10 mars 2025

Webinaire

Gérer ses données dans le cycle de vie d'un projet

Gaëlle Jaouen, Eva Legras

AgroParisTech

Contexte global autour des données de la recherche (DR)

Enjeux autour des données :

- Accélération de la production et de la manipulation de DR (essor technologique)
- Définition d'un cadre légal et politique d'accessibilité et de réutilisabilité des DR
- Volonté de maitriser les conditions de circulation et d'utilisation des DR (sécurisées, fiables, traçables…)

Volonté de la Recherche (et de notre établissement) de :

- Formaliser des règles communes de gestion et d'ouverture
- Accompagner les agents vers de bonnes pratiques et la connaissance des droits et devoirs
- Soutenir la validation des résultats de recherche
- Préserver le patrimoine de données et les DR à venir (coûteuses, uniques/non-reproductibles)

Lancement de plusieurs politiques DR intra-organismes : AgroParisTech (1er Janvier 2021), INRAE, IRD, CNRS, …

Définition des DR (politique AgroParisTech)

Les données de la recherche sont l'ensemble des enregistrements factuels (informations numériques, textuelles, visuelles ou sonores…) collectés, observés ou créés dans le cadre d'une activité de recherche.

Ces données sont nécessaires à la construction de la recherche, à l'établissement et à la validation des résultats de recherche.

Lorsque les données n'ont pas encore été traitées ou contextualisées, on parle de données brutes.

Entre données brutes et résultats finaux, les transformations successives des données peuvent servir de base à d'autres travaux de recherche et peuvent donc, de ce fait, être également considérées comme données de la recherche.

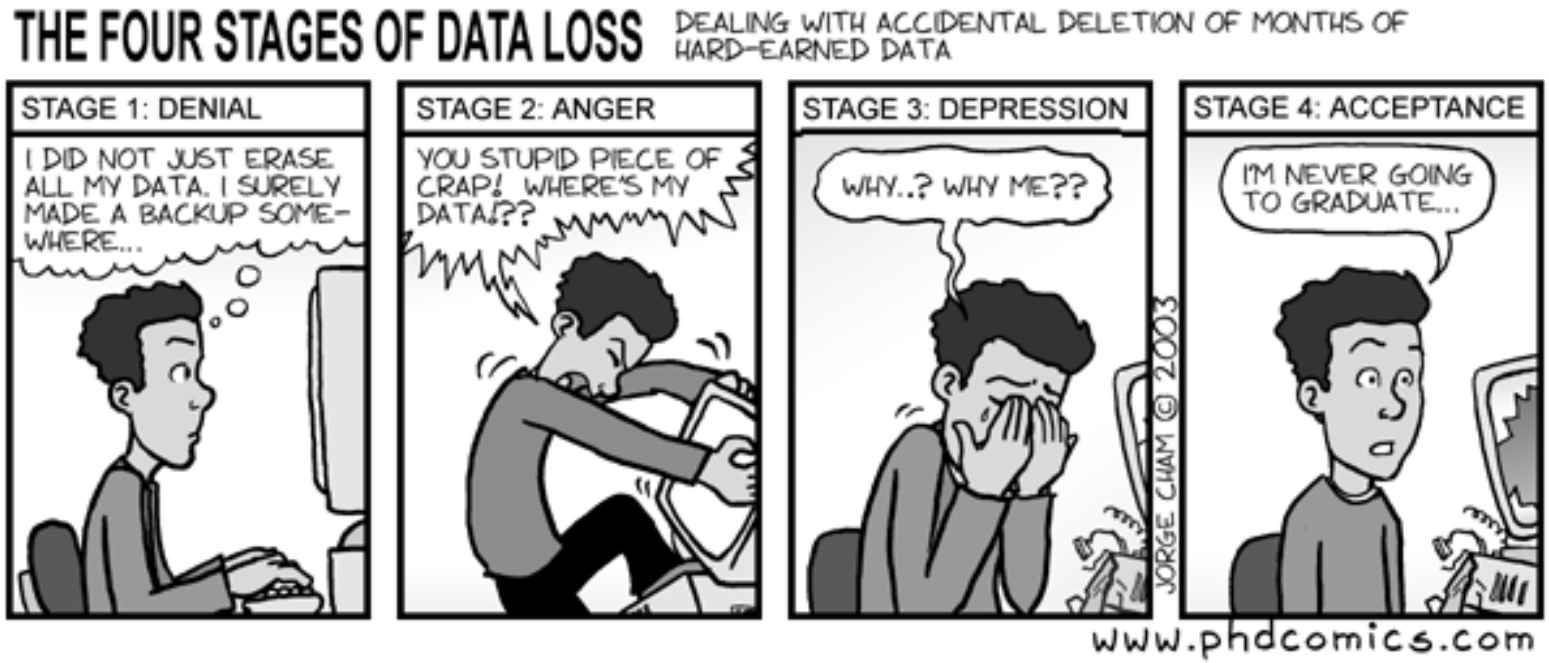
Quand s'y intéresse-t-on?

- Gestion d'un projet
- Gestion d'un laboratoire, d'une équipe, d'une plateforme, d'une unité de recherche
- Gestion du patrimoine d'un établissement

