Quel lien entre qualité et données ?

Alain Rivet
alain.rivet@cermav.cnrs.fr
Henri Valeins
henri.valeins@rmsb.u-bordeaux.fr

Atelier-données
5 juillet 2021

Quel lien entre qualité et données ?

- Qualité, données ?
- Gouvernance des données
- La non qualité des données
- Management des données et management de la qualité

Nous amorçons ce printemps de data.gouv.fr sur la question de la qualité des données.

La qualité des données est un élément essentiel du succès de l'open data : l'ouverture des jeux de données n'entraîne pas directement leur réutilisation. Ce constat s'explique notamment par les difficultés que rencontrent les réutilisateurs lorsqu'ils souhaitent s'approprier les données ouvertes.

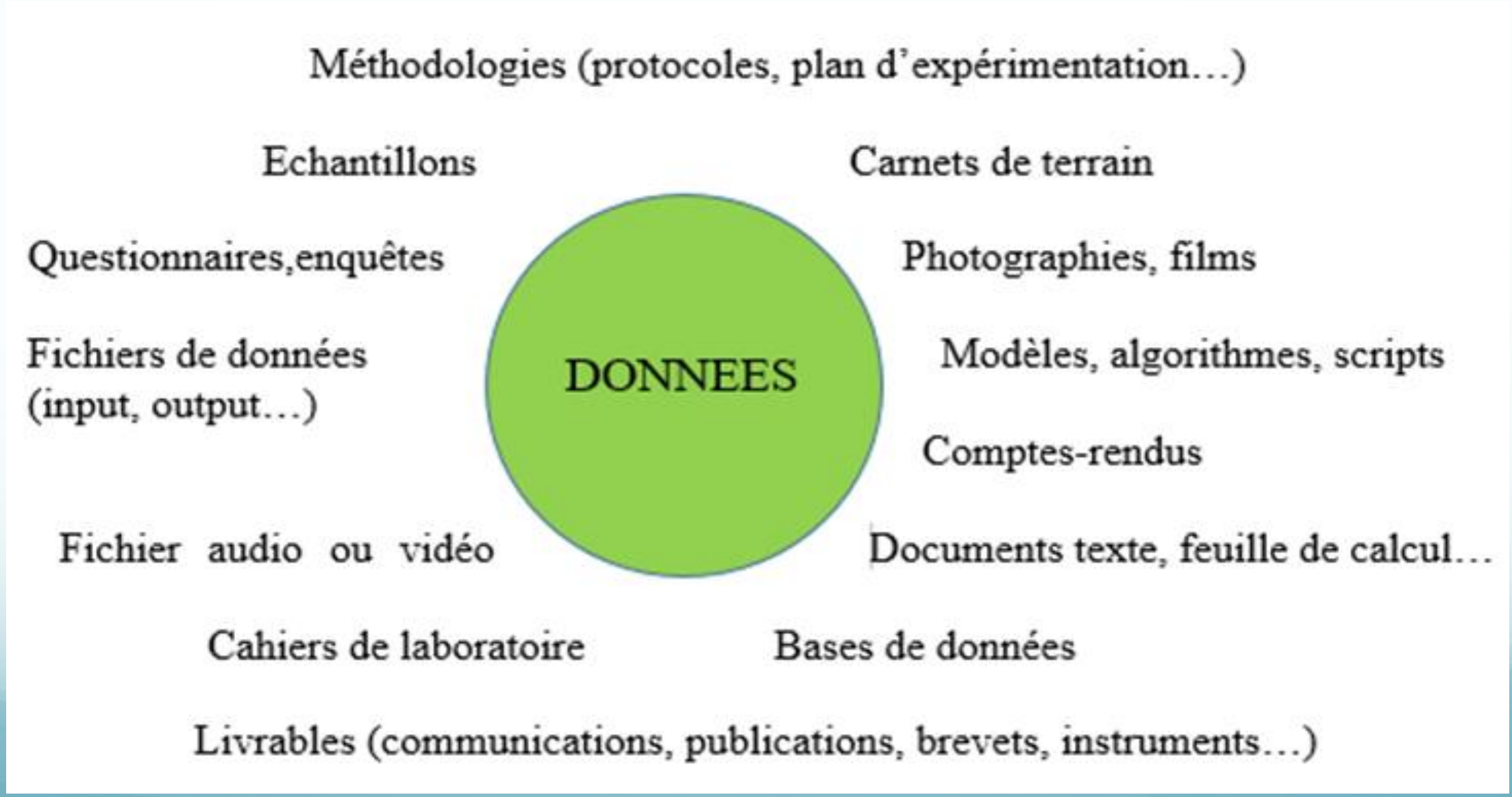
D'ailleurs, l'analyse de l'enquête auprès des usagers (905 répondants de juin à septembre 2020) pointe une **véritable attente des utilisateurs de la plateforme sur la qualité des données**. Les répondants remontent des problèmes de mise à jour avec des jeux de données souvent obsolètes, une documentation insuffisante ou inexacte quand elle existe, la multiplicité de jeux de données ou encore le manque d'échanges entre producteurs et réutilisateurs de données malgré le système de commentaires de data.gouv.fr. **En somme, la qualité n'est pas suffisamment au rendez-vous.**

