Définir une approche qualité des données de recherche -

Frédéric de Lamotte, Véronique Stoll, Romain David, Carlo Maria Zwölf, Françoise Genova, Emilie Lerigoleur

Ouvrir la science !

Définir une approche qualité des données de recherche - Guide pratique

Comité pour la science ouverte

Collège Données de la recherche

Juillet 2025

Définir une approche qualité des données de recherche
Guide pratique

Frédéric de LAMOTTE – Pilotage du Collège Données de la recherche
INRAE

Véronique STOLL – Pilotage du Collège Données de la recherche
Observatoire de Paris - PSL

Romain DAVID – Pilotage du groupe de travail
ERINHA

Carlo Maria ZWÖLF – Pilotage du groupe de travail
Observatoire de Paris – PSL

Françoise GENOVA
CNRS

Emilie LERIGOLEUR
GEODE CNRS

Juillet 2025

DOI : 10.52949/86

Conception graphique : opixido

Sommaire

Objet du guide

Ce guide pratique vise à établir une approche qualité pour les données de recherche. Il propose un cadre méthodologique structuré permettant d'initier ou de renforcer une démarche qualité tout au long du cycle de vie des données. Construit à partir d'analyses de concepts, de pratiques et de recommandations, il offre un cadre de référence cohérent avec les principes FAIR (*Findable, Accessible, Interoperable, Reusable*).

Ses objectifs principaux sont :

- Présenter un ensemble de principes et de bonnes pratiques garantissant la fiabilité, la traçabilité et la réutilisabilité optimale des données de recherche ;
- Faciliter l'évaluation et le suivi de la qualité des données à chaque étape de leur cycle de vie.
- L'approche qualité proposée est organisée autour de six phases : préparation, collecte/production, traitement/analyse, préservation/stockage, partage/diffusion et archivage.

Le guide s'appuie sur une approche interdisciplinaire et ouverte, reposant sur :

- Une revue de la littérature scientifique sur la qualité des données ;
- Les travaux de la *Task Force FAIR Metrics and Data Quality* de l'European Open Science Cloud (EOSC), notamment le rapport *Towards a Data Quality Framework for EOSC*[1] ;
- Des études de cas pratiques.

La démarche est inscrite dans une dynamique d'amélioration continue inspirée du cycle de Deming (PDCA).

Audience cible

Ce vade-mecum s'adresse à l'ensemble des acteurs souhaitant mettre en place ou consolider une démarche qualité appliquée aux données de recherche, notamment :

- Les porteurs de projets responsables de la rédaction des plans de gestion de données (PGD) ;
- Les producteurs de données (collecte, expériences, observations, simulations) ;
- Les gestionnaires de données, chargés de la gouvernance et de la gestion opérationnelle ;
- Les curateurs, responsables de la préparation et de la diffusion des données ;
- Les archivistes, en charge de la préservation à long terme.

Il présuppose une acculturation préalable à la gestion des données et vise particulièrement les acteurs souhaitant structurer et renforcer leurs pratiques.

Structuration du guide

Le document fournit des outils opérationnels concrets pour la mise en œuvre d'une approche qualité :

- Une check-list qualité structurée en sept dimensions :
    - Caractérisation et organisation des données,
    - Définition des objectifs qualité,
    - Documentation (pertinence, précision, adéquation, couverture, cohérence, crédibilité, actualité),
    - Traçabilité et réutilisation,
    - Conformité réglementaire,
    - Interopérabilité technique,
    - Finalisation et évaluation globale ;
- Un référentiel de rôles et responsabilités pour clarifier les fonctions des différents acteurs ;
- Des recommandations opérationnelles détaillées pour chaque phase du cycle de vie ;
- Un exemple concret d'application de la check-list, facilitant son appropriation par les équipes ;

[1] Lacagnina, C., David, R., Nikiforova, A., Kuusniemi, M. E., Cappiello, C., Biehlmaier, O., Wright, L., Schubert, C., Bertino, A., Thiemann, H., & Dennis, R. (2023). TOWARDS A DATA QUALITY FRAMEWORK FOR EOSC (1.0.0). Zenodo. https://doi.org/10.5281/zenodo.7515816

- Des éléments réutilisables pour des évaluations externes (par ex. *CoreTrustSeal*) ou pour mesurer l'alignement avec les principes FAIR, en complément d'outils comme le *FAIR Data Maturity Model*.

Utilisation du guide

Le guide peut être exploité comme :

- Un support à la rédaction de plans de gestion de données ;
- Un outil d'aide à la gestion des métadonnées ;
- Une grille d'évaluation de l'avancement des actions qualité ;
- Une base pour l'obtention de certifications ou l'évaluation FAIR ;
- Un cadre d'auto-évaluation adaptable aux exigences institutionnelles, disciplinaires ou de financeurs.

En proposant une approche structurée et pragmatique, le guide permet d'identifier les exigences fondamentales ainsi que les obstacles récurrents, de mettre en avant des leviers concrets pour améliorer la qualité, de garantir la fiabilité, la traçabilité et la réutilisabilité optimale des données, et de s'adapter à différents contextes tout en maintenant un socle commun de bonnes pratiques.