Le cycle de vie des données

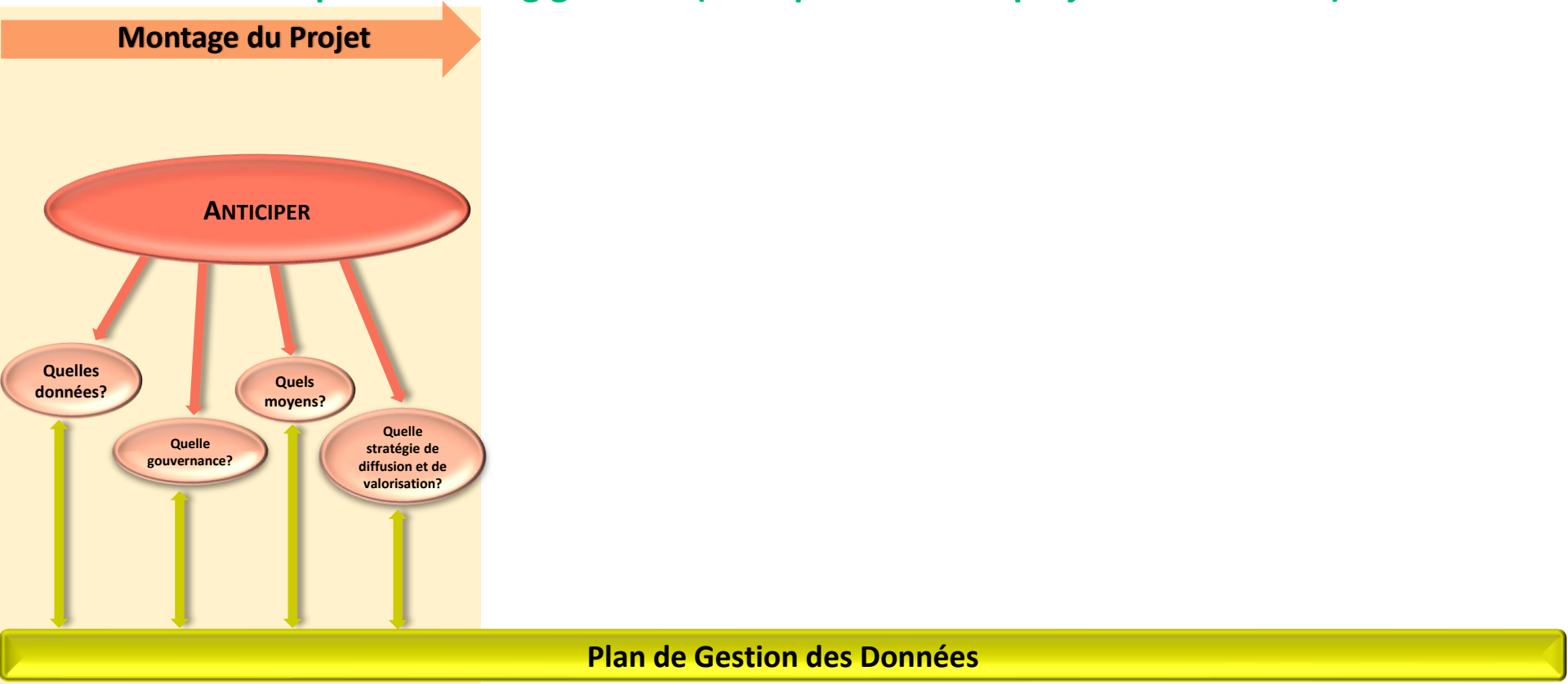
Adapté d'après University of Sidney et URFIST de Paris.

Les enjeux: de l'importance de l'anticipation …



Pour éviter ça, reprenons point par point!

Principes clés et logigramme (exemple : cas d'un projet de recherche)



Plan de Gestion des Données

Document pense-bête, synthétique, collaboratif et itératif, répertoriant tous les aspects de la gestion des DR du projet : production, documentation, organisation, stockage, partage, protection, ouverture, archivage, moyens humains et financiers

Objectifs :

- Soulever les points à discuter pour le bon déroulement du projet
- Eviter les mauvaises surprises et les oublis
- Prévoir les moyens et ressources nécessaires, anticiper les coûts
- Se mettre d'accord au sein du projet, fixer les règles

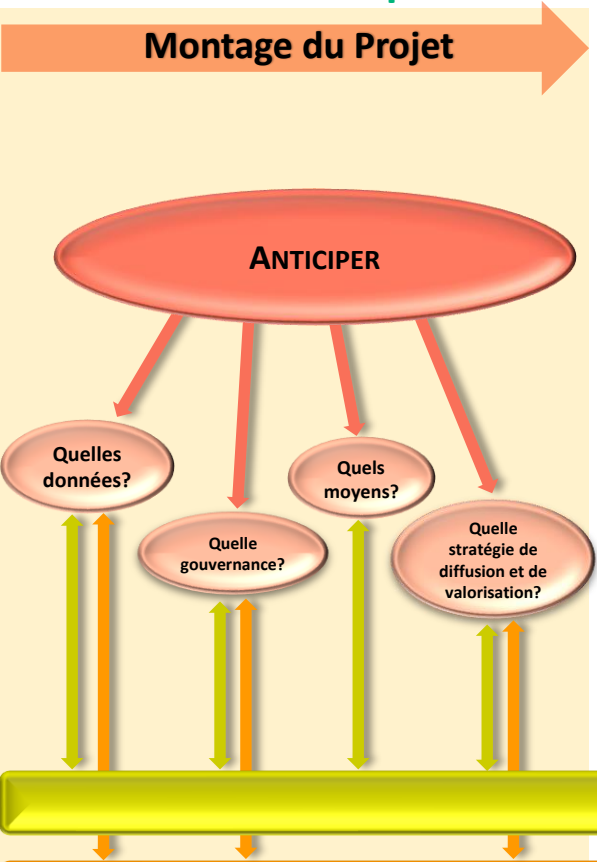
Plusieurs trames disponibles (ANR, Horizon Europe, institutionnelles, Trames AgroParisTech Projet et Entité) rassemblées sur DMP Opidor

Précautions :

- Lister les types de données et leur volume prévisionnel pour anticiper leurs particularités et les précautions à prendre (cadre juridique, infrastructures…)
- Désigner un/des responsable(s) du PGD dans le projet

Fiche pratique : Plan de gestion de données

Principes clés et logigramme (exemple : cas d'un projet de recherche)



Plan de Gestion des Données

Droits ? Devoirs ? Cadre légal, réglementaire et éthique

Cadre légal de protection et d'ouverture des données de la recherche

Ensemble de lois et règlements (français/européens) pour encourager l'ouverture de la science :

- Protéger les droits des producteurs de données (paternité, traçabilité) et des sujets d'études (confidentialité, sensibilité, anonymat)
- Cadrer les droits des utilisateurs
- Stimuler les collaborations

L'ouverture des données publiques numériques achevées est maintenant une norme et non une option !

- Données de la recherche publique = document administratif
- Les décisions reviennent aux établissements, pas au chercheur/ingénieur

Quelques cas particuliers (qu'il faut justifier), notamment : données sensibles, secret industriel, propriété intellectuelle

« Aussi ouvert que possible, aussi fermé que nécessaire »

Attention : Anticipation et entente entre partenaires sont fondamentales, surveiller les conditions des contrats et conventions

Fiche pratique : Cadre juridique données de la recherche

Les questions juridiques et réglementaires, un point important à anticiper !

Définir la titularité des données : qui est responsable de quoi sur les données du projet ?
⇒ Nature des données ? Producteurs ? Partenaires ? Convention ?