Données de la recherche

- « Enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche.».
- « Ce terme ne s'applique pas aux éléments suivants : carnets de laboratoire, analyses préliminaires et projets de documents scientifiques, programmes de travaux futurs, examens par les pairs, communications personnelles avec des collègues et objets matériels (par exemple, les échantillons de laboratoire, les souches bactériennes et les animaux de laboratoire tels que les souris). »

« Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics » OCDE, 2007

La diversité des données



Les concepts

- Définition ISO 9000 de la qualité
  - « Aptitude d'un ensemble de caractéristiques intrinsèques à satisfaire des exigences »

Caractéristique
Trait distinctif (qualitatif ou quantitatif)

Caractéristique intrinsèque
Trait permanent

Satisfaire
Satisfaction par rapport à une demande

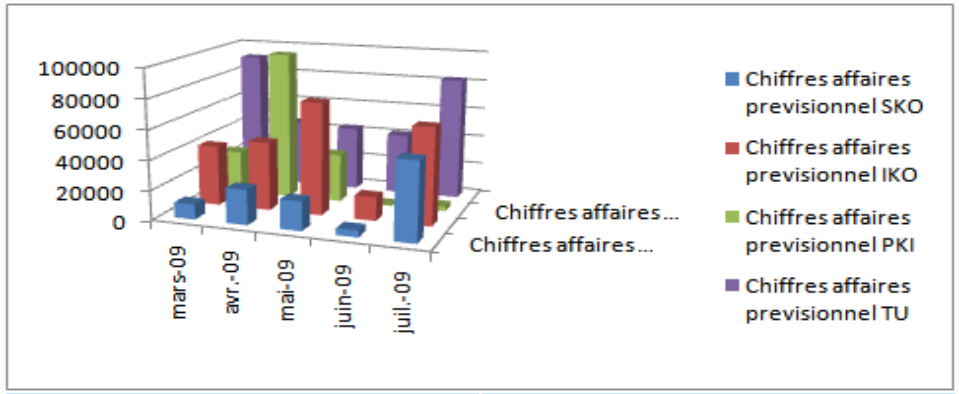
Exigence
Besoin formulé ou imposé

- Pratiquement
  - La qualité d'un produit signifie qu'il est adapté au besoin qu'il est censé satisfaire

Différentes visions de la qualité

- Chercheur : publier dans revues fort à facteur d'impact, avoir une reconnaissance de ses pairs
- Ingénieur de plateforme : produire des données robustes
- Spécialiste IST : disposer de données FAIR
- Economiste : favoriser l'innovation, la création d'emplois

