = vérifier la présence **éventuelle erreurs** dans votre jeux de données et de les corriger

- Une étude récente a démontré que le processus de préparation des données pour l'analyse peut prendre **60 à 80% du temps** total nécessaire pour un projet centré sur des données

- **Données propres/Données nettoyées** sont des données prêtes à être analyser ou publier

Les erreurs de données les plus fréquentes

- **Le mauvais format pour les dates**

Exemples: 8-sep-2013 → 8/9/2013 → 9/8/2013

- **Les échelles numériques cassées**

Exemples: 1200000→ 1,2M alors que 800000→ 800000

- **Les fautes d'orthographe/saisie (Valeurs aberrantes/valeurs hors plage)**
- **Les enregistrements en double**
- **Les données redondantes**

Exemples des colonnes des données combinées ou encore reproduites pour améliorer la lisibilité de l'utilisateur et qui ne sont (Totaux)

Eléments de la qualité des données

Lorsque nous parlons de la qualité des données, les aspects suivant sont a prendre en compte

1. Complétude
2. Unicité
3. Temporalité
4. Validité
5. Exactitude
6. Fiabilité/Cohérence

1. Complétude

- La proportion de données qu’on a par rapport au **potentiel de "100% complet” :**
- Qu'est-ce qui est **requis** ?
- Quelle est la raison possible pour **les valeurs manquantes** ?
- Un champ vide, ne veut pas forcément dire que les données sont incomplètes

- Par exemple, la variable enceinte n'a pas de sens pour les hommes.

2. Unicité

- Chaque **enregistrement doit être unique**
- Par exemple : Ne pas enregistrer le même point d’eau 2 fois.

3. Temporalité

- Les données reflètent la **réalité de la période** soumise à l'analyse

- Par exemple : Si je veux faire une analyse sur l'évolution de la fonctionnalité d'un point d'eau sur 5 ans, il me faut des données de 5 années consécutives.

4. Validité

- Les données sont valides si elles correspondent **au format, au type et à l'étendue de leur définition** :

- Par exemple : Si par exemple la réponse pour l'âge d'un enquêté doit être un numéro, allant de 0 à 130, les valeurs en-dessous et au-dessus sont probablement invalides.

- Ou si le format de la réponse doit être un nombre et que la réponse est un texte, par exemple "deux ans" au lieu de "2".

5. Exactitude

- Les données **reflètent les caractéristiques de l'objet réel ou des objets réels qu'elles doivent représenter** :

- Par exemple :
- Si on demande le niveau d'éducation en cours et que la personne à 45 ans, une réponse disant "école primaire" est susceptible d'être invalide.

- Données géographique ou la exactitude/précision est importante

6. Fiabilité/Cohérence

- Est-ce que les données sont mesurées et collectées de manière cohérente en fonction de définitions et de méthodes normalisées ?

- Par exemple : Si à T0 la question suivante a été posée:
- Quelle est la fonctionnalité de la pompe ?
- Totalement fonctionnel
- Partiellement fonctionnel
- En panne
- Abandonnée
- La question et les réponses doivent être exactement pareilles au moment de mesure T+1

Checklist de nettoyage des données

Vérifiez si chaque colonne a des en-têtes uniques et descriptifs;

Vérifiez si chaque colonne est formatée de manière cohérente;

Vérifiez s'il y a des lignes en double ou des lignes manquantes;

Dans le cas de données textuelles, vérifiez si les données sont cohérentes en termes de cas;

Vérifiez s'il y a des fautes d'orthographe;

Vérifiez si les données ont des espaces supplémentaires;

Vérifier les valeurs numériques et le signe décimal - virgule contre point;

Vérifiez si les colonnes sont disposées dans l'ordre approprié ou logique;

Vérifiez s'il y a des cellules vides qui ne devraient pas l'être;

Vérifier s’il y’a des données à caractère personnel.

Pratique: nettoyage des données

Nettoyer votre jeux de données pur la publication

- Considérer les groupes formés ;
- et les jeux de données identifiés précédemment ;
- Choisissez l'outil un outil familier;
- Nettoyer votre jeux de données en vous servant de la checklist
- Convertir votre fichier de jeux de données propres a un format ouvert

What is your biggest data challenge?

Find out more on www.Akvo.org

Ilyasse KABORE
Technical Account Manager

Email ilyasse@akvo.org
Twitter @KABOREIlyasse

Camille Clerx
Project Officer

Email camille@akvo.org
Twitter @ camille_Xmedia