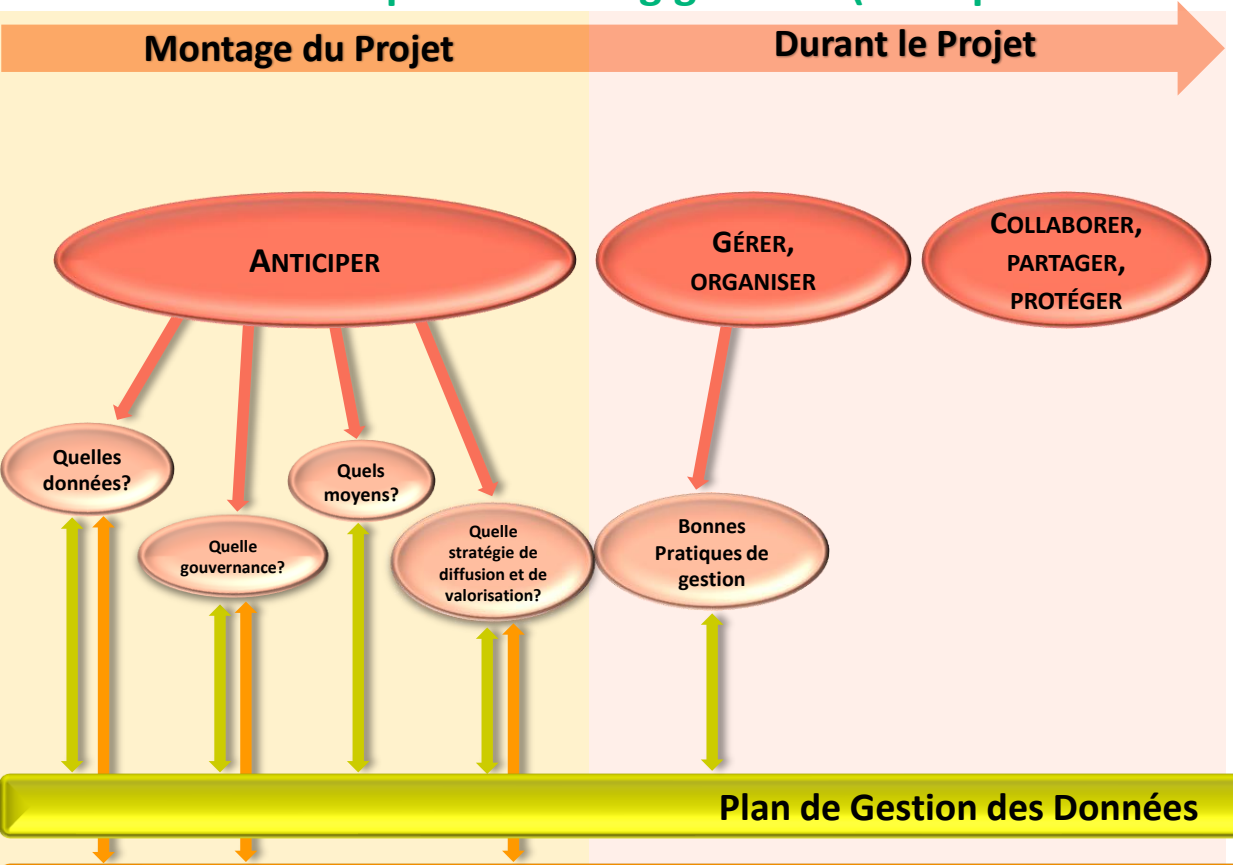
Identifier les obligations à remplir depuis l'étape de collecte des données jusqu'à l'ouverture des résultats (politiques institutionnelles, politiques des agences de financement) :

Qui peut voir et réutiliser les données du projet, à quel moment ?
⇒ secret ; confidentialité
⇒ communicabilité ; ouverture (à anticiper)

Quels sont les livrables à fournir en cours ou en fin de projet ?
⇒ Plan de gestion de données ?
⇒ Données associées aux publications ?
⇒ Ouverture par défaut ?

Sous quelles conditions particulières les données peuvent être collectées/utilisées ?
⇒ Enjeux éthiques
⇒ Enjeux de sécurité
⇒ Enjeux réglementaires

Principes clés et logigramme (exemple : cas d'un projet de recherche)



Bonnes Pratiques de Gestion

Ensemble de règles et réflexes pour optimiser :

- La gestion des fichiers et des données à l'intérieur des fichiers (documentation, nommage, organisation, versionnement, homogénéité)
- La collaboration (formats ouverts, outils collaboratifs)
- La limitation des risques de confusions, d'erreurs, de pertes de données (sécurité du stockage, politique de sauvegardes)