Définitions

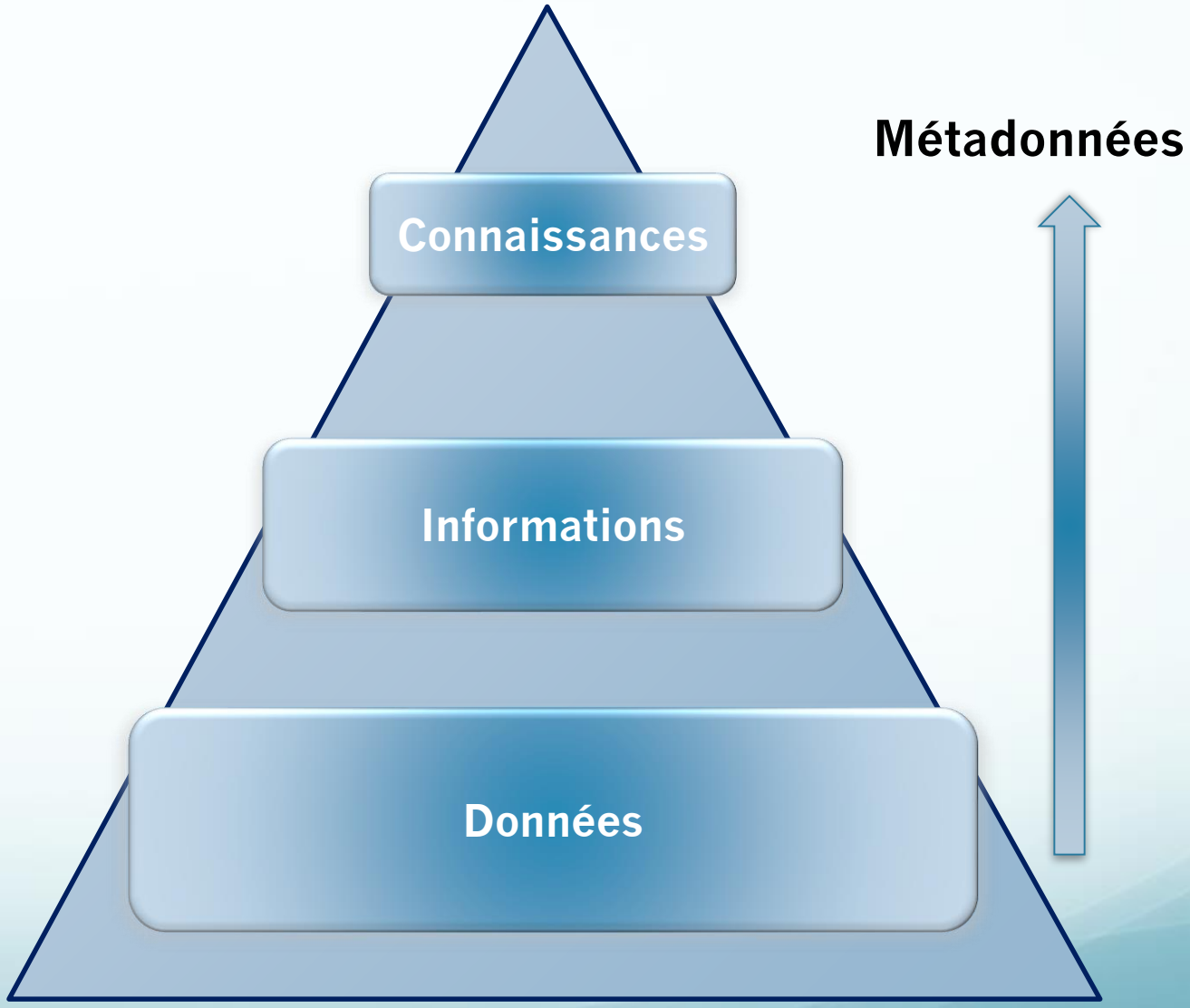


	MARS 2009
Chiffre d'Affaires	13 245 €

13 245 2009

MARS

Affaires

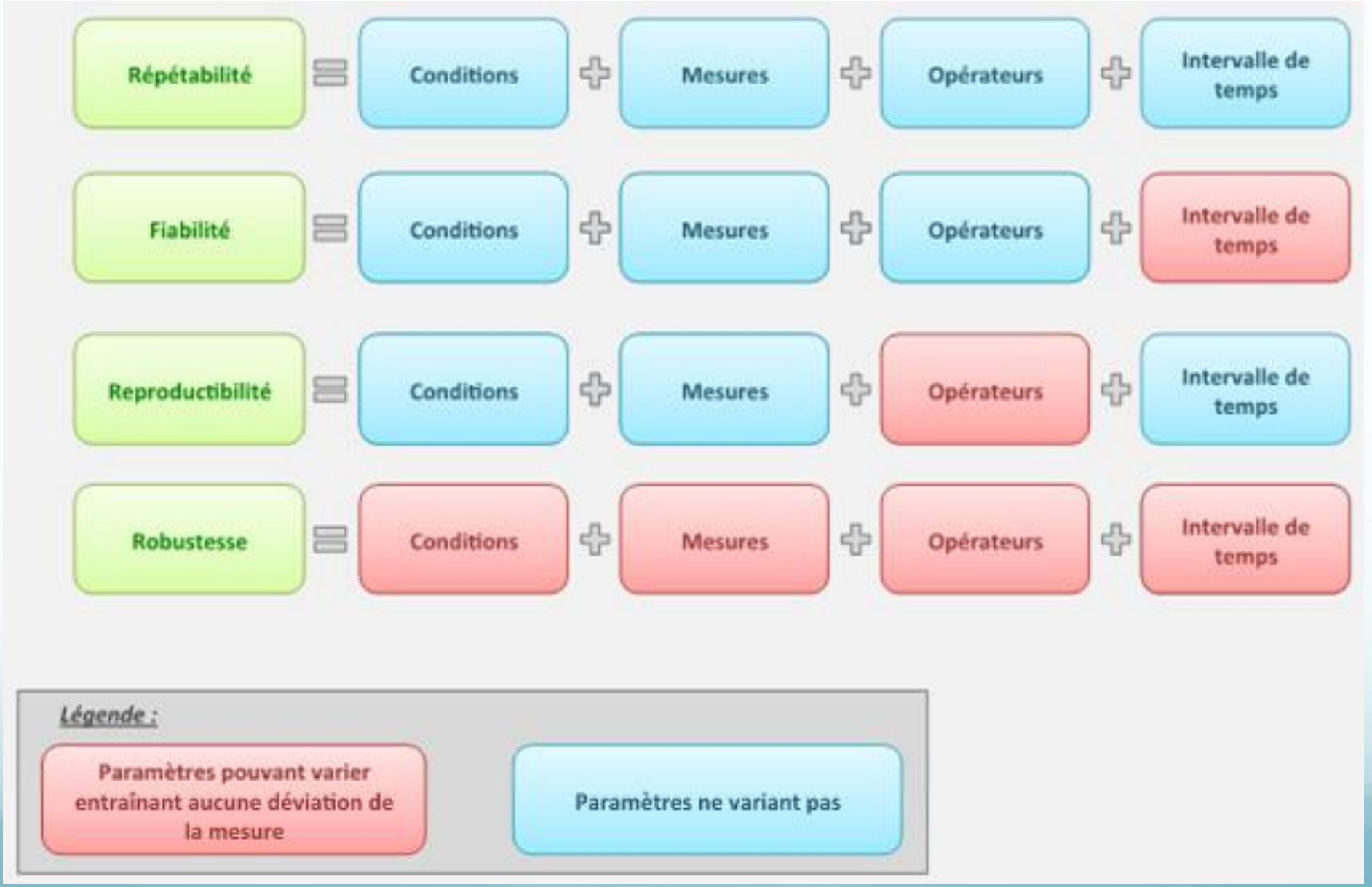


La qualité de la recherche

- La recherche : produit et traite des connaissances
  - Produit final mal connu
  - Maîtriser l'ensemble des moyens d'acquisition, de conservation et de