Qualité des données de la production à l'analyse

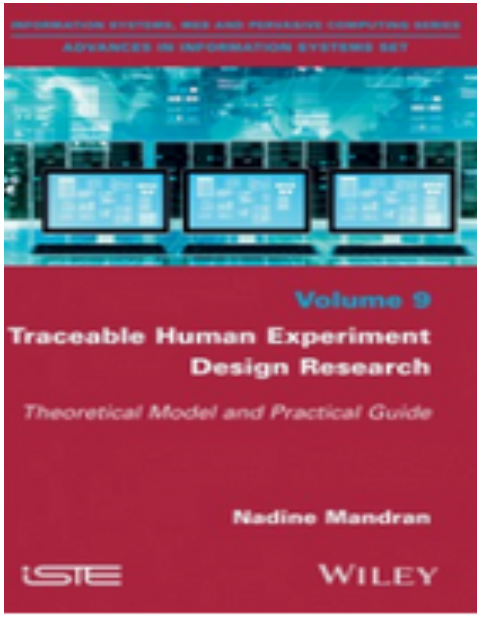
Nadine Mandran

LIG/CNRS

Création : février 2022

Quelques mots de présentation …

Nadine Mandran, CNRS-LIG



- Statistiques, SHS, Démarche qualité
- Ingénieure encadrant des doctorant.e.s sur les méthodes de production et analyse de données
- Recherche dans les méthodes de conduite de la recherche en informatique centrée humain
- https://thedre.imag.fr

Une thèse

- à identifier un sujet,
- **à rassembler des documents (données) sur ce sujet**
- **à les analyser, les organiser,**
- à construire une proposition, **à l'évaluer**
- se faire comprendre, défendre ses idées

Aussi

- on oubliera ainsi de parler comme un poète
- pas non plus obligé de tout lire sur le sujet mais le cœur de cible

Une thèse est une expérience de travail, mais c'est aussi un pari, un jeu, une chasse au trésor.

Le vrai défi dans une thèse, c'est de la vivre comme un défi, de se mettre soi-même au défi.

Avant propos …

Méthodologie ou méthode ?

- Méthodologie : étude de la méthode, conception de méthodes

  *« **Methodology** is the systematic, theoretical analysis of the methods applied to a field of study. It comprises the theoretical analysis of the body of methods and principles associated with a branch of knowledge. Typically, it encompasses concepts such as paradigm, theoretical model, phases and quantitative or qualitative techniques » (Berg, 2009).*

- Méthode : ensemble de tâches, d'outils, de procédures assemblés et organisés dans le temps pour atteindre un but.

  Method is "an integrated collection of procedures, techniques, product descriptions, and tools, for effective, efficient, and consistent support of the engineering process" (Harmsen et al., 1997).

- **Nous utilisons uniquement le terme méthode que nous qualifions par d'autres termes : méthode de conduite de la recherche, méthode de production de données, méthode d'analyse de données** …

Tour de table

- Votre discipline ?
- Vos données ?
- Le niveau de sensibilité de vos données ?
- Quelles questions vous vous posez sur la qualité des données ?
- Quel intérêt pour votre travail de thèse ?

Organisation du cours

- Données
- Répétabilité, Reproductibilité, Traçabilité ?
- Data Management Plan
- Temporalité pour production des données
- Critères de qualité des données
- 3 Guides

Produire et analyser des données

- Données existantes
  - Sociologie: Etudes existantes, Recensement de la population
  - Informatique : Base des Tweets
  - Histoire : Archives, Textes de lois
- Données à produire
  - Sociologie : Questionnaires d'intention de votes
  - Informatique : Tests utilisateurs d'un site web
  - Histoire : Faire des interviews

Produire et analyser des données

- Données factuelles
  - Sociologie: Recensement, comptage
  - Informatique : Capteurs de présence
- Données déclaratives
  - Sociologie : Questionnaires d'intention de votes
  - Informatique : Avis sur l'utilisabilité d'un site web
  - Histoire : Interviews
- Données documentaires
  - Histoire : Documents historiques