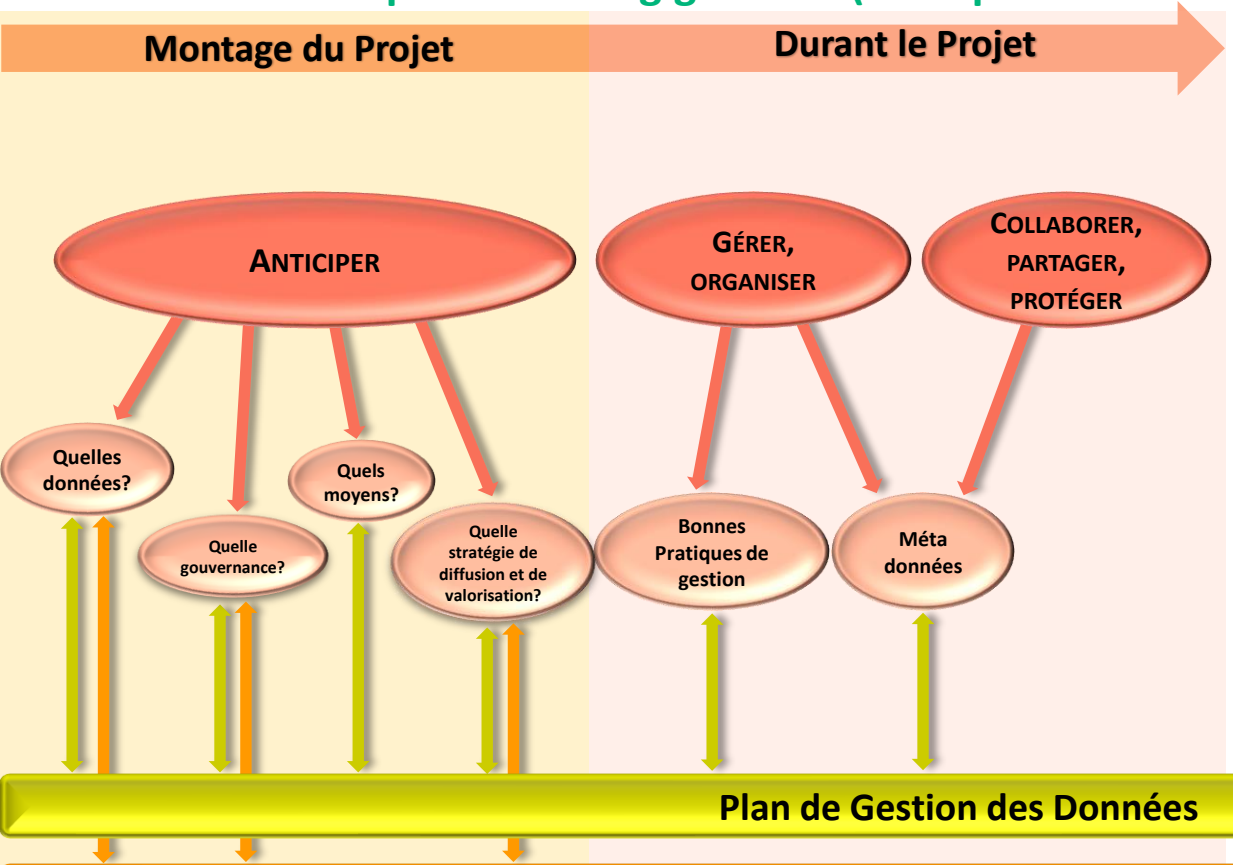
Gagner du temps et… éviter les catastrophes !

Les précautions à prendre :

- Rester réalistes et tenir compte des moyens disponibles (humains, techniques, financiers)
- Tous les membres du projet suivent les mêmes règles pour que tout le monde se comprenne!
- Documenter autant que possible pour être transparents et accessibles aux nouveaux collaborateurs

Fiche pratique : Bonnes pratiques de gestion

Principes clés et logigramme (exemple : cas d'un projet de recherche)



Métadonnées

Tous les éléments de description des DR = Carte d'identité d'un jeu de données :

- contexte d'acquisition (pourquoi, comment, auteur(s)…)
- unité de mesure
- dates de collecte, de révision
- format de fichier
- Etc.

Pour assurer la compréhension et la réutilisabilité éclairée des DR, conserver les informations de genèse des DR, faciliter les collaborations

Les précautions à prendre :

Choisir un standard de métadonnées (Dublin Core, Darwin core, EML, DataCite…) adapté au projet et à la communauté scientifique, avoir des métadonnées les plus riches possibles, utiliser autant que possible des vocabulaires contrôlés.

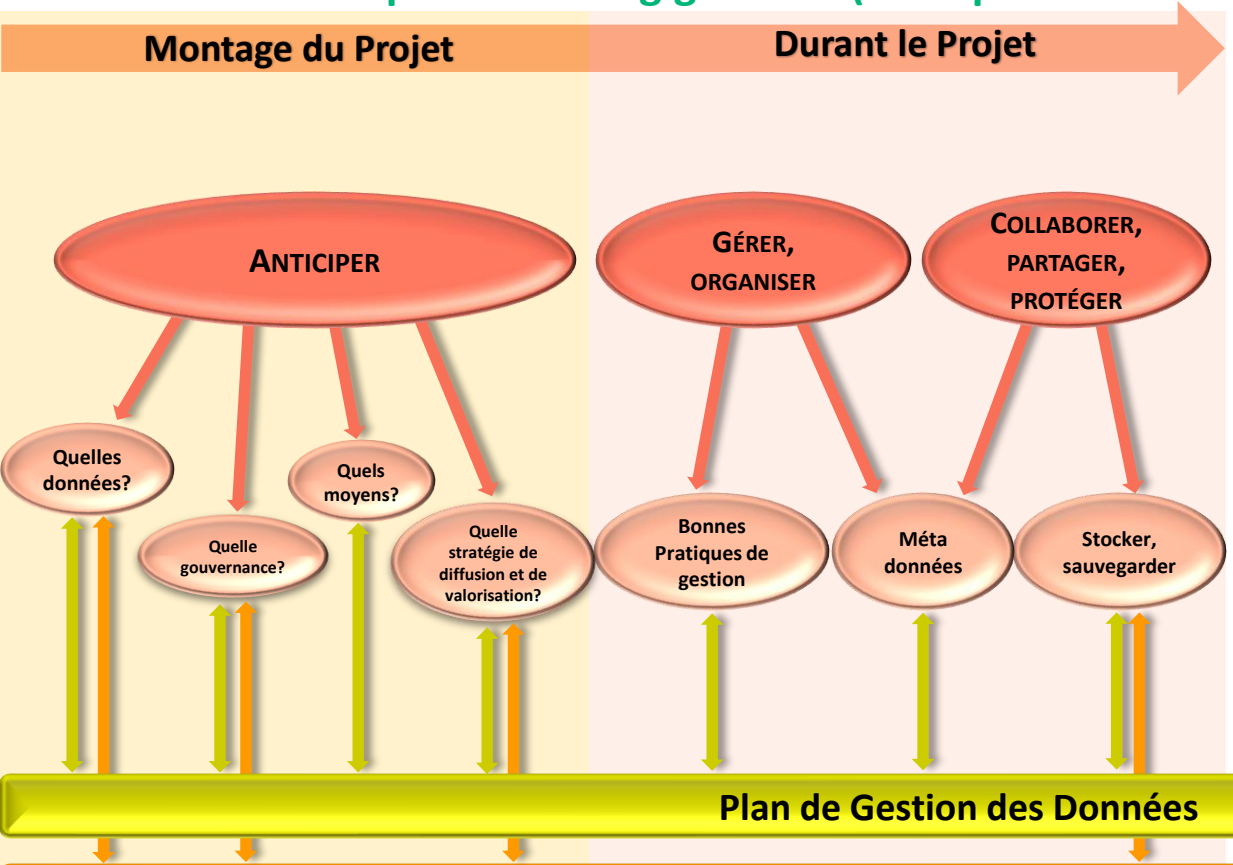
Ressources :

Standards de métadonnées : RDAlliance ; DCC ; FairSharing

Thésaurus : Loterre ; vocabulaires ouverts INRAE

Fiche pratique : Métadonnées

Principes clés et logigramme (exemple : cas d'un projet de recherche)



Stratégie de stockage

Elle démarre par la question « magique » à se poser pour chacun des cycles de vie du projet de recherche.