diffusion des résultats (notions de validité des mesures)
  - Evaluation par les pairs

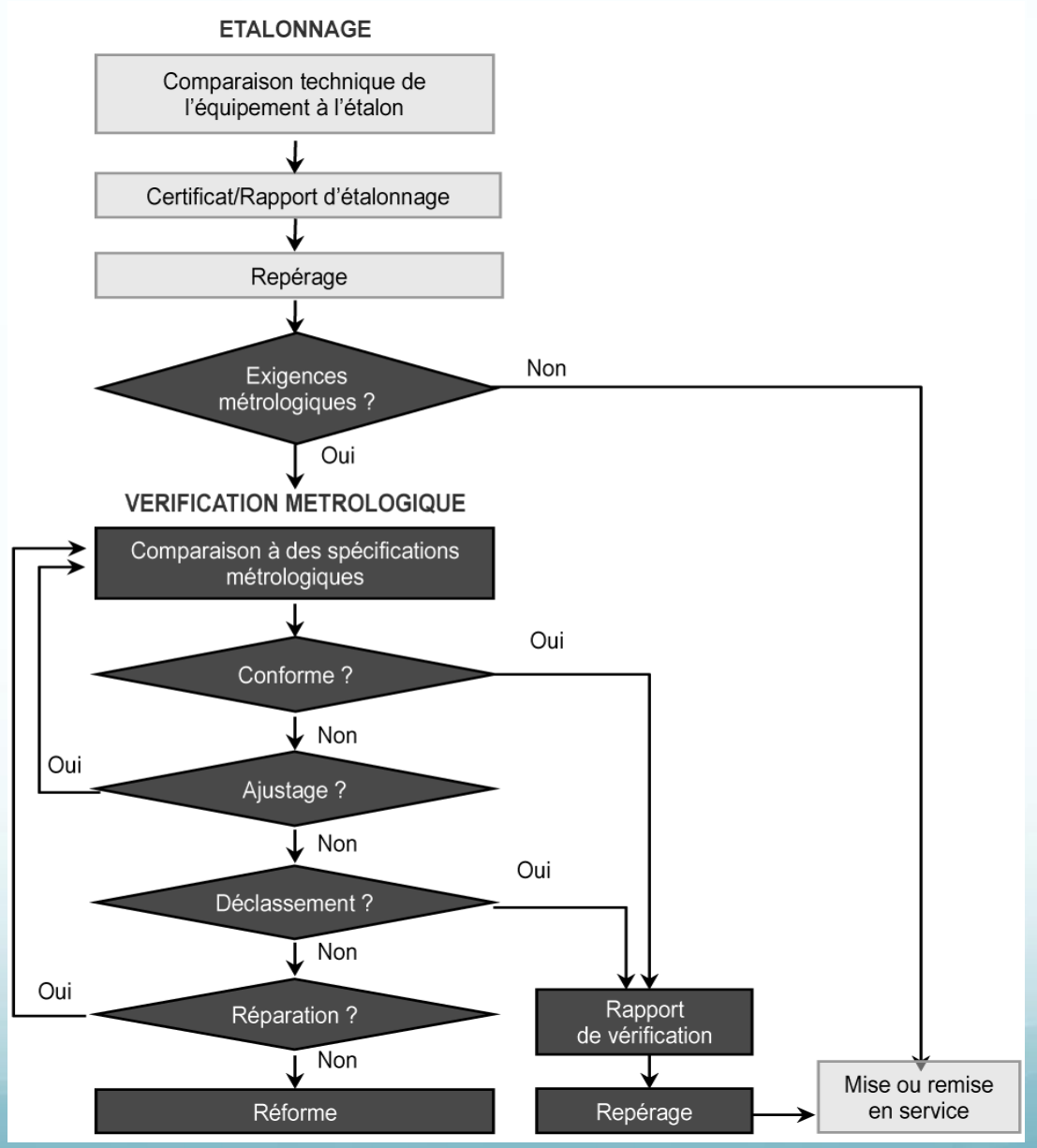
- La confiance en la qualité d'une recherche consiste à établir et vérifier que les différentes étapes d'une étude pourront être répétées en obtenant le même résultat par des chercheurs différents et/ou à des moments différents.