Produire et analyser des données

- Données Quantitatives => Quantifier
  - Sociologie: Recensement, comptage
  - Informatique : Capteurs de présence, traces d'activités
- Données Qualitatives => Comprendre
  - Sociologie : Interviews
  - Informatique : Focus groups
  - Histoire : Interviews
-

Produire et analyser des données

- Multiples
- De nature différentes
- Dépendants des disciplines

Répétabilité vs Reproductibilité

- ISO 3534-1
  - **Conditions de répétabilité**
    - Conditions où les résultats d'essais indépendants sont obtenus par la **même méthode sur des individus d'essai identiques** dans le même laboratoire, par le même opérateur, utilisant le même équipement et pendant un court intervalle de temps.
  - **Conditions de reproductibilité**
    - Conditions où les résultats d'essai sont obtenus **par la même méthode sur des individus d'essais identiques dans différents laboratoires,** avec différents opérateurs et utilisant des équipements différents

Répétabilité vs Reproductibilité

- Les protocoles de production des données sont reproductibles (?) :
  - Ils peuvent être réutilisés avec d'autres utilisateurs dans des conditions différentes ou similaires
  - Nécessite de tracer le processus de production et de traitement des données
  - Nécessité de tracer les évolutions de la production scientifique (connaissance scientifique et des outils)

Data Management Plan (modèle ANR)

- Description des données
  - Comment de nouvelles données ou des données existantes seront utilisées ?
  - Quelles données : types, format, volumes ?

Source : https://opidor.fr/
https://dmp.opidor.fr/public_templates

Data Management Plan (modèle ANR)

1- Documentation et qualité des données

- Quelles métadonnées, quelle documentation ?
- Quelles méthodes de production ?
- Quelles mesures de contrôle de la qualité ?

Source : https://opidor.fr/
https://dmp.opidor.fr/public_templates

Data Management Plan (modèle ANR)

2- Stockage et sauvegarde pendant le processus de recherche

- Comment les données et les métadonnées seront stockées et sauvegardées ?
- Comment la sécurité des données et la protection des données sensibles seront-elles assurées ?

Source : https://dmp.opidor.fr/public_templates

Data Management Plan (modèle ANR)

3-,Exigences légales et éthique, code de conduite

- Règlement de protection des données ?
- Cnil ?
- Comité d'éthique ?

Source : https://opidor.fr/https://dmp.opidor.fr/public_templates

Data Management Plan (modèle ANR)

4- Partage des données et conservation à long terme

- Comment et quand les données seront-elles partagées ? Quelles sont les restrictions ?
- Comment les données à conserver seront-elles sélectionnées ?
- Où seront-elles archivées ?
- Quels méthodes et outils seront nécessaires pour accéder et utiliser les données ?
- Comment l'attribution d'un identifiant unique et pérenne (DOI) sera-t-elle assurée ?

Source : https://opidor.fr/
https://dmp.opidor.fr/public_templates

Data Management Plan (modèle ANR)

5- Responsabilité et ressources en matière de gestion des données

- -Qui ? Rôle et responsabilités concernant la gestion des données ? De la production à l'archivage ?
- Quelles seront les ressources pour la préparation du partage des données ? (temps, budget, matériel)

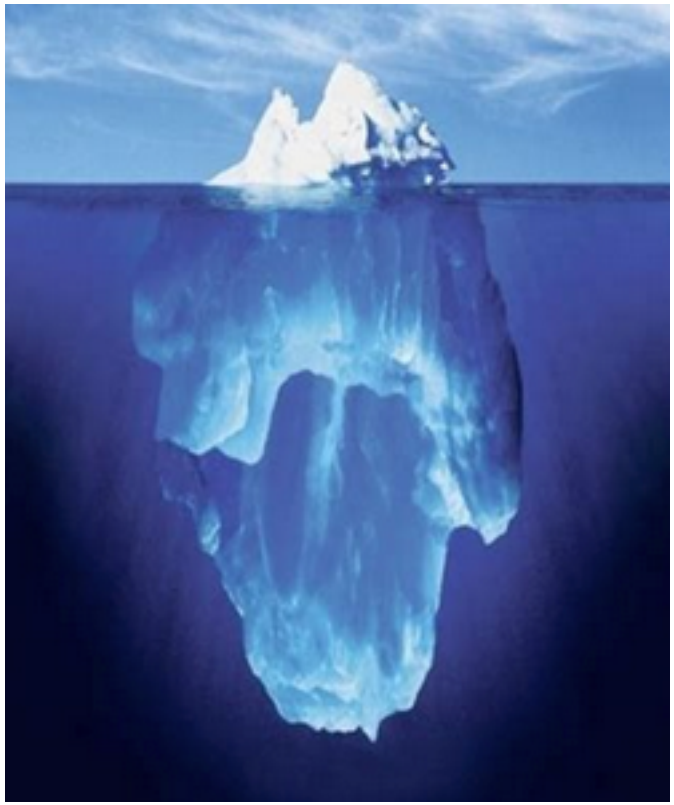
Source : https://opidor.fr/
https://dmp.opidor.fr/public_templates