=> Le projet a besoin de

Quel(s) volume(s) de données

Pour stocker quel type de données

Selon quelles modalités d'accès

Pour quelle durée

Avec quel niveau de sensibilité

Selon quels modèles économiques et écologiques

Avec quelles contraintes réglementaires et contractuelles

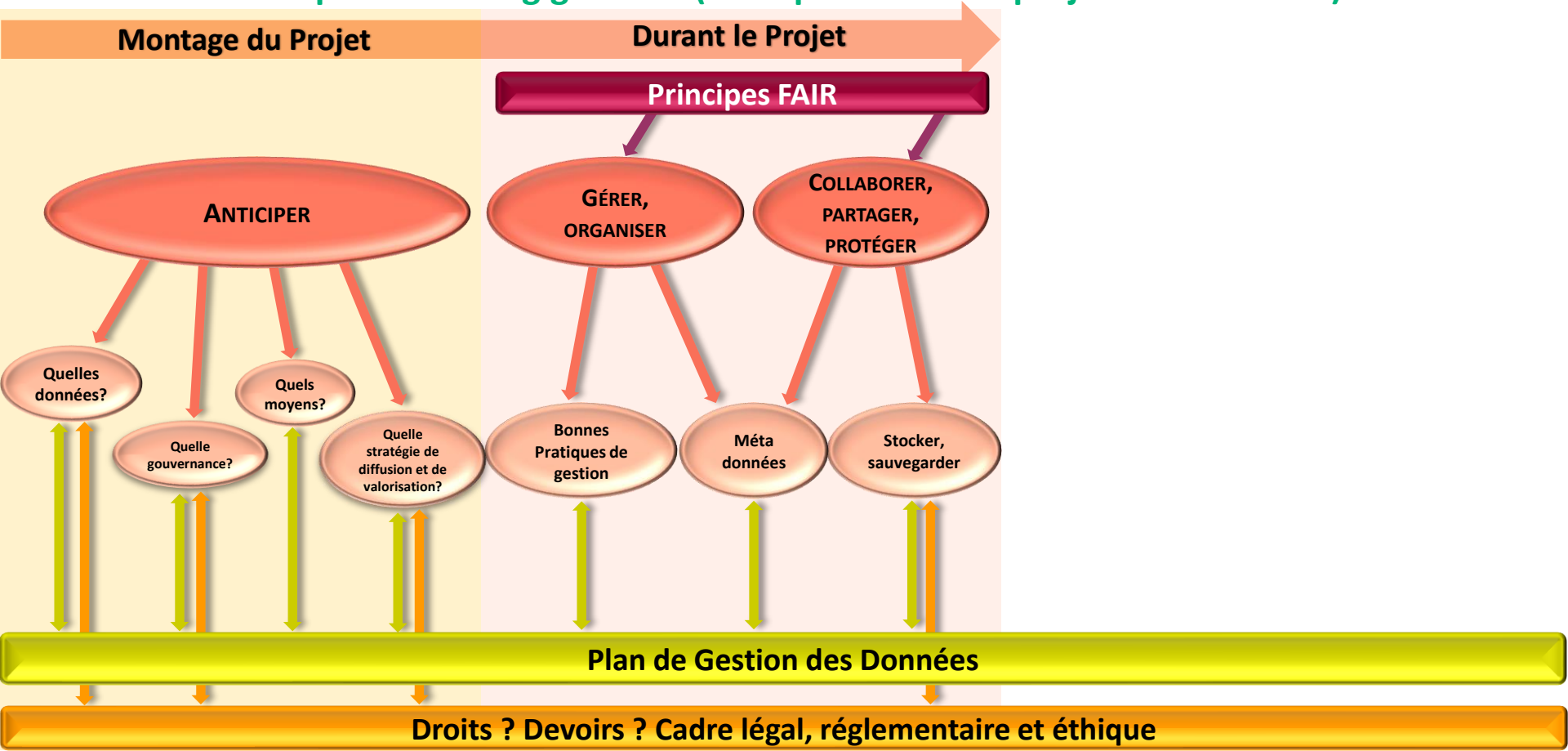
La réponse à cette question permet de sélectionner des types de stockage candidats, à anticiper par intégration au projet et/ou à l'identification de solutions existantes.

Identifier cette stratégie ne veut pas dire qu'une réponse pourra y être apporter de manière simple et rapide

=> *l'anticipation est la meilleure façon de garantir la faisabilité de cette stratégie*

Fiche pratique : Stratégie de stockage

Principes clés et logigramme (exemple : cas d'un projet de recherche)



AgroParisTech

Principes FAIR

Qu'est-ce que c'est ?

Encore un pense-bête! 4 ensembles de précautions à prendre pour que les DR soient Faciles à trouver, Accessibles, Interopérables et Réutilisables

À quoi ça sert ?

Standardiser (formats libres et ouverts, vocabulaires, unités…), documenter (métadonnées), tracer (identifiant pérenne…)

Principes FAIR ≠ ouverture obligatoire : des données FAIR peuvent être confidentielles

Les gros avantages ?