Qualité intrinsèque des données



Management de la qualité : l'assurance d'une recherche robuste et fiable, Réseau Inserm Qualité

Etalonnages et vérifications



Qualité des jeux de données

Plusieurs éléments permettent d'évaluer le niveau de qualité d'un jeu de données :

- Des éléments sur les données elles-mêmes et leur structure : format de fichier, structure du fichier…
- Des éléments attestant du potentiel de réutilisation et de croisement des données :
  - Le respect de standards, référentiels et schémas déjà établis ;
  - La présence de données et colonnes pivots pour lier les données à un référentiel (Le code SIRET ou SIREN par exemple).
- Des éléments qui accompagnent les données (documentation claire et rigoureuse, gestion des versions et des mises à jour des données…

Le cycle de vie de la donnée

- Disposer de données FAIR :
  - « Findable, Accessible, interopérable et Reusable »
  - Facile à trouver, Accessible, Interopérable et Réutilisable

Gouvernance des données

- Connaissances : « Ensemble disponible d'informations constituant une conviction justifiée et ayant une forte certitude d'être vraies » (ISO 9001)

- Passer de structures centrées sur les processus qui s'exécutent indifféremment des données manipulées vers des structures qui mettent les données au cœur de leurs actions

Quels sont les risques de la non qualité des données ?

On peut considérer que des données ne sont pas de qualité lorsqu'elles sont:

- **Inexactes**: Ce qui peut entraîner des conclusions erronées et donc mettre en doute l'intégrité scientifique
- **Non-conforme**: Données qui ne répond pas au référentiel choisi, ce qui implique que l'on a défini ce qu'est une donnée conforme
- **Non-contrôlée**: Il faut vérifier que les données soient exactes et conformes pour les considérer comme fiables
- **Non-sécurisée ou non-fiable**: c'est une donnée non contrôlée

Rétractations

- En 2016, 33 % des rétractations étaient dues à des erreurs dans les données

CAMPOS-VARELA (I.) et RUANO-RAVIÑA (A.). – Misconduct as the main cause for retraction. A descriptive study of retracted publications and their authors, Gaceta Sanitaria, vol. 33, issue 4, p. 356-360, juill.-août 2019.

Table 2
Reasons for retraction and proportion of misconduct by category.

Reason of retraction	Articles, n (%)	Misconduct, n (%)	
Plagiarism	354 (32.7)	Yes	354 (100)
		No	0
		Uncertain	0
Data	352 (32.5)	Yes	129 (36.6)
		No	1 (0.3)
		Uncertain	222 (63.1)
Review process compromised	152 (14.1)	Yes	152 (100)
		No	0
		Uncertain	0
Authors	64 (5.9)	Yes	42 (65.6)
		No	0
		Uncertain	22 (34.4)
Journal	47 (4.3)	Yes	0
		No	44 (93.6)
		Uncertain	3 (6.4)
Ethical	23 (2.1)	Yes	19 (82.6)
		No	1 (4.4)
		Uncertain	3 (13.0)
Conflicts of interest	7 (0.7)	Yes	7 (100)
		No	0
		Uncertain	0
Other	54 (5.0)	Yes	4 (7.4)
		No	8 (14.8)
		Uncertain	42 (77.8)
Unknown	29 (2.7)	Yes	0
		No	0
		Uncertain	29 (100)

Coût de la non qualité

- Harvard Business Review estimait en 2017 qu'une tâche effectuée avec une donnée erronée engageait un coût 100 fois supérieur à celui d'une tâche réalisée à partir d'une donnée initialement vérifiée et correcte.
- Selon l'analyse Gartner 2020 sur les solutions de gestion de qualité des données, plus de 25 % des données critiques des plus grandes entreprises sont erronées, au point que le coût moyen d'une mauvaise qualité des données pourrait s'élever à 11M€ par an pour les organisations. Les répercussions économiques, positives ou négatives, sont donc à considérer avec la plus grande attention.