Qualité des données

- La qualité des données est critique par essence.
- Tâche effectuée avec une donnée erronée engage un coût 100 fois supérieur à celui d'une tâche réalisée à partir d'une donnée initialement vérifiée et correcte.(Harvard Business Review , 2017)
- Plus de 25 % des données critiques des plus grandes entreprises sont erronées,
- Le coût moyen d'une mauvaise qualité des données pourrait s'élever à 11M€ par an pour les organisations.(analyse Gartner 2020)

https://datavalue-consulting.com/bonnes-pratiques-qualite-des-donnees/

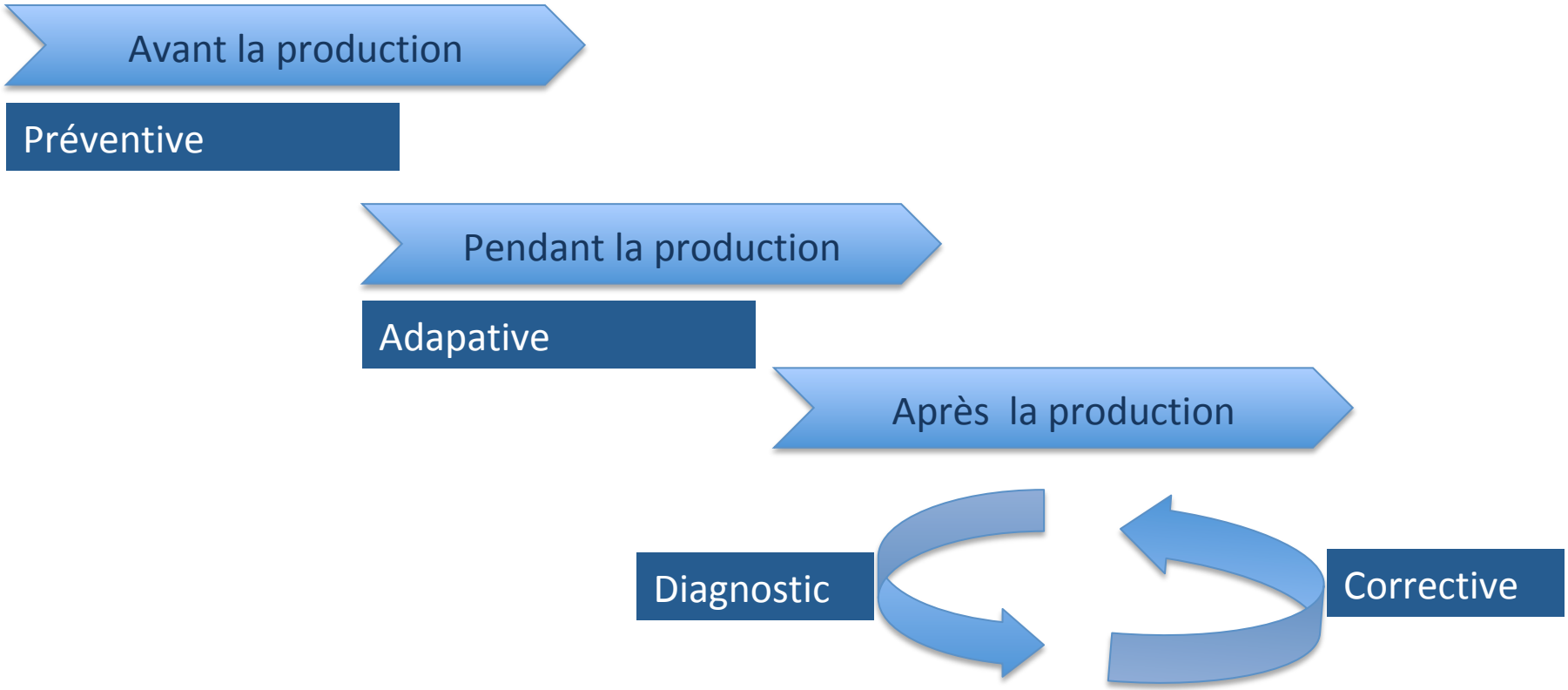
Pré traitement et analyse des données quantitatives