Méthodologie

Après avoir clarifié les rôles et responsabilités des différents acteurs impliqués dans un projet de recherche, le document propose une démarche qualité structurée, accompagnée de recommandations opérationnelles pour chacune des phases du cycle de vie du projet, de la conception à la valorisation des données. S'inscrivant dans une dynamique d'amélioration continue — principe fondamental de toute approche qualité —, cette démarche est appuyée par une check-list permettant d'évaluer de manière systématique l'état d'avancement des actions mises en œuvre et d'identifier les axes d'amélioration à chaque itération. Un exemple concret vient illustrer l'utilisation de cet outil au travers d'un cas appliqué aux différentes étapes d'un projet de recherche, facilitant ainsi son appropriation par les équipes.

Par ailleurs, plusieurs éléments présentés dans le document peuvent être à ce stade réutilisés dans d'autres contextes, notamment dans le cadre d'évaluations menées en vue de l'obtention de certifications, telles que le *Core Trust Seal* (https://www.coretrustseal.org), ou pour apprécier le degré d'alignement avec les principes FAIR. À ce titre, le *FAIR Data Maturity Model*, élaboré par la *Research Data Alliance* (RDA), peut constituer un outil d'auto-évaluation complémentaire, adapté à des contextes disciplinaires variés[2]. Inversement, les efforts réalisés pour répondre aux exigences de ces évaluations externes contribuent à renforcer durablement la qualité des données produites ainsi que l'efficacité des processus internes de gestion des données.

[2] FAIR Data Maturity Model Working Group. *FAIR Data Maturity Model. Specification and Guidelines (1.0)*. Zenodo, (2020). https://doi.org/10.15497/rda00050

Mettre en place un processus qualité des données : acteurs et rôles

Le tableau ci-dessous présente le référentiel des rôles adoptés dans le cadre de ce guide, en s'appuyant notamment sur l'expérience des membres du Collège des Données. Il sert de base pour structurer les responsabilités et articuler les actions à mener selon les phases du cycle de vie des données.

Rôles et fonctions des acteurs d'un projet

Rôle	Description	Fonctions
Production de données (P)	À l'origine de la création des données	Concevoir et mettre en œuvre les protocoles de collecte de données
		Réaliser les expériences ou observations ou simulations pour obtenir les données primaires
		Documenter la méthodologie et les conditions de production des données
		Effectuer un premier contrôle qualité des données recueillies
Gestion de données[3](G)	Gestion et gouvernance opérationnelles des données	Définir et mettre en œuvre les politiques de gestion des données
		Organiser et structurer les bases de données
		Établir les standards de métadonnées et de documentation
		Gérer le stockage, la sauvegarde et la sécurité des données
		Assurer la conformité avec les réglementations en vigueur
		Faciliter l'accès et le partage des données au sein de l'équipe de recherche
Curation et diffusion des données (C)	Préparation des données pour la diffusion et la gestion de leur partage	Vérifier et enrichir les métadonnées
		Standardiser les formats de données pour faciliter l'interopérabilité
		Anonymiser les données sensibles si nécessaire
		Choisir les plateformes appropriées pour le partage des données
		Définir les licences et conditions d'utilisation des données, en accord avec les tutelles
		Mettre en place des identifiants pérennes (DOI) pour les jeux de données
		Promouvoir les données auprès de la communauté scientifique
Archivage (A)	Préservation à long terme des données.	Sélectionner les données à conserver sur le long terme
		Mettre en œuvre les stratégies de préservation numérique
		Gérer les migrations de formats pour assurer la pérennité des données
		Maintenir l'accessibilité des données archivées

[3] Dans un souci de simplification, la fonction de *gestionnaire des données* fusionne les rôles de *data manager* et *data steward*.

Dans le cadre d'un projet de recherche impliquant la production, la gestion et la diffusion de données, plusieurs rôles sont mobilisés à différentes étapes du cycle de vie des données[4]. Ces rôles, parfois assumés par une même personne ou répartis entre plusieurs membres de l'équipe projet, participent de manière complémentaire à la mise en œuvre d'une démarche qualité cohérente et efficiente.

Il est important de rappeler que, dans de nombreux projets de recherche — notamment les plus modestes —, une même personne peut assurer plusieurs fonctions liées à la gestion des données, le plus souvent le chercheur lui-même. Une certaine souplesse dans la répartition des tâches est donc nécessaire. Néanmoins, il demeure essentiel que l'ensemble des fonctions clés soient effectivement prises en charge afin de garantir une gestion rigoureuse et efficiente des données tout au long du projet.

Définir la qualité au cours des différentes phases d'un projet de recherche

Une approche qualité dans la gestion des données s'inscrit dans une démarche d'amélioration continue, fondée sur une approche itérative comparable au cycle de Deming[5]. Cette approche est particulièrement pertinente pour la mise en œuvre des principes FAIR, et s'applique à l'ensemble du cycle de vie des données, depuis la phase de préparation jusqu'à leur archivage[6]. Elle implique une coordination étroite entre les acteurs tout au long des étapes du projet, afin de garantir la fiabilité, l'interopérabilité et la réutilisabilité des données produites.

Les principales phases concernées sont les suivantes :

- Préparation (planification, sensibilisation, formation) ;
- Collecte ou production des données ;
- Traitement et analyse ;
- Préservation et stockage ;
- Partage, diffusion et documentation ;
- Archivage.

Phase 1 – Préparation : sensibilisation et planification

Cette phase initiale permet d'identifier les besoins, de prendre en compte les contraintes éventuelles, et de poser les fondations d'une gestion des données conforme aux principes FAIR. Elle est déterminante pour assurer la cohérence des étapes ultérieures et anticiper les exigences de diffusion, de documentation ou d'archivage.