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

Principes de la démarche qualité

C'est une démarche ***organisationnelle*** avec comme bases :

- Une approche processus
- L'amélioration continue
- La traçabilité

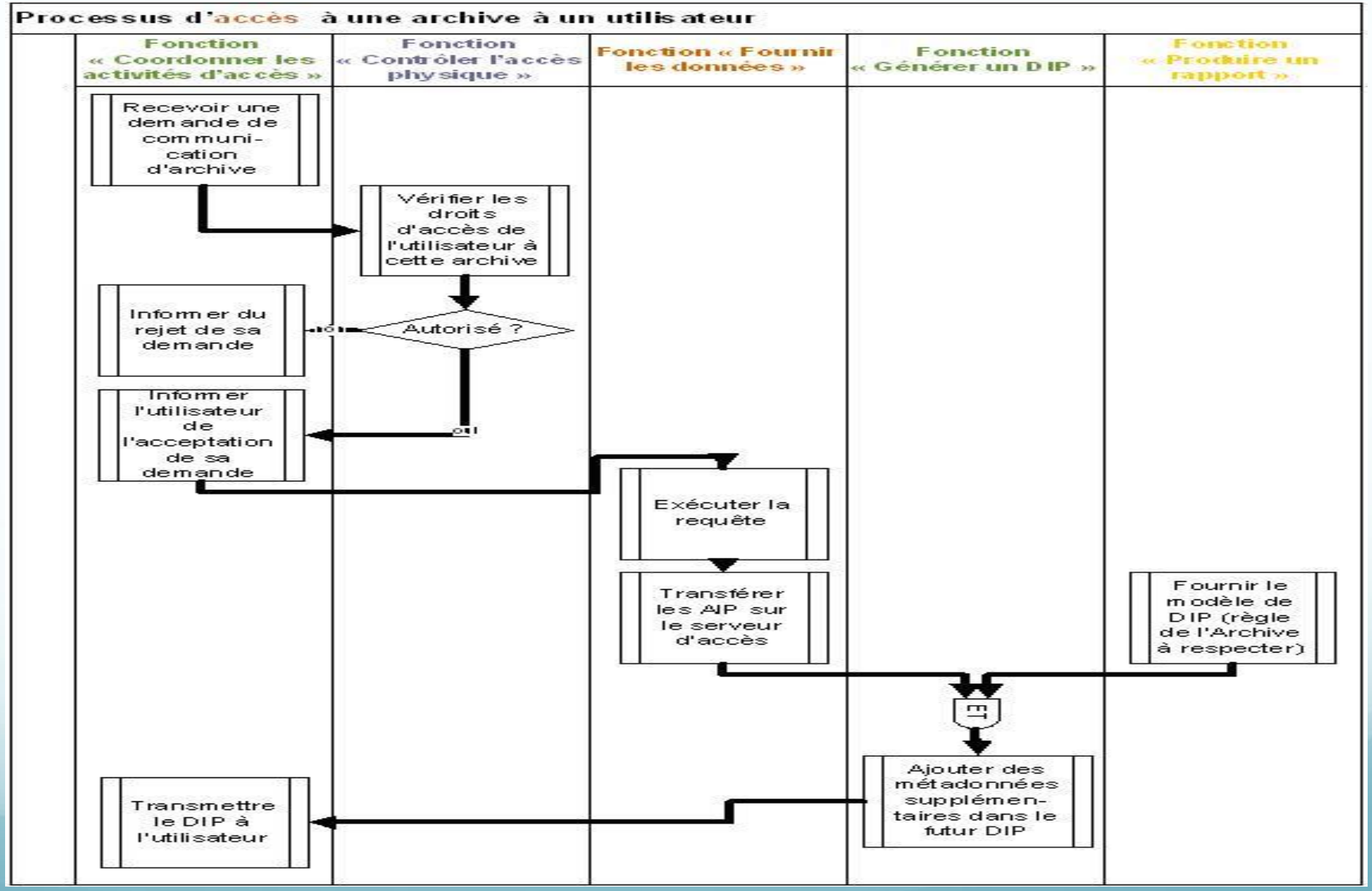
Pour pouvoir mettre en place et suivre une démarche qualité il faut utiliser un référentiel

Approche processus

- Quels sont les processus de production des données ?
- Quels sont les processus de gestion des données ?
- Quels sont les processus de contrôle des données ?
- Quels sont les processus de conservation des données ?
- Quels sont les processus de d'accessibilité des données ?
- Quels sont les processus de diffusion des données ?
- …

