

DOSSIER DE VALIDATION

Ingénieur en science des données

spécialisé en infrastructure data

pour le Titre RNCP 39586

Bloc 3: Élaborer et piloter un projet Data

| Nom Prénom | CORIN Gaëtan |
| --- | --- |
| Nom Prénom du tuteur | MOULIN Alexis |
| Niveau visé | RNCP 7 |
| Date de la soutenance | Septembre 2025 |
| Lieu de la soutenance | Toulouse |

Table des matières

[**C3.1.1 : Définir les objectifs à atteindre et le périmètre du projet, en analysant les contraintes techniques et réglementaires, en étudiant le contexte et les enjeux afin de dimensionner le projet en termes de délai et budget. 3**](#_jy30bt2p9iwd)

[**C3.1.2 : Dimensionner le projet en évaluant la charge de travail et les ressources nécessaires (humaines, matérielles) au regard des exigences attendues et des contraintes préalablement définies afin d’estimer le temps et le budget nécessaires à la faisabilité du projet. 3**](#_b3ft6usja70a)

[**C3.1.3 : Rédiger la documentation projet, en identifiant les parties prenantes concernées, en prenant en compte l’ensemble des caractéristiques du projet, afin de clarifier et formaliser les attendus. 3**](#_xu4v09r5mxto)

[**C3.2.1 : Planifier l’exécution du projet en organisant la répartition et l'ordonnancement des activités, le planning prévisionnel de réalisation et les ressources nécessaires à son exécution, en prenant en considération les personnes en situation de handicap afin de suivre les différentes phases du projet. 3**](#_llo8lkahs3yt)

[**C3.2.2 : Suivre l’avancement du projet en mettant en place un outil de suivi (logiciel de suivi, tableau de bord), en définissant les indicateurs (qualitatifs et/ou quantitatifs) pour chaque jalon défini dans le planning, en réalisant des reportings et des comptes rendus de réunion afin d’anticiper les aléas éventuels. 3**](#_ueljvy826n95)

[**C3.3.1 : Évaluer les besoins en compétences de l’équipe projet, en collaborant avec le service Ressources Humaines, en établissant un plan de développement des compétences et en orientant les membres de l’équipe vers des formations adaptées, afin de renforcer l’équipe responsable de mener à bien le projet DATA. 4**](#_hrv3uxlrcrci)

[**C3.3.2 : Piloter l’équipe projet en affectant les missions à réaliser, en prenant en compte les spécificités des membres de l’équipe, en intégrant les spécificités d’un contexte multiculturel, international, en utilisant les différentes techniques de communication et d’animation managériale pour favoriser le bon fonctionnement de l'équipe. 4**](#_iouoshj55hut)

[**C3.3.3 : Procéder aux arbitrages et aux réajustements nécessaires à partir de l’analyse des écarts entre le prévisionnel et l’état du projet à date, en utilisant des outils d’aide à la décision (ex : logigramme) afin de garantir la performance du projet dans le respect des objectifs de qualité, coûts et délai. 4**](#_wu78rnjl1i7c)

[**C3.4.1 : Mettre en place un système de veille technologique et réglementaire en matière de science des données et d’Intelligence Artificielle à l’aide de recherches documentaires, de plateformes de partage, de webinars afin d’être alerté des évolutions qui impacteraient les pratiques métier. 4**](#_wbnmobkfxy4j)

[**C3.4.2 : Intégrer dans ses pratiques métier les enjeux en termes de données responsables, de responsabilité sociétale et environnementale (RSE), de sécurité, d’éthique et de confidentialité des données en se tenant informé des évolutions du cadre juridique, à travers une recherche documentaire ou en étant accompagné par des juristes afin d’agir dans le respect de la législation. 4**](#_z17r6lt5s45)

# C3.1.1 : Définir les objectifs à atteindre et le périmètre du projet, en analysant les contraintes techniques et réglementaires, en étudiant le contexte et les enjeux afin de dimensionner le projet en termes de délai et budget.

**Problématique:**

Le pôle data du gouvernement français possède un jeu de données officiel sur l'évolution du prix des différentes essences vendu en France. Ces données sont rassemblées en moyenne hebdomadaire.Malheureusement, ils ne connaissent pas la méthodologie qui a permis de calculer ces moyennes hebdomadaires à l'époque où celle-ci ont été calculées.