***Acteurs mobilisés** : producteurs de données, gestionnaires, curateurs, responsables de diffusion.*

Les actions clés à mettre en œuvre à ce stade incluent :

- La définition des objectifs scientifiques et des critères de qualité des données ;
- L'identification des contraintes applicables (éthiques, techniques, juridiques) ainsi que des bénéfices attendus de la démarche qualité[7] ;
- L'élaboration d'un plan de gestion des données (PGD), couvrant l'ensemble du cycle de vie des données ;
- La définition des conditions de réutilisation des données, incluant une réflexion sur les licences à appliquer ;
- La sélection des standards, protocoles et métadonnées nécessaires à la structuration et à la réutilisabilité des données ;

---

[4] Ray, Joyce M., ed. Research Data Management: Practical Strategies for Information Professionals. DGO-Digital original. Purdue University Press, 2014. https://doi.org/10.2307/j.ctt6wq34t

[5] PDCA – Plan (Planifier) : Définir les objectifs et préparer le plan d'action ; Do (Faire) : Mettre en œuvre le plan d'action ; Check (Vérifier) : Contrôler les résultats obtenus ; Act (Agir) : Ajuster et améliorer le processus en fonction des résultats.

Sokovic, M., Pavletic, D., Pipan, K. K. "Quality improvement methodologies–PDCA cycle, RADAR matrix, DMAIC and DFSS." *Journal of Achievements in Materials and Manufacturing Engineering* 43, no. 1 (2010): 476-483. https://api.semanticscholar.org/CorpusID:16093281

[6] David, R., Mabile, L., Specht, A., et al. "RDA – SHAring Reward and Credit (SHARC) Interest Group R.D.A. FAIRness Literacy: The Achilles' Heel of Applying FAIR Principles." *Data Sci J* 19, no. 1 (2020): 32. http://doi.org/10.5334/dsj-2020-032

[7] David, R., Mabile, L., Specht, A., et al. "RDA – SHAring Reward and Credit (SHARC) Interest Group R.D.A. FAIRness Literacy: The Achilles' Heel of Applying FAIR Principles." *Data Sci J* 19, no. 1 (2020): 32. http://doi.org/10.5334/dsj-2020-032

- L'identification de métriques permettant de mesurer le niveau de conformité des données et métadonnées aux principes FAIR ;
- La répartition claire des rôles et des responsabilités entre les parties prenantes, en fonction des différentes phases du projet.

Phase 2 – Collecte, acquisition ou production des données

***Acteurs mobilisés** : producteurs de données, gestionnaires de données*

Cette phase opérationnelle repose sur l'application rigoureuse des protocoles définis lors de la préparation du projet. Elle vise à garantir la qualité, la traçabilité et la sécurité des données dès leur création ou leur acquisition. Les mécanismes de contrôle doivent être adaptés en continu à partir des anomalies détectées, dans une logique d'amélioration progressive.

Les actions à mettre en œuvre à ce stade comprennent :

- L'application stricte des protocoles de collecte préalablement définis, conformément aux standards établis ;
- La mise en œuvre des dispositifs de contrôle qualité, incluant :
  - L'activation de contrôles à la saisie ou à l'importation des données,
  - La validation des données entrantes,
  - La détection précoce des erreurs ou incohérences,
  - L'enrichissement des données, lorsque cela est possible, avec des indicateurs de précision (tels que des marqueurs de qualité ou "flags"),
  - L'application des mesures nécessaires à la protection des données sensibles ou à caractère personnel ;
- La production ou le renseignement des premières métadonnées descriptives, selon des formats riches, standardisés et interopérables ;
- La mise en place de mécanismes d'identification unique pour chaque donnée collectée (par des PID internes ou externes), afin d'en assurer la traçabilité et la gestion pérenne.

Phase 3 – Traitement, analyse et documentation des données

Cette phase constitue une étape clé pour assurer la qualité, la cohérence et la réutilisabilité des données. Elle repose sur des processus de vérification, de curation et de documentation, qui doivent être ajustés en continu selon les difficultés rencontrées ou les améliorations identifiées. Une documentation rigoureuse des traitements est indispensable pour garantir l'interopérabilité, la reproductibilité scientifique et la conformité aux principes FAIR.

***Acteurs mobilisés, en appui des équipes de recherche** : gestionnaires de données, assistés des producteurs (pour les opérations proches de la production) et des curateurs (notamment en vue de la diffusion).*

Les actions recommandées à ce stade incluent :

- La vérification de la cohérence et de l'intégrité des données, à l'aide de techniques appropriées et d'outils spécialisés (par exemple : OpenRefine[8]) ;
- L'identification des anomalies ou incohérences via des opérations de profilage, pouvant donner lieu à la production de rapports spécifiques ;
- La documentation systématique des modifications et corrections apportées aux données ;
- La conversion des données dans des formats ouverts, pérennes et largement diffusés. En l'absence de format de référence reconnu, il peut être pertinent de proposer plusieurs formats ;
- L'enrichissement des jeux de données à l'aide d'identifiants persistants (URI), de vocabulaires contrôlés et d'ontologies validées par la communauté disciplinaire concernée ;
- La standardisation ou l'harmonisation des vocabulaires utilisés, afin de favoriser l'interopérabilité sémantique ;
- La documentation détaillée des méthodes et processus mis en œuvre, en s'appuyant sur des standards reconnus (par exemple ISO) et des ontologies adaptées (ex. : Prov-O pour la provenance[9]) ;

[8] https://openrefine.org

[9] https://www.w3.org/TR/prov-o/

- L’adoption, lorsque cela est possible, d’un système de *versionning* adapté à tous les objets numériques manipulés (données, scripts, documents) ;
- La fourniture d’exemples d’analyses et d’interprétations illustrant les méthodes utilisées, en précisant les limites identifiées quant à la qualité des données.