Approche processus

Exemple du processus d'accès aux archives du CINES



L'amélioration continue

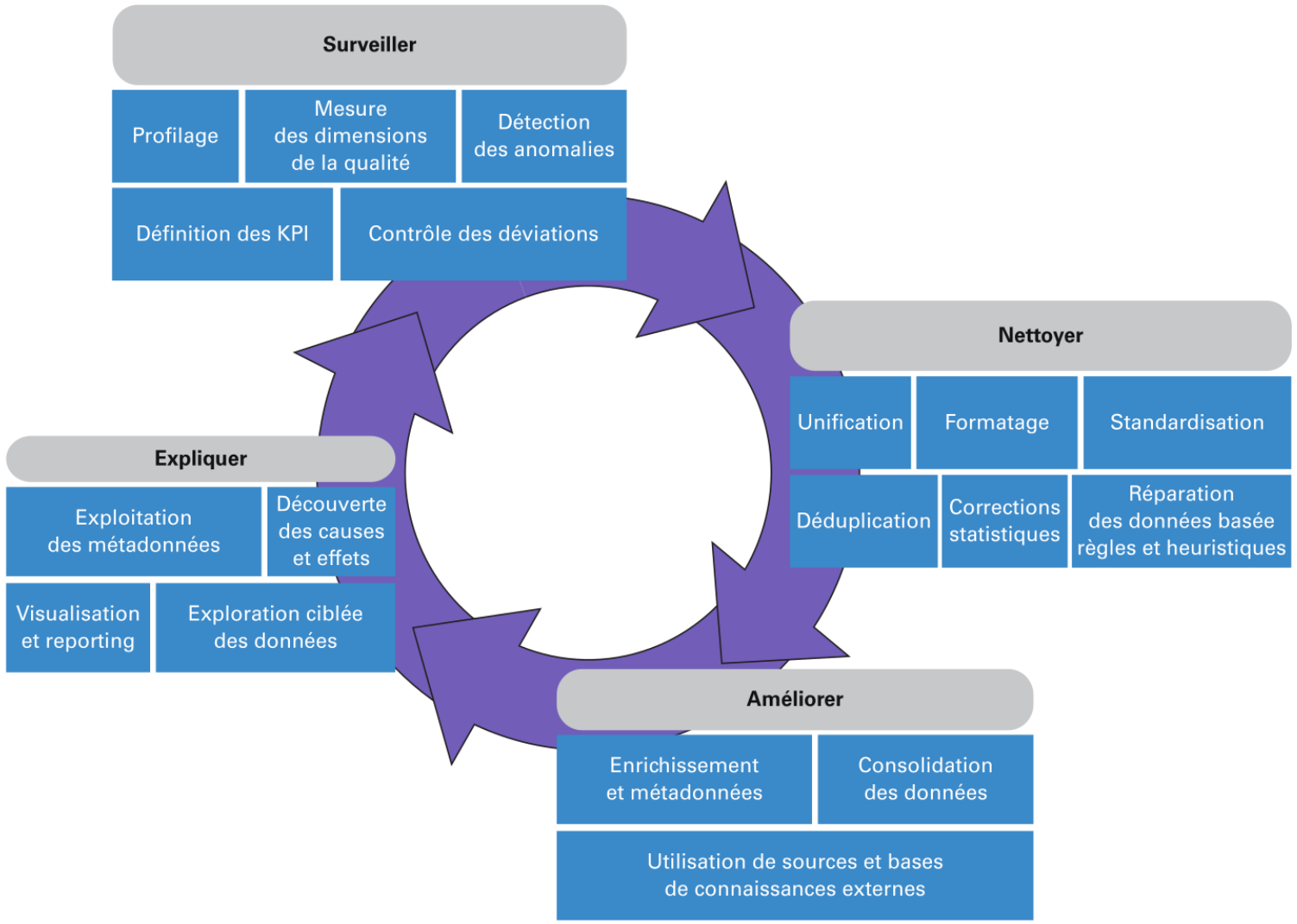
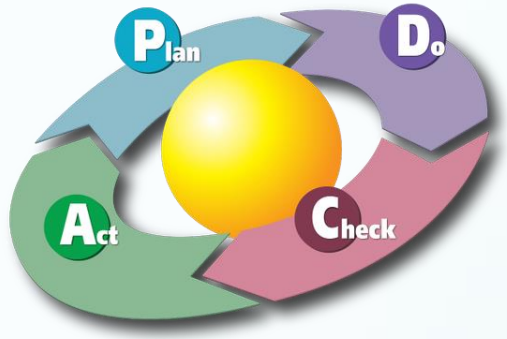


Figure 4 – Cercle vertueux pour la gestion de la qualité des données



Source: Qualité des données, Laure BERTI-ÉQUILLE, Techniques de l'Ingénieur, 2018, Réf: H3700 V2

Indicateurs (Traçabilité)

L'amélioration continue ne peut exister que si on peut la mesurer, il faut donc des indicateurs pour suivre le dispositif de maintien de la qualité des données

- des critères intrinsèques aux données elles-mêmes
- des critères de services liés à l'utilisation de ces données
- des critères de sécurité liés à l'ensemble du dispositif de gestion des données

Les métadonnées (Traçabilité)

Définition: est une donnée servant à définir ou décrire une autre donnée quel que soit son support (papier ou électronique)

- Pour la recherche les premières métadonnées sont celles qui sont saisies dans le cahier de laboratoire
- Les métadonnées sont utiles à l'interopérabilité
- Inconvénient / difficulté: Comment définir un standard ?