Ils se posent donc la question de la véracité et fiabilité des transformations des données anciennement récoltés, et souhaiteraient qu' un recalcul soit fait sur d'autres données journalières existantes afin de vérifier la crédibilité des transformations passées.

Ils aimeraient aussi une vision plus fine des cours des différentes essences vendus en France, afin d'en tirer des conclusions et avoir une meilleure visibilité en terme journalière. Cet aperçu devra être mis à jour régulièrement.

**Objectif:**

Réalisation d'un audit sur la qualité des données officiels hebdomadaires en comparaison à un autre jeux de données journalier fournis par le gouvernement..

Recherche de décalage de prix sur les données historiques qui suggérerait un changement de méthode de calcul des moyennes hebdomadaires officielles.

Analyses sur une granulométrie temporelle plus fine qu' actuellement afin d'en tirer des conclusions.

Le rendu sera fait sous format de visualisations claires qui pourra être automatiquement rafraîchi lors de nouvelles données.

La base de données ainsi que les visualisations doivent pouvoir être sauvegarder et restaurer dans un environnement externe, afin d’assurer la facilité de déploiement du système dans d'autres environnements.

**Cadre réglementaire:**

La “doctrine cloud au centre” de l'État incite les applications gouvernementales à être déployées sur des environnements cloud de confiance.

Il sera donc nécessaire de construire l’application de manière à être entièrement déployé sur des clouds réputés.

**Environnement:**

L'environnement est laissé libre durant la réalisation du projet. Le projet devra ensuite être entièrement dockerisé afin d'être déployé sur les services cloud.  
Les données pourront aussi être stockées dans des clouds, car il ne s’agit pas de données sensibles. Nous ne travaillons qu' avec des données OpenData.

**Contraintes:**

Il est obligatoire de ne croiser que les données gouvernementales issus des différents sites internet de l'Etat, afin d'assurer une viabilité des données reconnue par celui-ci.

**RSE:**

Une attention particulière devra être faite sur le chargement de données.

En effet, il faut éviter le surchargement inutile sur l'ensemble de l'historique des données chaque jour, dans une optique de préservation écologique.

Les problèmes de confidentialité des données suivant la loi RGPD sont faiblement impactant sur ce projet, car l'intégralité des données est déjà disponible en OpenData et anonymisé. Les données sources sont donc déjà considérées comme respectant la loi RGPD.

# C3.1.2 : Dimensionner le projet en évaluant la charge de travail et les ressources nécessaires (humaines, matérielles) au regard des exigences attendues et des contraintes préalablement définies afin d’estimer le temps et le budget nécessaires à la faisabilité du projet.

Le projet sera réalisé par 2 Data-Engineers de niveau intermédiaire à plein temps ayant chacun un ordinateur suffisamment performant pour faire fonctionner des programmes avec des grands jeux de données.

**La charge de travail** est estimé comme ceci:

| Collecte du besoin et des attentes | 2 jours |
| --- | --- |
| Planification de la méthode de travail (Agile) | 2 jours |
| Collecte et analyse initiale des données brutes journalières | 2 jours |
| Nettoyage et Transformation des données brutes journalières | 5 jours |
| Chargement des données brutes journalières en base de données | 3 jours |
| Collecte et analyse initiale des données officielles hebdomadaire | 2 jours |
| Nettoyage et Transformation des données officielles hebdomadaire | 2 jours |
| Chargement des données officielles hebdomadaire en base de données | 2 jours |
| Création d’un jeu de données dénormalisé pour faciliter la visualisation | 2 jours |
| Analyses et création de graphiques de visualisation | 5 jours |
| Implémentation d’outils de sauvegarde et de restauration des bases de données | 2 jours |
| Implémentation d’outils de sauvegarde et de restauration des graphiques de visualisations | 2 jours |
| Rédaction de documentations techniques permettant la maintenance de l’application | 3 jours |
| **TOTAL** | **34 jours** |

La charge detravail estimée est donc de 34 jours travaillés.

La charge de la base de données en cloud sera financée par le client lors du déploiement, et pourra être en local lors du développement.

La charge des sauvegardes de bases de données et de graphique de visualisations stockées en cloud sera financée par le client lors du déploiement.