Phase 4 – Préservation et stockage

La phase de préservation vise à garantir la conservation fiable, sécurisée et durable des données et de leurs métadonnées, dans le respect de l’intégrité des contenus produits au cours des étapes antérieures. Les procédures mises en œuvre doivent éviter toute altération des fichiers ou des informations associées, en s’appuyant sur des environnements techniques adaptés et éprouvés.

***Acteurs mobilisés** : gestionnaires de données, en lien avec les curateurs responsables de la diffusion.*

Les actions à engager comprennent :

- Le stockage des données et des métadonnées dans un environnement sécurisé, techniquement adapté aux spécificités des données concernées ;
- La mise en place de mécanismes de fiabilité et de redondance (sauvegardes, réplications), garantissant la pérennité des contenus ;
- L'application de mesures de protection adaptées pour les données sensibles, avec des contrôles réguliers de leur efficacité (tests d’accès, de récupération ou d’intégrité)[10] ;
- La documentation exhaustive des procédures de stockage, incluant le cas échéant : l’organisation des fichiers, les structures d’accès, les jeux de métadonnées, les catalogues ou index utilisés, ainsi que les vocabulaires ou dictionnaires de référence. Cette documentation doit être maintenue à jour afin de garantir une compréhension et une exploitation durables des données archivées.

Phase 5 – Partage et diffusion

Cette phase vise à assurer une diffusion optimisée des données, dans des conditions garantissant leur accessibilité, leur traçabilité et leur réutilisation. L’objectif est de maximiser la visibilité des jeux de données et de favoriser leur exploitation par d’autres équipes, dans le respect des exigences éthiques et juridiques.

***Acteurs mobilisés :** curateurs, responsables de la diffusion.*

Les principales actions à conduire sont les suivantes :

- Attribution d’identifiants pérennes (PID) tels que DOI ou Handle, permettant une citation stable et une identification univoque des jeux de données ;
- Définition explicite des licences d’accès et de réutilisation, en cohérence avec les choix documentés lors de la phase de planification (cf. phase 4.1) ;
- Mise en accessibilité des données, de manière ouverte et libre dès que cela est possible, ou via une politique AAA (Authentication, Authorization, Accounting) adaptée en cas de restrictions[11] ;
- Publication des métadonnées descriptives accompagnant les données, y compris les conditions d’accès et de réutilisation. Des protocoles appropriés doivent être mis en œuvre pour les données sensibles ;
- Enrichissement régulier des métadonnées, en phase avec l’évolution des standards communautaires et des usages. Cette dynamique permet d’augmenter le potentiel de réutilisation des données ;
- Utilisation d’entrepôts certifiés ou d’infrastructures disciplinaires reconnues pour garantir la sécurité, l’interopérabilité et la pérennité du dépôt[12] ;
- Mise en place de mécanismes normalisés de citation des données, assurant leur indexation automatique et leur rattachement aux auteurs (via ORCID), aux structures (via ROR), et aux réseaux de publication (via Crossref, Scholix) ;

[10] https://www.cnil.fr/sites/cnil/files/2024-03/cnil_guide_securite_personnelle_2024.pdf

[11] Bressoud, T., White, D. "Authentication and Authorization." In *Introduction to Data Systems*, Springer, (2020). DOI: https://doi.org/10.1007/978-3-030-54371-6_24

[12] Frédéric de Lamotte, Véronique Stoll, Cécile Arènes, Marie-Emilia Herbet, Stéphane Debard, et al.. Sélectionner un entrepôt thématique de confiance pour le dépôt de données : méthodologie et analyse de l'offre existante. Comité pour la Science Ouverte. 2024. DOI : 10.52949/52

- Documentation transparente des processus qualité, y compris les éventuelles limites ou biais identifiés au cours du cycle de vie des données ;
- Suivi de l'accès et de la réutilisation, via des outils de traçabilité ou d'analyse d'usage, permettant d'évaluer l'impact des données diffusées ;
- Mise en œuvre de mécanismes d'alerte post-publication, afin de notifier les utilisateurs en cas de découverte ultérieure de problèmes de qualité affectant les jeux de données ;
- Indexation optimisée des données, à travers des métadonnées interopérables et des systèmes automatiques respectant les exigences de découvrabilité (critère F4[1,2]).

Phase 6 – Archivage

Phase ultime du cycle de vie des données, l'archivage vise à assurer la pérennité des données ainsi que celle des pratiques qualité qui leur sont associées. Elle constitue une étape déterminante pour garantir la réutilisabilité, l'intégrité et la lisibilité des données sur le long terme.