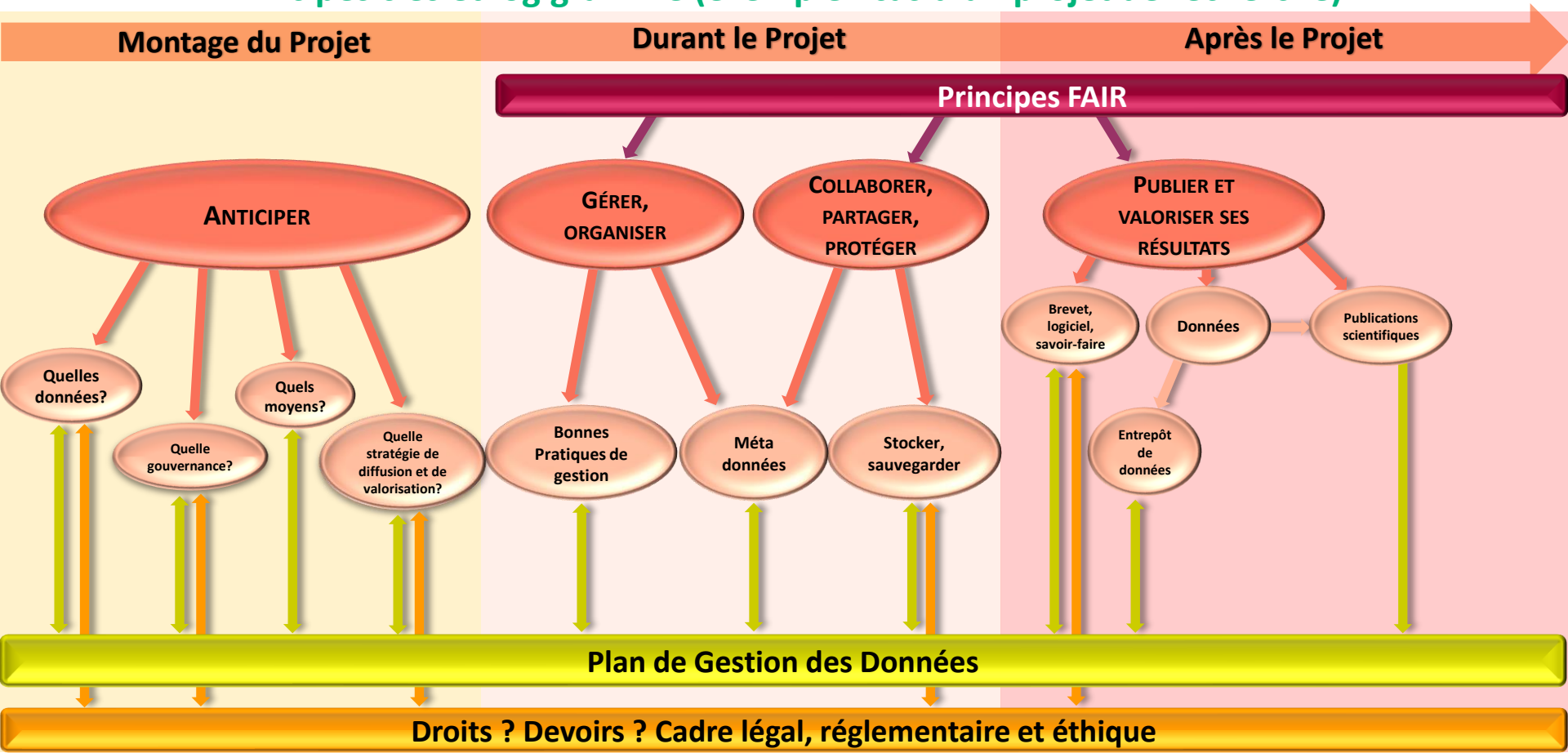
Facilite la conservation, la traçabilité, la publication, les collaborations pour l'ouverture et même avant!

Les précautions à prendre ?

Respecter les standards

Fiche pratique : Principes FAIR

Principes clés et logigramme (exemple : cas d'un projet de recherche)



Entrepôts de Données

Outils clés en main pour publier ses données = base de données en ligne + principes FAIR

Publier ≠ Partager

Publier = Ouvrir durablement, efficacement et de façon juridiquement cadrée ses DR à une communauté ciblée ou plus large

Tout est guidé et cadré : grilles de métadonnées, formatage des données, attribution d'identifiants pérennes (DOI, PURL…), outils de recherche et de découverte…

- Choisir son entrepôt : Thématique/Généraliste? Gouvernance? Durabilité? Certification? (ex CoreTrustSeal)
- S'assurer de son droit à ouvrir ses données
- Lier ses données à ses publications (via identifiants pérennes)

Annuaires d'entrepôts : re3data, FairSharing, OAD, OpenDOAR

Exemples d'entrepôts : disciplinaires : GBIF (biodiversité), European Nucleotide Archive (génétique), Nakala (SHS) ; **généralistes** : Recherche Data Gouv ; **institutionnels** : Cirad DataVerse, IRD DataSuds

Fiche pratique : Entrepôts de données

Entrepôts de Données

Ouvrir la science !

COSO, 2024

Méthode d'identification des entrepôts thématiques de confiance établie par le Comité pour la Science Ouverte (mars 2024) :

7 critères d'exclusion :