En prenant en compte la tailles des deux jeux de données traités, l’estimation du prix des sauvegardes de base de données lors de la période de développement sur un stockage S3 est estimée à un volume de 50Go par mois(surestimé) pour un montant moyen de 1 euros par mois. Il s’agit donc d’une charge négligeable.

Quant à la charge de l’outil de visualisation, celui-ci pourra être un outil Opensource.

En prenant en compte le fait que les 2 Data-Engineer travaillerons sur le projet a plein temps a 500 euros brut par jours, la durée totale du projet est estimée à environ **3 semaines et demi**, pour un montant total de **17 500 euros brut**.

La faisabilité de ce projet semble optimiste, car les différentes contraintes, cadre réglementaire, environnement et RSE ne semblent pas être de point bloquant quant à la capacité de conception et réalisation du projet.

# C3.1.3 : Rédiger la documentation projet, en identifiant les parties prenantes concernées, en prenant en compte l’ensemble des caractéristiques du projet, afin de clarifier et formaliser les attendus.

Lors de la réalisation de ce projet, un référent métier sera désigné au sein du pôle data du gouvernement afin d’être un point de contact durant tout le long de la réalisation.

L’un des Data-Engineer sera élu chef de projet et sera le responsable de la réalisation. L’autre Data-Engineer sera un exécutant tout en étant un acteur de la méthode Agile.

En prenant en compte les caractéristiques du projet, les attendus sont divisés en plusieurs parties:

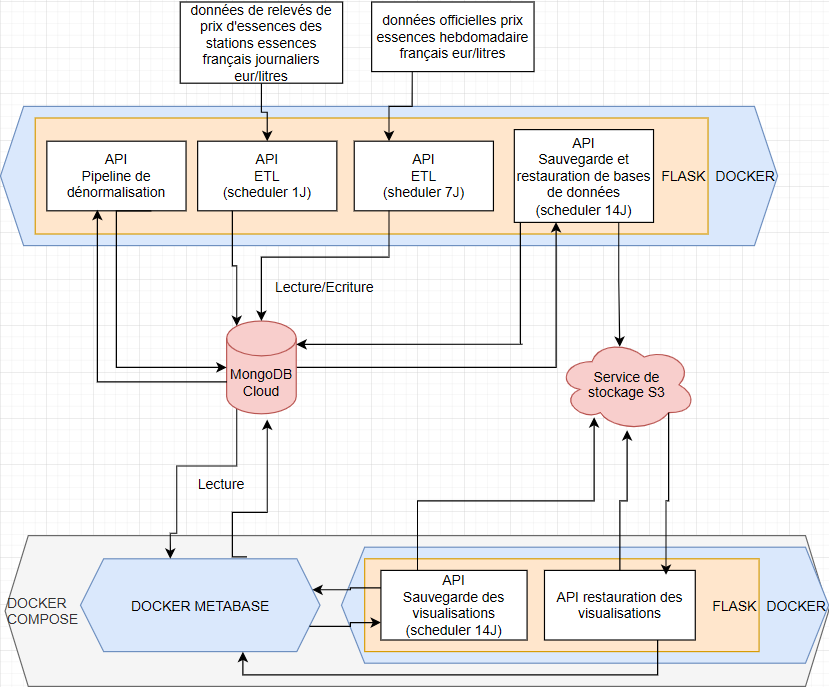
* Le docker contenant la partie d’extraction, transformation et chargement des deux flux de données en bases de données cloud. Il doit aussi permettre d’extraire les données traitées vers un espace de stockage externe cloud, ainsi que les restaurer. Tout doit être fonctionnel via des APIs et des planificateurs de taches.
* Les dockers contenant l’outil de visualisation. Il doit pouvoir se connecter en temps réel à la base de données cloud, et mettre à jour ces visualisations. Il doit pouvoir extraire les visualisations vers un espace de stockage externe cloud, ainsi que les restaurer. Tout doit être fonctionnel via des urls, APIs et des planificateurs de taches.
* Des documentations détaillées sur le fonctionnement des 2 dockers, sur le fonctionnement de l’outil de visualisation, sur les données sources ainsi que sur les données stockées en bases de données.