***Acteurs concernés** : archivistes de données, en lien éventuel avec les responsables de la gestion des données, les référents techniques et les services d'appui à la recherche.*

A) Pérennisation initiale

- Mettre en œuvre des processus visant à maintenir la qualité des données dans le temps (contrôle de validité, procédures de gestion de la qualité, des accès, des licences, etc.) ;
- Anticiper les conditions de réutilisation future des données, notamment en prévoyant des mécanismes de mise à jour, d'enrichissement ou d'annotation à partir de nouvelles ontologies ou standards disciplinaires ;
- Associer aux jeux de données une documentation complète, stable et intelligible, garantissant leur compréhension et leur exploitation future ;
- Prévoir, le cas échéant, des stratégies de migration pour répondre aux évolutions technologiques et assurer la continuité d'accès.

B) Consolidation et maintenance

- Réaliser des vérifications régulières de l'intégrité des données archivées ;
- Veiller à la lisibilité, à la pérennité et à l'interopérabilité des formats de conservation ;
- Maintenir les conditions de réexécution et de reproduction des résultats associés aux données (notamment dans le cas de données liées à un code ou un environnement logiciel) ;
- Suivre l'évolution des modalités de stockage et adapter les infrastructures pour garantir une accessibilité durable ;
- Garantir un accès pérenne aux données archivées, dans le respect des contraintes légales, réglementaires et éthiques, y compris celles relatives aux données sensibles ;
- Engager une amélioration continue des pratiques d'archivage et de diffusion, en intégrant les retours d'expérience et les avancées technologiques.

Une check-list « qualité des données »

La check-list ci-dessous a été élaborée en s'appuyant sur les travaux de la Task Force *FAIR Metrics and Data Quality* de l'European Open Science Cloud (EOSC). Elle reprend et adapte les éléments issus du rapport *Towards a Data Quality Framework for EOSC* [13], afin de proposer un outil opérationnel et structurant pour accompagner les porteurs de projets dans l'évaluation et le suivi de la qualité des données tout au long de leur cycle de vie.

Caractérisation des données : organisation et leviers d'amélioration

Une première étape essentielle consiste à caractériser les données générées ou utilisées dans le cadre du projet, afin d'identifier les exigences associées à leur qualité. Cette caractérisation permet de poser les bases d'une organisation structurée et de définir les actions à engager pour atteindre une amélioration concrète et mesurable de la qualité des données.

Ce point d'étape sur l'organisation vise à clarifier la répartition des rôles, les responsabilités associées, les ressources mobilisées ainsi que les modalités de coordination entre les parties prenantes. Il permet également d'identifier les outils, référentiels et procédures déjà en place, et d'évaluer dans quelle mesure ils peuvent être renforcés ou adaptés pour soutenir la démarche qualité.

- **1.1** – Évaluer le caractère des données (structurées, semi-structurées non-structurées) ;
- **1.2** – Identifier l'origine des données (mesures, manuscrits…) ;
- **1.3** – Évaluer les éventuelles contraintes de temps d'accès sur les données (disques rapides vs. stockage sur bandes).
- **1.4** – Évaluer l'éventuel caractère sensible des données (haut, moyen, bas risque)[14] ;
- **1.5** – Établir l'ensemble des métadonnées minimales nécessaires pour caractériser les données. Cet ensemble doit garantir une description suffisante pour permettre une réutilisation pertinente des données, dans des conditions conformes aux principes FAIR, tout en respectant, lorsque cela est applicable, les standards et référentiels reconnus par la communauté scientifique de référence.

[13] Lacagnina, C., David, R., Nikiforova, A., et al. *Towards a Data Quality Framework for EOSC (1.0.0)*. Zenodo, (2023). https://doi.org/10.5281/zenodo.7515816

[14] https://cyber.gouv.fr/sites/default/files/document/anssi_essentiels_données_traitements_sensibles_1.0.pdf.pdf *Arrêté du 26 avril 2024 modifiant l'arrêté du 11 juin 2018 portant approbation du référentiel d'accréditation des organismes de certification et du référentiel de certification pour l'hébergement de données de santé à caractère personnel* (Journal Officiel de la République Française, 16 May 2024), no. 113, p. 10. NOR: TSSD2325104A

Données structurées : Les données structurées sont des informations organisées selon un format prédéfini et standardisé, généralement stockées dans des bases de données relationnelles. Elles sont facilement interprétables et exploitables par des machines, grâce à un schéma structuré comportant des champs clairement définis.

Données semi-structurées : Les données semi-structurées se situent entre les données structurées et non structurées. Bien qu'elles ne se conforment pas à un modèle relationnel strict, elles intègrent des balises ou marqueurs permettant de distinguer les différents éléments. Des exemples courants incluent les formats JSON et XML.

Données non structurées (dé-structurées) : Les données non structurées sont des informations qui ne se conforment pas un format ou un modèle prédéfini. Elles peuvent inclure du texte libre, des images, des vidéos, des fichiers audio, ...

Données sensibles : Les données sensibles sont des informations qui nécessitent une protection spéciale en raison de leur nature délicate et du risque potentiel qu'elles représentent si elles sont divulguées (ex, localisation des espèces menacées, secret militaire et commercial...).

Définition des objectifs

Avant d'engager une démarche qualité autour des données, il est essentiel de clarifier les objectifs poursuivis. Cette étape préliminaire permet d'orienter les actions à mettre en œuvre, en cohérence avec les usages visés et les exigences des communautés concernées.

2.1 – Déterminer le niveau d'engagement qualité souhaité : Il convient d'identifier si l'objectif est de mettre en place un contrôle qualité, de type a posteriori (détection d'éventuelles anomalies ou erreurs), ou de garantir la qualité, en intégrant en amont un ensemble de bonnes pratiques permettant de prévenir les défauts ;

2.2 – Définir les garanties attendues pour assurer la réutilisabilité des données.