Analyse : 10%

Pré-traitement : 80% du temps

Quatre étapes

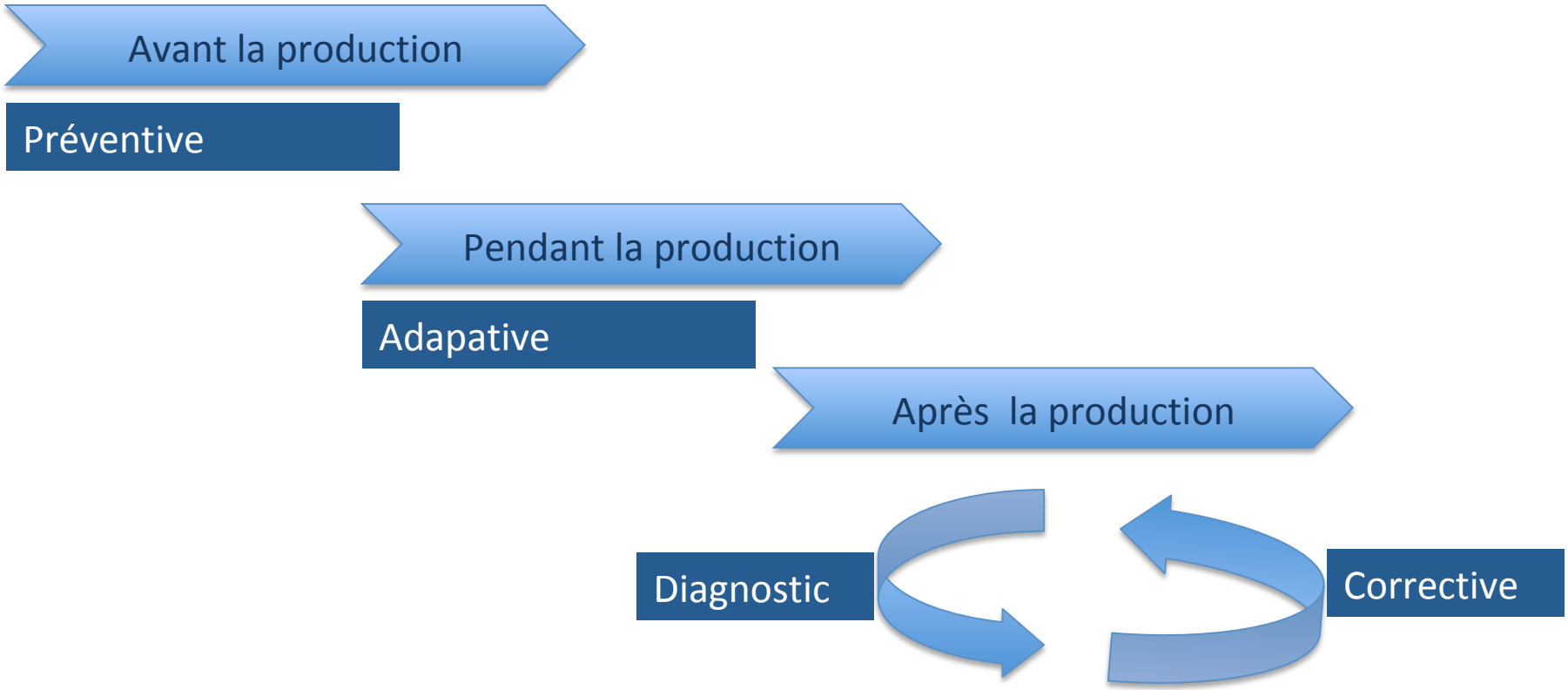


Inspiré de Berti-Equille, L. (2007). Measuring and modelling data quality for quality-awareness in data mining. In Quality measures in data mining (pp. 101-126). Springer, Berlin, Heidelberg.

Quatre étapes

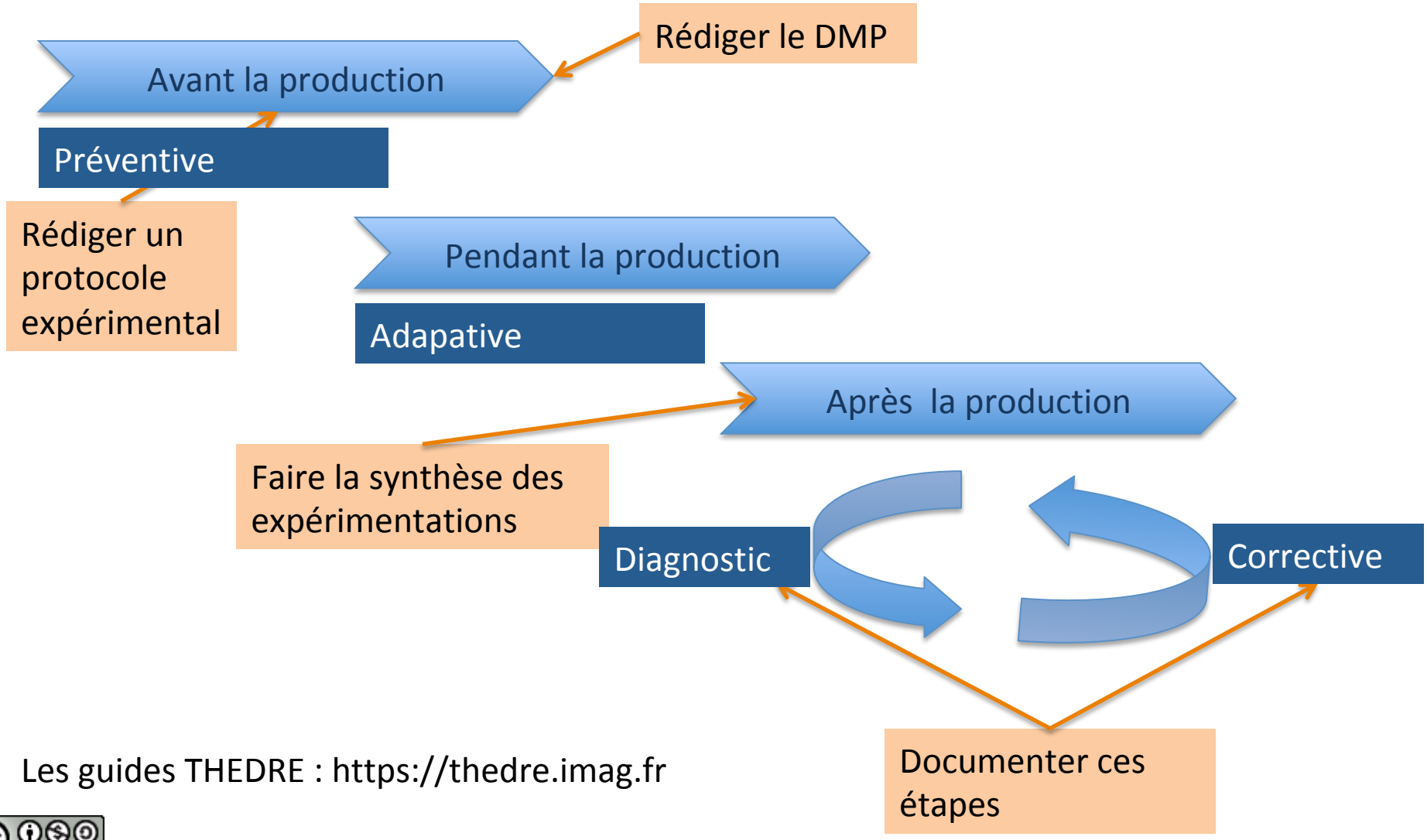
- L'approche **préventive** permet un contrôle en amont avant la production des données (p.ex., un test de production des données par un capteur garantira que les données produites en temps réel sont correctes).
- L'approche **adaptative** permet la vérification des données en temps réel (p.ex.: pendant la capture de données une application permet d'identifier les données aberrantes, par exemple une augmentation soudaine de la température sur un capteur).
- Les approches **diagnostiques et correctives** sont menées après la production de données.
- L'approche **diagnostique** comprend la comparaison avec la réalité sur le terrain et la gestion des métadonnées.
- L'approche **corrective** comprend, entre autres, la correction par rapport à la réalité du terrain, l'imputation de données manquantes, le redressement et l'élimination des doublons.