La création de base de données en cloud ainsi que la création d’espace de stockage externe cloud n’est pas considéré comme un rendu, car ceux-ci seront implémentés par l’équipe du pôle data du gouvernement. Cependant, il est nécessaire d’avoir une base de données fonctionnelle et un espace de stockage externe fonctionnel lors de développement afin de tester la solution et que l’application soit prête à l’emploi.

Le code sera réalisé en anglais, mais la documentation technique devra être français.

En prenant en compte le fait que le projet est à destination de personnes travaillant dans le domaine du traitement de données, les documentations et cahiers des charges pourront contenir des mots techniques.

Voici l’architecture complète attendu pour ce projet:



# 

# C3.2.1 : Planifier l’exécution du projet en organisant la répartition et l'ordonnancement des activités, le planning prévisionnel de réalisation et les ressources nécessaires à son exécution, en prenant en considération les personnes en situation de handicap afin de suivre les différentes phases du projet.

La méthodologie choisie pour la réalisation de ce projet est la Méthode Scrum, en utilisant des outils de planifications et visualisations. La réalisation de cette méthode se fera par l’élection d’un Data-Engineer qui tiendra le rôle de référent vers le client et de Scrum Master en plus de son travail de Data-Engineer.

Le Data-Engineer référent devra diviser le travail en tickets, animer un Poker Planning avec l’autre Data-Engineer afin d’évaluer la complexité de chaque ticket, puis réaliser un diagramme de Gantt permettant d’organiser le planning théorique et la répartition des tâches en prenant en compte la matrice RACI.

Cette méthode de travail en Scrum sera réalisée avec 6 sprint de 3 jours travaillés, avec un Daily tous les matin, ainsi qu’un rétro planning a la fin de chaque sprint pour faire un point sur l’avancement du projet. Des présentations clients seront faites à la fin de chaque sprint afin de leur fournir une visualisation de l’avancement du projet. Cette méthodologie permet d’impliquer le client et de travailler en agilité avec lui tout en renforçant la dynamique collective et l'implication de chacun.

Cette méthode agile telle que décrite permet de présenter régulièrement des versions potentiellement testables ou utilisables du projet au client. Elle permet une meilleure estimation des tâches, un meilleur suivi de l’avancement et une forte réactivité face aux imprévus. Le client restera joignable durant toute l’exécution du projet afin de répondre rapidement à de potentiels points bloquants.

Avant la répartition des tâches, la matrice RACI est réalisé pour bien comprendre les compétences et les rôles de chacun:

**R** Responsable Celui qui réalise concrètement la tâche, l’exécutant.

**A** Approbateur Celui qui est redevable du résultat, valide ou prend la décision finale.

**C** Consulté Celui qui est sollicité pour donner un avis ou un conseil.

**I** Informé Celui qui est tenu au courant de l’avancement ou du résultat.

| **Tâches** | **Data Engineer reférent** | **Data Engineer 2** | **Référent métier du pôle data** | **Autres personnes du pôle data** |
| --- | --- | --- | --- | --- |
| Collecte du besoin et des attentes | R | C | A | C |
| Planification de la méthode Agile | R | A | A | I |
| Collecte et analyse initiale des données brutes journalières | A | R | C | I |
| Nettoyage et Transformation des données brutes journalières | A | R | A | I |
| Chargement des données brutes journalières en base de données | A | R | C | I |
| Collecte et analyse initiale des données officielles hebdomadaire | R | A | C | I |
| Nettoyage et Transformation des données officielles hebdomadaire | R | A | C | I |
| Chargement des données officielles hebdomadaire en base de données | R | A | C | I |
| Création d’un jeu de données dénormalisé pour faciliter la visualisation | A | R | C | I |
| Analyses et création de graphiques de visualisation | A | R | C | I |
| Implémentation d’outils de sauvegarde et de restauration des bases de données | R | A | A | I |
| Implémentation d’outils de sauvegarde et de restauration des graphiques de visualisations | R | A | A | I |
| Rédaction de documentations techniques permettant la maintenance de l’application | R | A | A | I |

Le diagramme de Gantt est ensuite réalisé en prenant en compte la matrice RACI et les notes de complexités donné a chaque ticket lors du poker planning.