2.3 – Les données sont-elles disponibles lorsque les utilisateurs en ont besoin ?

2.4 – Utilisez-vous des procédures validées par votre communauté pour garantir la fiabilité des données ?

2.5 – Les utilisateurs disposent-ils des informations nécessaires pour : i) évaluer le contenu et la qualité des données ; ii) comprendre les modalités de leur réutilisation dans un contexte donné ?

2.6 – Évaluer la préparation des données pour des usages spécifiques**.**

2.7 – Les besoins pour lesquels les données sont produites sont-ils clairement exprimés ?

2.8 – Le périmètre de validité des données est-il fourni, avec une identification explicite des limites, des points forts et des éventuelles zones d'incertitude ?

Documentation

Afin d'évaluer et d'assurer la qualité des données, il est indispensable de fournir une documentation explicite attestant de leur conformité aux phénomènes qu'elles représentent sous forme numérique**.** Cette documentation permet d'apprécier dans quelle mesure les données répondent aux exigences attendues selon plusieurs dimensions fondamentales :

3.1 – Pertinence : Les données doivent comporter une ou plusieurs valeurs représentatives de chaque phénomène observé, conformément aux objectifs du projet.

3.2 – Précision : Le niveau de granularité des données doit rendre compte de manière fiable des variations observées.

3.3 – Adéquation : Les attributs et les valeurs doivent correspondre à la réalité des phénomènes mesurés ou observés.

3.4 – Couverture : L'ensemble des attributs attendus doit être renseigné, afin d'assurer une complétude satisfaisante des données ;

3.5 – Cohérence : Les données doivent être exemptes de contradictions internes et compatibles avec d'autres sources ou jeux de données connexes ;

3.6 – Crédibilité : Les données doivent être considérées comme fiables, justifiables et dignes de confiance par les utilisateurs finaux, en particulier lorsqu'elles sont diffusées ou partagées ;

3.7 – Actualité : Les attributs doivent être maintenus à état à jour, et demeurer valides au regard de l'évolution des phénomènes décrits ou des usages envisagés ;

Traçabilité et réutilisation

La traçabilité et la réutilisabilité sont des dimensions clés pour garantir la qualité, la transparence et la pérennité de l'usage des données. Elles nécessitent la mise à disposition d'un ensemble d'informations claires et structurées, notamment :

4.1 – Fournir les modalités pour la citation des données, afin de garantir leur traçabilité et de valoriser leur production.

4.2 – Définir explicitement les conditions de réutilisation des données (licences applicables) et préciser les dispositions en matière de propriété intellectuelle.

4.3 – Documenter les transformations subies par les données, ainsi que leur provenance, idéalement en s'appuyant sur des standards reconnus, tels que *Prov-O*.

4.4 – Attester que les données ont été produites, gérées et diffusées conformément aux standards en vigueur dans la ou les disciplines concernées.

4.5 – Fournir des contacts à jour pour questions ou signalements sur les données.

Conformité réglementaire

Le respect du cadre juridique et réglementaire constitue un fondement indispensable de la démarche qualité, en particulier lorsqu'il s'agit de données sensibles ou à caractère personnel :

5.1 – Garantir que l'accès aux données est restreint aux seuls utilisateurs autorisés.

5.2 – Veiller à la conformité avec le Règlement général sur la protection des données (RGPD).

5.3 – Respecter tous les accords de confidentialité, droits de propriété intellectuelle et obligations contractuelles.

5.4 – Appliquer les politiques de sécurité en vigueur, notamment celles relatives aux données à caractère sensible.

5.5 – Fournir un guide utilisateur complet décrivant :

- la structure des données et leurs différentes versions ;
- la finalité de leur production et les procédures associées (collecte, traitement, transformation) ;
- les éléments de documentation attestant de leur qualité et de leur conformité ;
- des exemples d'utilisation optimale ;
- les logiciels recommandés pour les manipuler, accompagnés, si possible, de sources accessibles et pérennes.

Interopérabilité technique

L'interopérabilité technique garantit que les données peuvent circuler, être comprises et réutilisées efficacement dans des environnements variés :

6.1 – S'assurer que les données sont lisibles et interprétables par des tiers (usage de conventions claires, de symboles partagés, de formats standards), et que des procédures d'import/export sont disponibles depuis les environnements de stockage.

6.2 – Utiliser, dans la mesure du possible, des formats ouverts et normalisés, en adéquation avec les normes et bonnes pratiques de la communauté scientifique concernée.

6.3 – Associer des métadonnées structurées aux données et, si possible, des vocabulaires contrôlés conformes aux standards de référence.

6.4 – Veiller à la portabilité des données, en garantissant qu'elles peuvent être déplacées ou dupliquées sans altération, et en maintenant l'intégrité de leurs métadonnées et identifiants pérennes.

6.5 – Mettre en œuvre des mécanismes de protection contre la perte de données, y compris la traçabilité des copies secondaires et des identifiants associés (par exemple via l'utilisation d'identifiants secondaires persistants[15]).