Le plus grand risque serait d'ignorer l'importance des métadonnées

Registre de métadonnées: Norme: ISO/CEI 11179

Intégration dans un processus de recherche

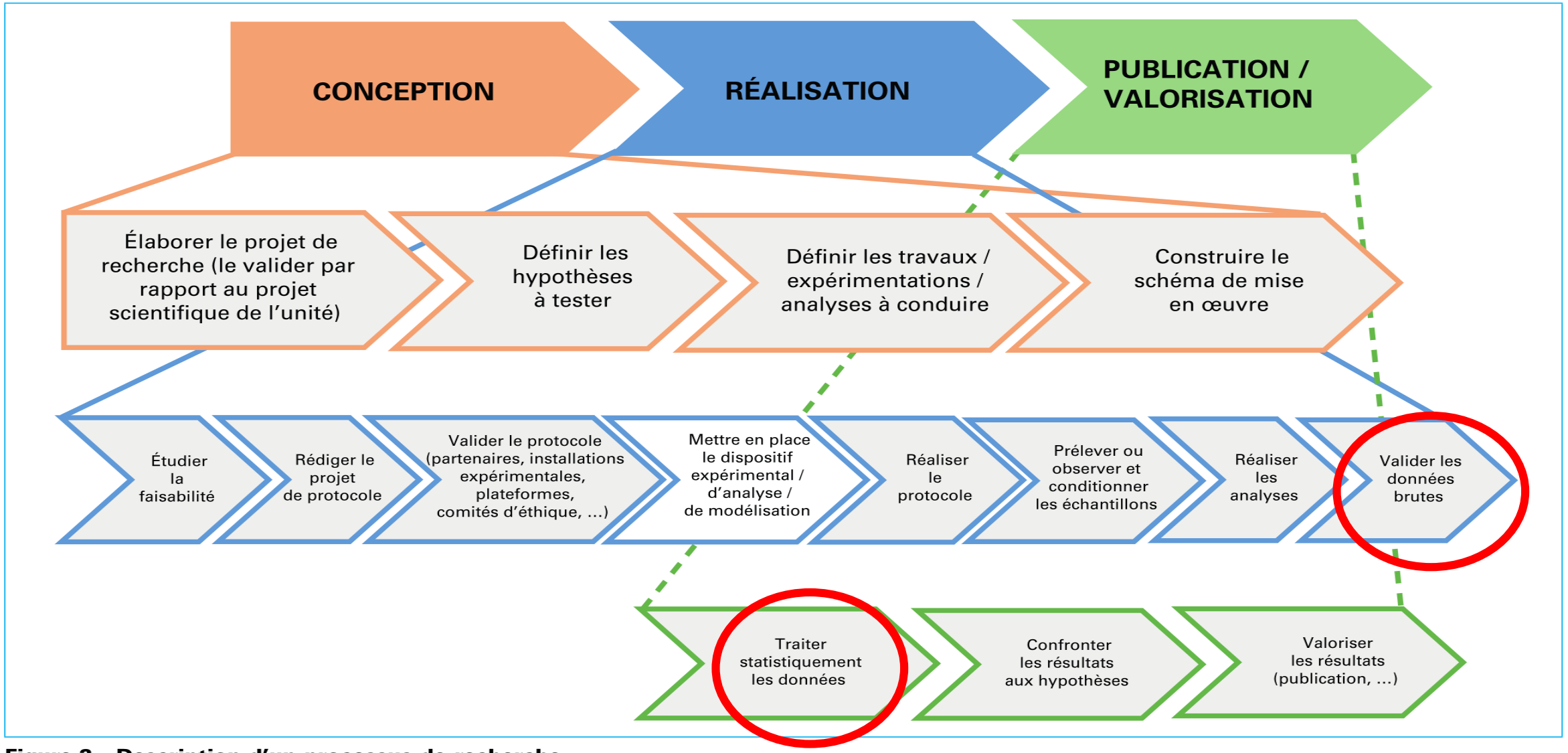


Figure 2 – Description d'un processus de recherche

Conclusion

Les principes de la qualité sont un moyen de faciliter la mise en place d'une gestion des données de qualité et ainsi de répondre aux différentes visions de la qualité des données.

- **Développer un nouveau modèle de gouvernance** :
  - Ensemble de pratiques qui contribuent à assurer la maîtrise du patrimoine des données à travers une démarche organisationnelle avec la définition de processus de prévention de la non-qualité, de remédiation des anomalies