**Data engineer référent** / **Data engineer 2**

| **Tâches** | | | **Sprint 1**  **(3 jours)** | | | | **Sprint 2**  **(3 jours)** | | | **Sprint 3**  **(3 jours)** | | | **Sprint 4**  **(3jours)** | | | **Sprint 5**  **(3 jours)** | | | **Sprint 6**  **(3 jours)** | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Collecte du besoin et des attentes | | |  | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Planification de la méthode Agile | | |  | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Collecte et analyse initiale des données brutes journalières | | |  | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Nettoyage et Transformation des données brutes journalières | | |  | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Chargement des données brutes journalières en base de données | | |  | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Collecte et analyse initiale des données officielles hebdomadaire | | |  | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Nettoyage et Transformation des données officielles hebdomadaire | | |  | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Chargement des données officielles hebdomadaire en base de données | | |  | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Création d’un jeu de données dénormalisé pour faciliter la visualisation | | |  | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Analyses et création de graphiques de visualisation | | |  | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |
| Implémentation d’outils de sauvegarde et de restauration des bases de données | | |  | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Implémentation d’outils de sauvegarde et de restauration des graphiques de visualisations | | |  | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Rédaction de documentations techniques permettant la maintenance de l’application | | |  | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

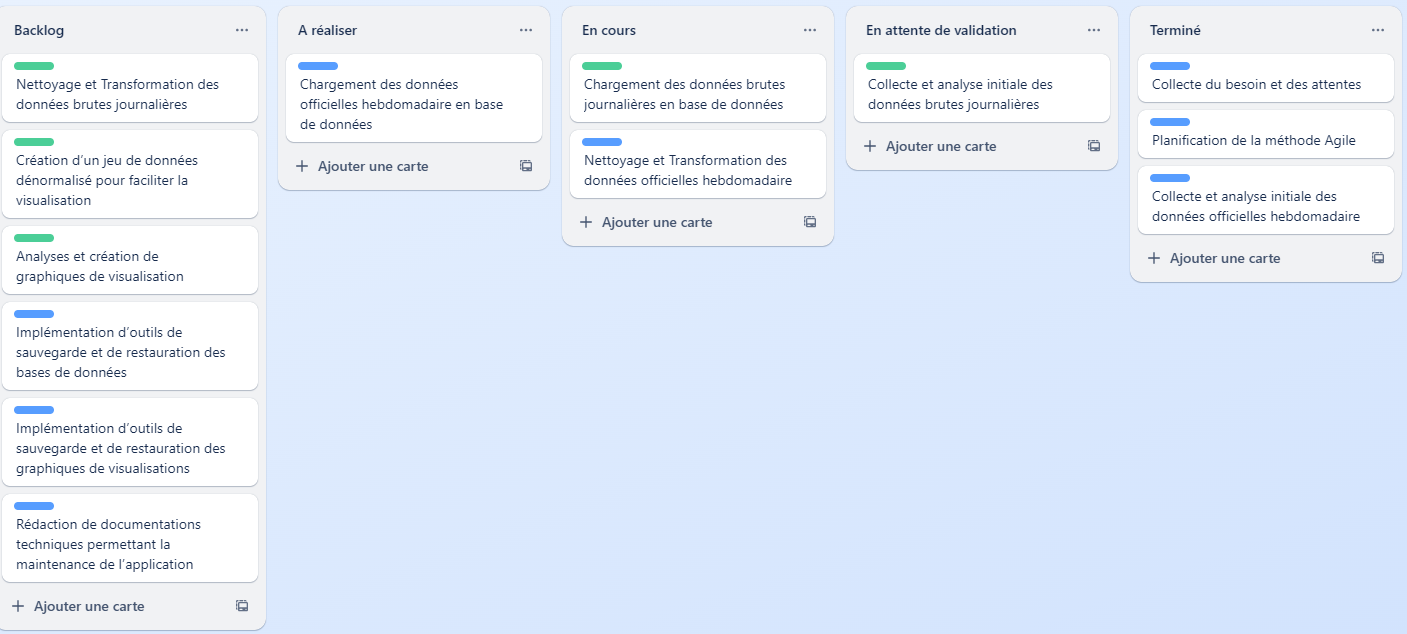
Si un collaborateur était handicapé, je ferais en sorte que l’équipe se rapproche du Référent handicap afin de mieux comprendre le handicap et recevoir des conseils et démarches nécessaires au bon accueil de la personne handicapé. Je demanderais aussi des conseils directement à la personnes handicapé, puis je mettrais en place les recommandations exprimées afin d’aménager au mieu l’espace de travail ainsi que la méthodologie de travail au handicap de la personne. Par exemple, en cas de handicap de trouble de déficit de l’attention, je pourrais mettre un bureau à part supplémentaire afin que la personne puisse s’isoler lorsqu’elle le souhaite. Je lui laisserais aussi la possibilité de télétravailler lorsqu’elle le considère nécessaire, et si nécessaire adapterais le planning de charge a ces capacités.