Finaliser la démarche qualité

Les éléments suivants visent à finaliser l'évaluation de la démarche qualité mise en œuvre et à apprécier son articulation avec les standards reconnus et les principes FAIR :

7.1 – Indiquer si des contrôles qualité spécifiques ont été réalisés, en lien par exemple avec des normes internationales (type ISO) ou dans le cadre d'une préparation à une certification (telle que le CoreTrustSeal). Préciser si ces éléments sont documentés, visibles et accessibles pour les utilisateurs potentiels des données ;

7.2 – Évaluer le niveau global de conformité avec les principes FAIR[16] :

- Combien de principes sont effectivement pris en compte ?
- À quel niveau de détail et de rigueur chaque principe est-il appliqué ?
- Quel est l'impact attendu de cette mise en œuvre sur la réutilisabilité des données à court, moyen et long terme ?

[15] Kahn, R., Wilensky, R. "A framework for distributed digital object services." *International Journal on Digital Libraries* 6, no. 2 (2006): 115-123. DOI : https://www.doi.org/10.1007/s00799-005-0128-x

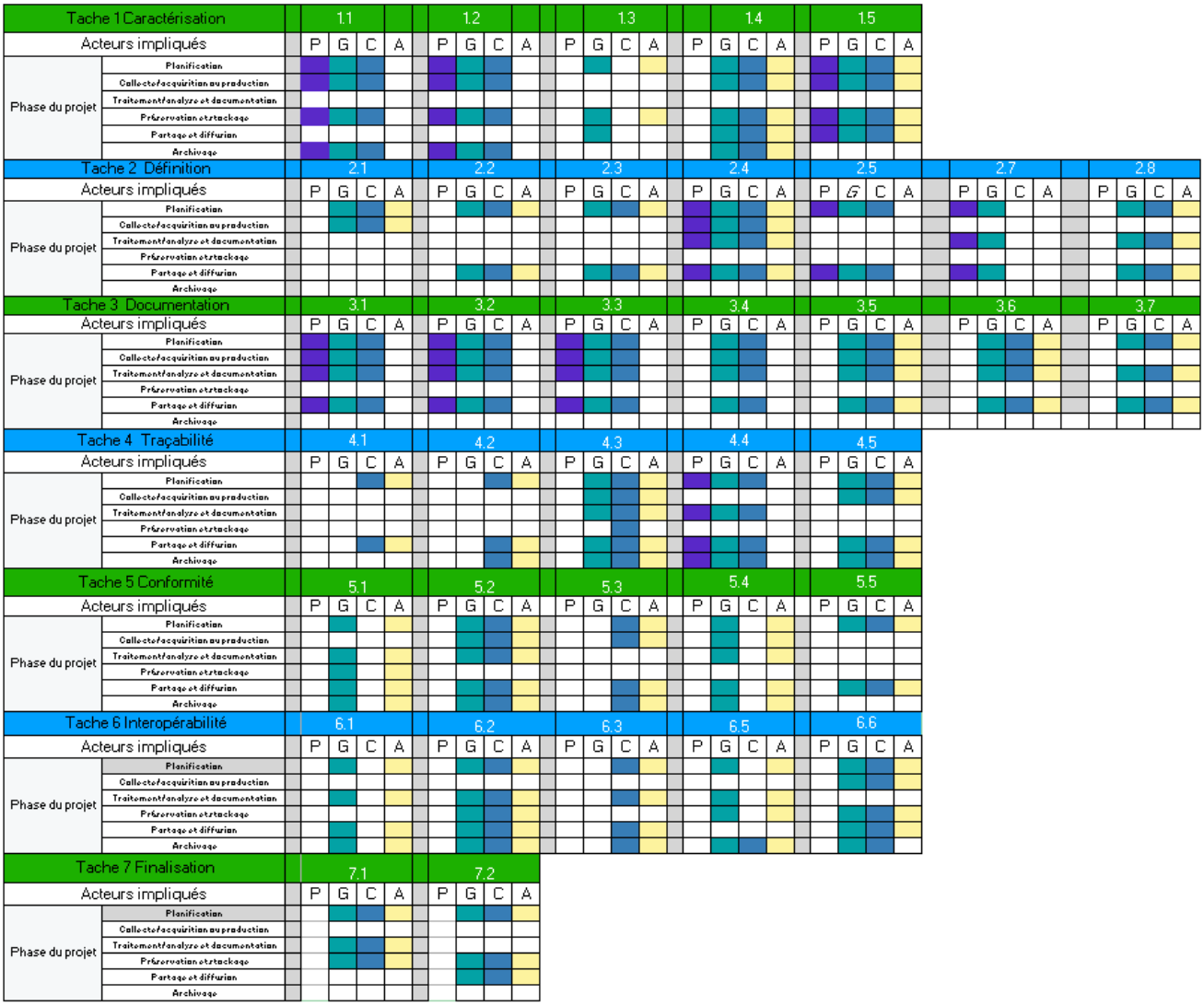
[16] Gaignard, A., Rosnet, T., De Lamotte, F. *et al.* FAIR-Checker: supporting digital resource findability and reuse with Knowledge Graphs and Semantic Web standards. *J Biomed Semant* 14, 7 (2023). https://doi.org/10.1186/s13326-023-00289-5

Annexe 1 : Dérouler la check-list au cours des différentes phases de vie d'un projet

Conçu par les auteurs du présent document, le tableau ci-après illustre l'application de la check-list qualité des données, telle que détaillée *supra*. Une même question peut être mobilisée à plusieurs reprises, à différentes phases, selon la nature du projet et les exigences associées.