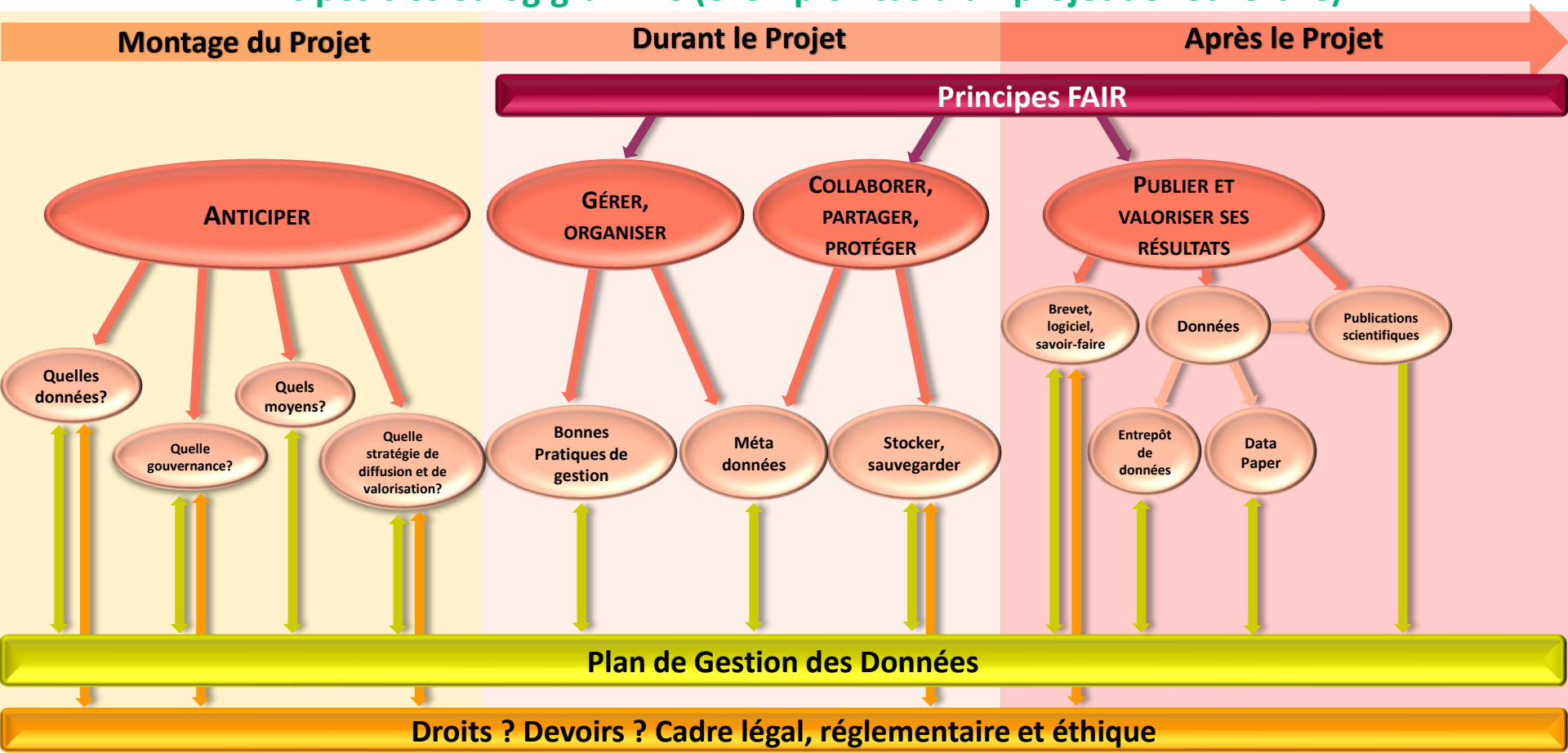
- Absence de modération des dépôts
- Absence d'identifiants pérennes
- Absence de garantie sur la pérennité de l'infrastructure
- Pratique de la cession de droits
- Politique tarifaire excessive
- Localisation des données hors UE pour certains types de données (personnelles notamment)
- Dépôt restreint par l'affiliation institutionnelle

7 critères de description :

- Champ disciplinaire
- Données acceptées
- Identifiant pérenne
- Pérennité des données
- Type de modération
- Possibilité d'ajouter un embargo
- Limite de volume

1ère liste sur la base de ces critères : 49 entrepôts disciplinaires identifiés

Principes clés et logigramme (exemple : cas d'un projet de recherche)



Article de Données ou Data Paper

Article qui décrit précisément un jeu de données : genèse précise et complète, potentiel de réutilisation, accès aux données. MAIS pas de résultats scientifiques, d'analyses ou d'interprétations ; peut être publié en association avec un article scientifique classique pour cela.

- Informe la communauté de l'existence et de la valeur de ces données
- Synthétise et simplifie la communication des données
- Complète le dépôt en entrepôt, connectable à un article scientifique
- Valorise le temps passé à gérer les données

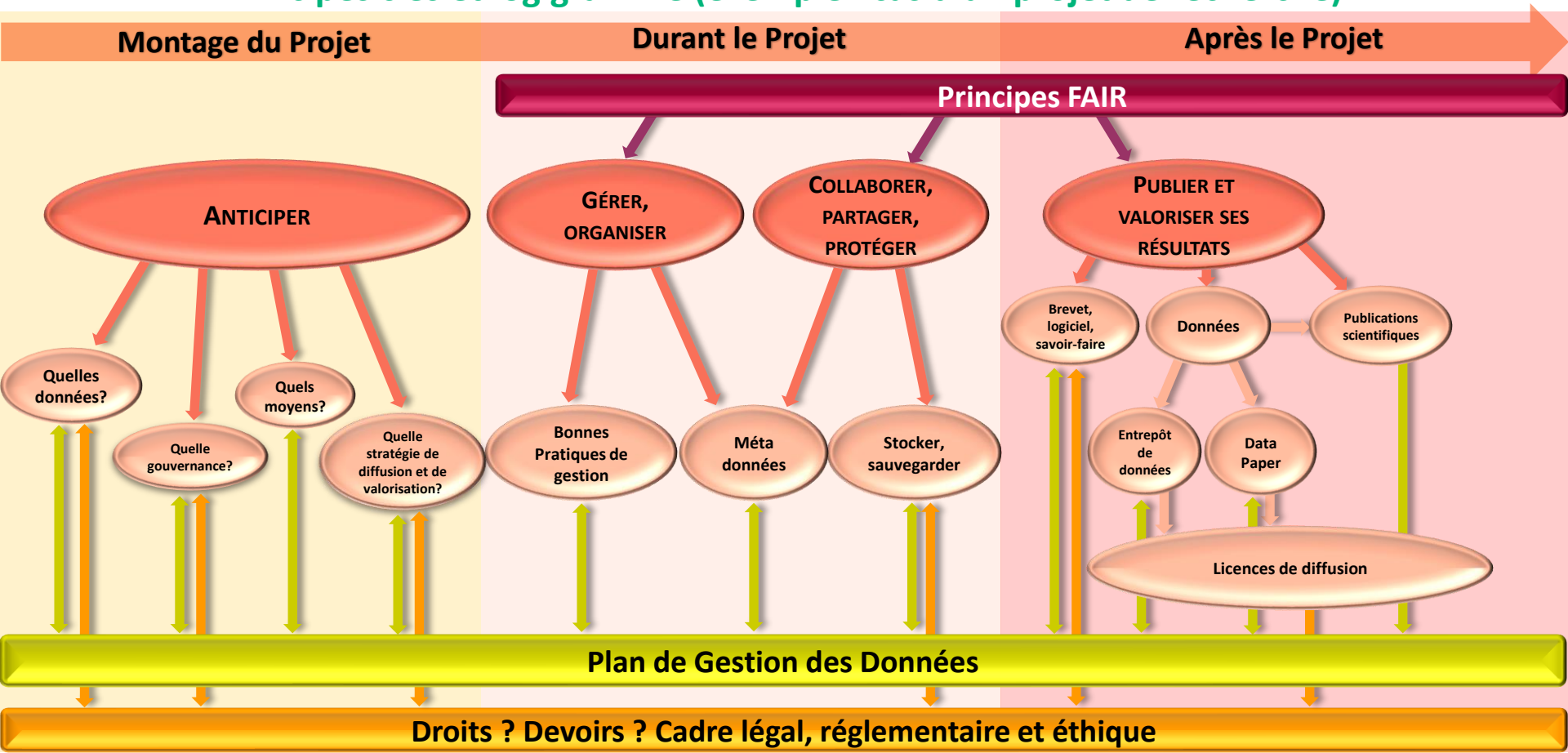
Les précautions à prendre :

- Choix du journal (thématique, audience, comité de lecture…)
- Choix de l'entrepôt (compatibilité)
- Respect des consignes du financeur

Ressources : listes de revues spécialisées publiées par INRAE, le Cirad, le GBIF…

Fiche pratique : Data Paper

Principes clés et logigramme (exemple : cas d'un projet de recherche)



Licences de Diffusion

Qu'est-ce que c'est ?