# C3.2.2 : Suivre l’avancement du projet en mettant en place un outil de suivi (logiciel de suivi, tableau de bord), en définissant les indicateurs (qualitatifs et/ou quantitatifs) pour chaque jalon défini dans le planning, en réalisant des reportings et des comptes rendus de réunion afin d’anticiper les aléas éventuels.

Les outils de suivi et de gestion de tickets les plus courants lors d’un développement de logiciel en méthode agile sont Azure, ou bien Jira.

Cependant, au vu de l’équipe réduite de 2 Data-Engineer, et de la dimension du projet a réalisé, le déploiement d’outils de suivi aussi lourds semble être démesuré par rapport au besoin de gestion et serait probablement contre-productif.

L’outil de gestion utilisé pour la réalisation de ce projet sera donc Trello. Il permet d’être facilement accessible, gratuit, et offre l’essentiel des fonctionnalités nécessaires pour le suivi et la gestion des tickets tels que la création de tickets, la prise de notes dans les tickets, l’assignation, et le déplacement à travers différents états. De plus, c’est un outil collaboratif ou plusieurs personnes peuvent y avoir accès.



Chaque couleur sur les tickets(vert et bleu) représente l’assignation de la tâche à un Data-Engineer.  
A chaque début de sprint, les tickets qui devront être réalisés passe de la colonne “Backlog” à la colonne “À réaliser”. Les tickets changent ensuite de colonne suivant l’avancé de chaque collaborateur. En fin de sprint, nous accordons une importance à chaque ticket en cours de réalisation ou terminé. Nous évaluons ensuite pourquoi certains tickets censés être terminés ne le sont pas encore, et évaluons l’état d’avancement par rapport au diagramme de Gantt. Ainsi, les délais et les coûts de développement sont maîtrisés.

Le client peut aussi avoir accès au Trello afin qu’il ait un accès direct à l'évolution de son projet, ce qui améliore grandement sa compréhension et sa visibilité de l'avancée. Cela lui permet d’avoir reporting de l’état actuel de l’avancement, et simplifie sa compréhension lors des comptes rendus de réunion à chaque fin de sprint.

# C3.3.1 : Évaluer les besoins en compétences de l’équipe projet, en collaborant avec le service Ressources Humaines, en établissant un plan de développement des compétences et en orientant les membres de l’équipe vers des formations adaptées, afin de renforcer l’équipe responsable de mener à bien le projet DATA.

blabla

# C3.3.2 : Piloter l’équipe projet en affectant les missions à réaliser, en prenant en compte les spécificités des membres de l’équipe, en intégrant les spécificités d’un contexte multiculturel, international, en utilisant les différentes techniques de communication et d’animation managériale pour favoriser le bon fonctionnement de l'équipe.

# C3.3.3 : Procéder aux arbitrages et aux réajustements nécessaires à partir de l’analyse des écarts entre le prévisionnel et l’état du projet à date, en utilisant des outils d’aide à la décision (ex : logigramme) afin de garantir la performance du projet dans le respect des objectifs de qualité, coûts et délai.

# C3.4.1 : Mettre en place un système de veille technologique et réglementaire en matière de science des données et d’Intelligence Artificielle à l’aide de recherches documentaires, de plateformes de partage, de webinars afin d’être alerté des évolutions qui impacteraient les pratiques métier.

# C3.4.2 : Intégrer dans ses pratiques métier les enjeux en termes de données responsables, de responsabilité sociétale et environnementale (RSE), de sécurité, d’éthique et de confidentialité des données en se tenant informé des évolutions du cadre juridique, à travers une recherche documentaire ou en étant accompagné par des juristes afin d’agir dans le respect de la législation.