P : production des données
G : gestion des données
C : curation des données
A : archivage

Une version éditable de ce tableau sous format excel se trouve en fichier annexe du présent document



Annexe 2 : Exemples d'application des critères de qualité des données

Cas d'usage 1 : Étude du bien-être des étudiants sous forme de questionnaire

Pertinence
Les questions doivent porter sur des dimensions directement liées au bien-être étudiant (stress, satisfaction, qualité du sommeil, etc.), et non sur des sujets non pertinents (par exemple, la marque de leur téléphone).

Précision
Les modalités de réponse doivent permettre de saisir des nuances (par exemple, une échelle de 1 à 10 pour évaluer le niveau de stress, plutôt qu'un simple oui/non), en particulier autour des valeurs critiques.

Adéquation
Les instruments de mesure (questionnaires, capteurs, etc.) doivent réellement évaluer le phénomène visé. Ainsi, une question sur le stress doit mesurer le stress, et non un autre concept.

Couverture
Tous les domaines pertinents sont renseignés : aucun attribut attendu (ex. qualité du sommeil) ne doit rester vide.

Cohérence
Les réponses doivent être logiquement compatibles (par exemple, il est incohérent de déclarer un sommeil optimal et un niveau de fatigue extrême simultanément).

Crédibilité
Les données doivent paraître plausibles pour l'utilisateur. Par exemple, un étudiant affirmant dormir zéro heure par nuit tout en se disant « en pleine forme » remet en cause la fiabilité des réponses.

Actualité
Les informations collectées doivent refléter la situation présente des étudiants, et non un état passé datant de plusieurs mois.

Cas d'usage 2 : Vérification expérimentale de la loi d'Ohm

Pertinence
On mesure la résistance, l'intensité du courant et la différence de potentiel, seules grandeurs pertinentes pour confirmer la loi d'Ohm.

Précision
On choisit un calibre de multimètre adapté pour distinguer des variations fines (par exemple, relever 10,00 Ω plutôt que 10 Ω).

Adéquation
Le multimètre doit être correctement étalonné et l'influence de la température ambiante prise en compte, car la résistance varie en fonction de la température.

Couverture
Les trois grandeurs (tension, courant, résistance) sont mesurées pour permettre un croisement des résultats et vérifier la cohérence globale.

Cohérence
Les mesures ne doivent pas se contredire : un courant de 0 A en présence d'une tension, ou une résistance négative, traduisent une incohérence.

Crédibilité
Une valeur aberrante (par exemple, –5 Ω pour un conducteur en cuivre) remettrait en cause la fiabilité de l'expérience.

Actualité
Les lectures sont effectuées en temps réel sur l'appareil de mesure. Aucune valeur préenregistrée ou obsolète ne doit être utilisée.

Bibliographie indicative

Wilkinson, M. D., Dumontier, M., Aalbersberg, IjJ., et al. "The FAIR Guiding Principles for scientific data management and stewardship." Sci Data 3 (2016): 160018. https://doi.org/10.1038/sdata.2016.18.

FAIR Data Maturity Model Working Group. FAIR Data Maturity Model. Specification and Guidelines (1.0). Zenodo, (2020). https://doi.org/10.15497/rda00050.

Ray, Joyce M., ed. Research Data Management: Practical Strategies for Information Professionals. DGO-Digital original. Purdue University Press, 2014. https://doi.org/10.2307/j.ctt6wq34t .

Sokovic, M., Pavletic, D., Pipan, K. K. "Quality improvement methodologies–PDCA cycle, RADAR matrix, DMAIC and DFSS." Journal of Achievements in Materials and Manufacturing Engineering 43, no. 1 (2010): 476-483. https://api.semanticscholar.org/CorpusID:16093281

David, R., Mabile, L., Specht, A., et al. "RDA – SHAring Reward and Credit (SHARC) Interest Group R.D.A. FAIRness Literacy: The Achilles' Heel of Applying FAIR Principles." Data Sci J 19, no. 1 (2020): 32. http://doi.org/10.5334/dsj-2020-032.

Wittenburg, P., Hellström, M., Zwölf, C.-M., et al. Persistent Identifiers: Consolidated Assertions. Status of November, 2017. Zenodo, (2017). https://doi.org/10.5281/zenodo.1116189.

Bressoud, T., White, D. "Authentication and Authorization." In Introduction to Data Systems, Springer, (2020). DOI: https://doi.org/10.1007/978-3-030-54371-6_24.

Lacagnina, C., David, R., Nikiforova, A., et al. Towards a Data Quality Framework for EOSC (1.0.0). Zenodo, (2023). https://doi.org/10.5281/zenodo.7515816.

Arrêté du 26 avril 2024 modifiant l'arrêté du 11 juin 2018 portant approbation du référentiel d'accréditation des organismes de certification et du référentiel de certification pour l'hébergement de données de santé à caractère personnel (Journal Officiel de la République Française, 16 May 2024), no. 113, p. 10. NOR: TSSD2325104A.

Kahn, R., Wilensky, R. "A framework for distributed digital object services." International Journal on Digital Libraries 6, no. 2 (2006): 115-123. DOI : https://www.doi.org/10.1007/s00799-005-0128-x.

ISO/IEC 2382:2015, "Information technology - Vocabulary'' -https://www.iso.org/standard/63598.html, https://www.iso.org/obp/ui/en/#iso:std:iso-iec:2382:ed-1:v2:en