Quatre étapes



Identifiez ce que vous mettez en place pour ces étapes ?

Quatre étapes



Les guides THEDRE : https://thedre.imag.fr

10 critères de qualité des données

1. **Pertinence** : capacité des données répondre aux besoins actuels et futurs des utilisateurs.

2. **Exactitude/Justesse** : mesure de la conformité des données par rapport à la réalité. (p.ex., la taille des individus ne peut pas être supérieur à 2,50 m ni inférieur 0,40 cm).

3. **Précision temporelle** : exactitude des données par rapport à l'instant qu'elles sont censées représenter. Le chercheur a besoin d'avoir des données qui décrivent une situation telle qu'elle est ou était à un moment précis. (p.ex., les bilans pour une entreprise sont enregistrés avec l'année de référence).

Di Ruocco, N., Scheiwiler, J. M., & Sotnykova, A. (2012). La qualité des données: concepts de base et techniques d'amélioration. *La qualité et la gouvernance des données, Hermes-Lavoisier, Paris*.

10 critères de qualité des données

4. **Accessibilité** : la facilité de localisation et d'accès aux données et aux métadonnées.

5. **Facilité d'interprétation :** facilité de compréhension des données, de leur analyse et de leur usage. Les données doivent être bien documentées pour être comprises sans ambiguïté.

6. **Unicité** : garantie qu'une entité du monde réel est représentée par un seul et unique objet, il s'agit de contrôler la présence des doublons.

7. **Cohérence** : absence d'informations conflictuelles. (p.ex., l'âge des enfants doit être inférieurs à celui de leurs parents).

Di Ruocco, N., Scheiwiler, J. M., & Sotnykova, A. (2012). La qualité des données: concepts de base et techniques d'amélioration. *La qualité et la gouvernance des données, Hermes-Lavoisier, Paris*.

10 critères de qualité des données

8. **Conformité à une norme :** respect d'une norme standardisée ou d'une convention de nommage (p.ex., la profession de la personne est codée selon la norme INSEE : PCS en 8 catégories).

9. **Complétude** : Ce critère est utilisé dans les approches préventives. Car il s'agit de contrôler si les objets nécessaires à la production des données sont présents dans le modèle de données. La complétude se juge en fonction selon 4 critères : entités, attributs, relations et occurrences. (p.ex., pour les entités, une base de données des clients est incomplète s'il manque l'adresse de facturation, p.ex., pour les relations, une personne peut aller dans plusieurs salles de cinéma, le modèle doit comporter une relation «voir des films » liant les entités « personne » aux entités « salles de cinéma ».

10. **Consistance** : Quand une entité est recopiée, il y a consistance si on retrouve les mêmes valeurs d'attributs dans toutes les bases.

Di Ruocco, N., Scheiwiler, J. M., & Sotnykova, A. (2012). La qualité des données: concepts de base et techniques d'amélioration. *La qualité et la gouvernance des données, Hermes-Lavoisier, Paris*.