Outil juridique pour spécifier publiquement les formes de diffusion et de réutilisation autorisées par le détenteur des droits d'un contenu

À quoi ça sert ?

A choisir ce qu'on autorise une fois pour toutes avec les DR publiées (reproduction, réutilisation, transformation…)

Les gros avantages ?

Synthétise les conditions, gratuit, facile d'utilisation. Protège les producteurs de données.

Les précautions à prendre ?

Vérifier les particularités des DR, les accords des partenaires, les conditions demandées par les financeurs du projet et les journaux de DataPapers

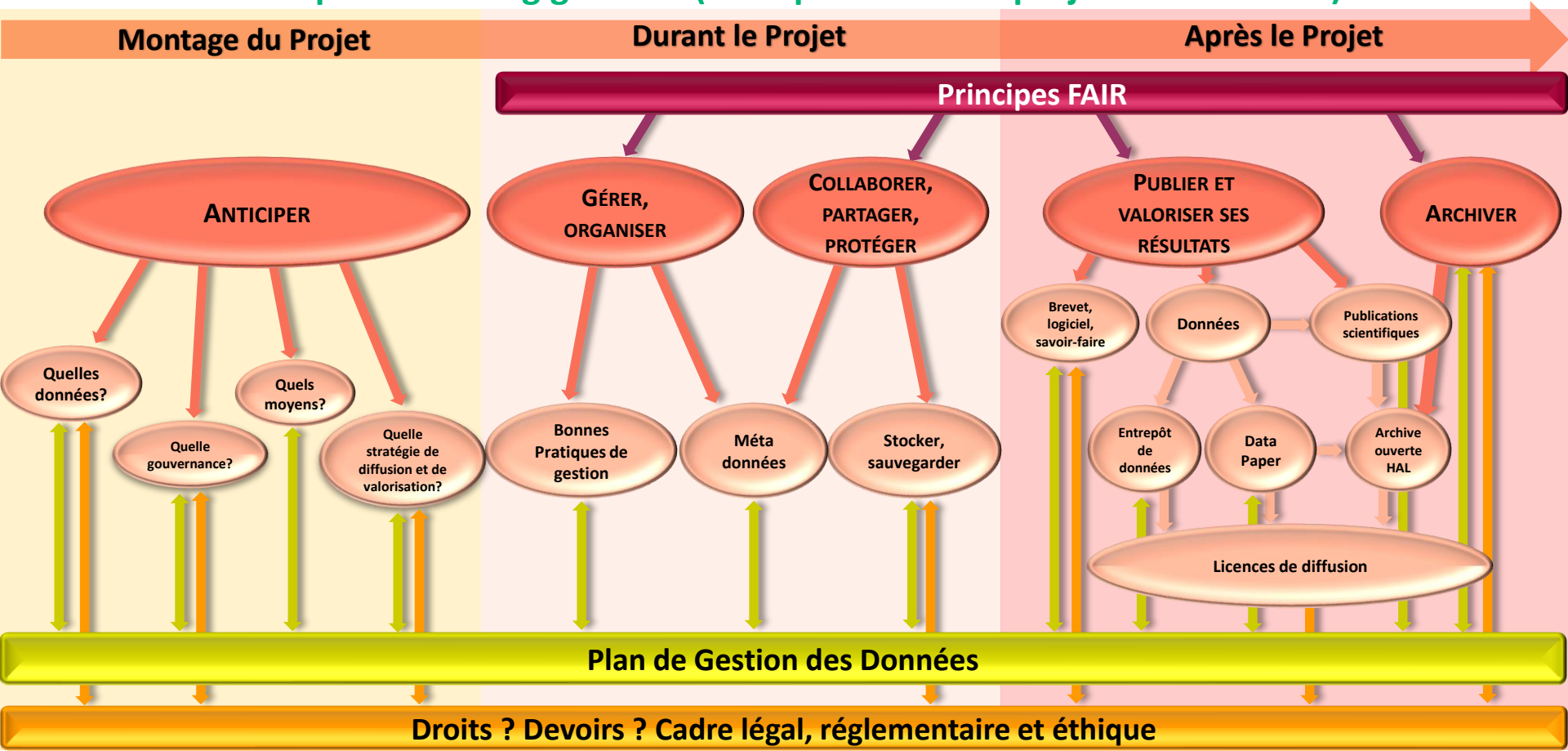
Licences courantes : Creative Commons (ex: 

).

Pour le droit français : https://www.data.gouv.fr/fr/licences

Fiche pratique : Licences de diffusion

Principes clés et logigramme (exemple : cas d'un projet de recherche)



Quelques conclusions…

→ Les données sont le carburant de la science ☺

- Gérer ses données de la conception du projet à l'ouverture pour:
  - Gagner du temps, être efficace
  - Éviter les erreurs et les catastrophes!
  - Attester la fiabilité des résultats
  - Produire des données réutilisables par soi et par les autres
  - Faire avec ses moyens techniques, humains, financiers et anticiper
- Ne jamais modifier les données brutes
- Documenter : noter et partager les démarches et conventions choisies
- S'entendre entre partenaires sur les règles de fonctionnement

→ Faire au mieux pour soi et pour le projet, rester réaliste et raisonnable!

AgroParisTech

Pour aller plus loin

Ressources atelier de la donnée ADOC
Ressources AgroParisTech
Ressources INRAe
Plateforme de ressources Doranum

Lectures complémentaires :

- UNESCO, Recommandation sur une science ouverte, 2021
- MESR, Deuxième plan national pour la science ouverte, 2021
- MESR, Série Passeport pour la science ouverte, 2023
- Coalition for Advancing Research Assessment https://coara.eu/

On en discute?

AgroParisTech