Les données de la recherche dans les appels à projets Horizon 2020

Produire un Data Management Plan

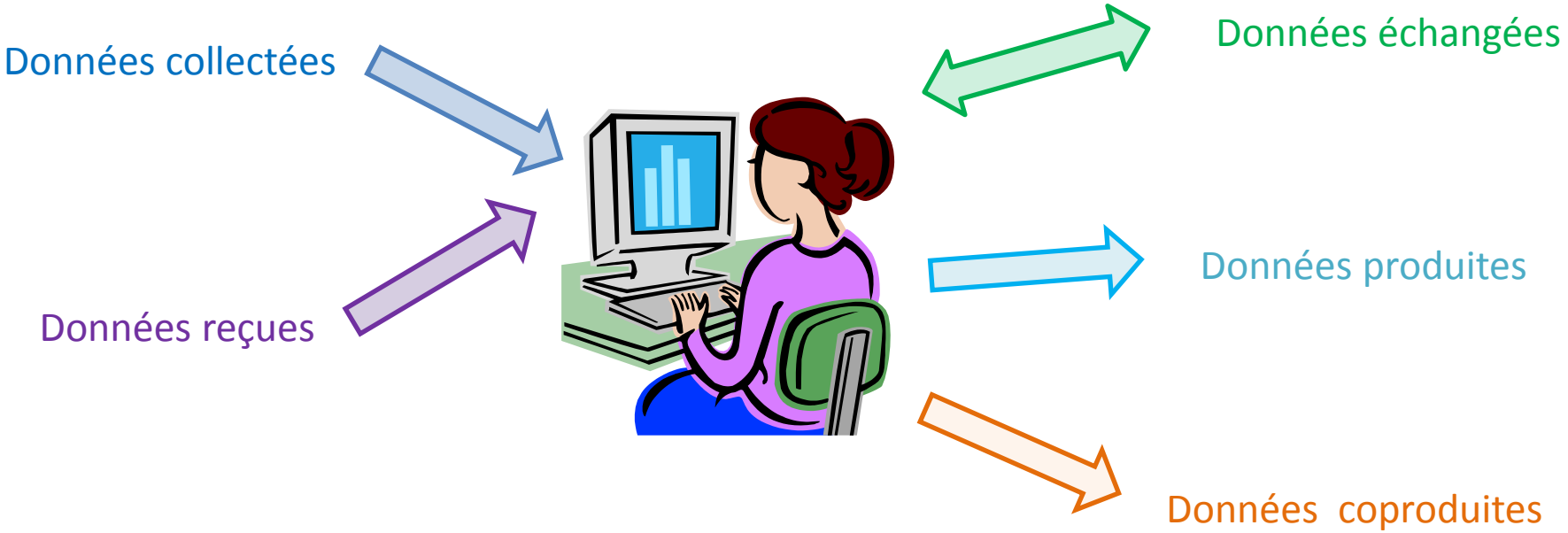
Définir les données de la recherche

Qu’est ce qu’une donnée de recherche ?

Une information factuelle enregistrée sur un support, produite ou collectée, selon divers procédés au cours d’un processus.

Au cours de son processus de recherche,
un chercheur…

… collecte des informations par différents moyens…



Les données de la recherche sont des enregistrements factuels (chiffres, textes, images et sons) produites, coproduites, reçues ou échangées dans le cadre d'un recherche scientifique.

Elles peuvent être nécessaires pour valider des résultats de recherche ou simplement utiles à leur compréhension.

… selon divers procédés parfois propre à une discipline…

Types de données	Définitions	Valeur et spécificité	Exemples
Données d'observation (*Observational data*)	Données obtenues en temps réel	Données souvent uniques et irremplaçables	Données atmosphériques, d'enquêtes, échantillons, neuro-image
Données expérimentales (*Experimental data*)	Données obtenues en laboratoire à partir d'équipements spécifiques	Données souvent reproductibles mais à des coûts dissuasifs	Séquence de génome, chromatographie, spectres RMN
Données de simulation (*Simulation data*)	Données générées à partir de modèles test	Données descriptives du modèle et les métadonnées ont + de valeur que les données de résultats	Modèles climatiques, modèles économiques
Données dérivées ou compilées (*Derived or compiled data*)	Données obtenues par compilation ou traitement des données brutes	Données reproductibles mais à des coûts parfois importants	Texte et *data mining*, bases de données compilées, modèles 3D
Données de référence ou données canoniques (*Reference or canonical data*)	Collections statiques ou organiques de jeux de données validées	Données généralement publiées ou qui ont fait l'objet d'une curation	Banque de données sur le génome, structure chimique, portail de données spatiales

… et stockées sur différents supports.

Documents numériques

Enregistrements vidéo/sonores
Bases de données
Séquences génomiques
Simulation modèles
Algorithmes
Codes informatiques
Traitement de texte
Fiches de lectures

…

Documents numérisés

Articles
Chapitres d'ouvrages
Schémas
Corpus de texte
Photographies

…

Documents papier

Notes manuscrites
Cahiers de laboratoire
Cahiers de manipulations
Tirages photographiques
Ouvrages

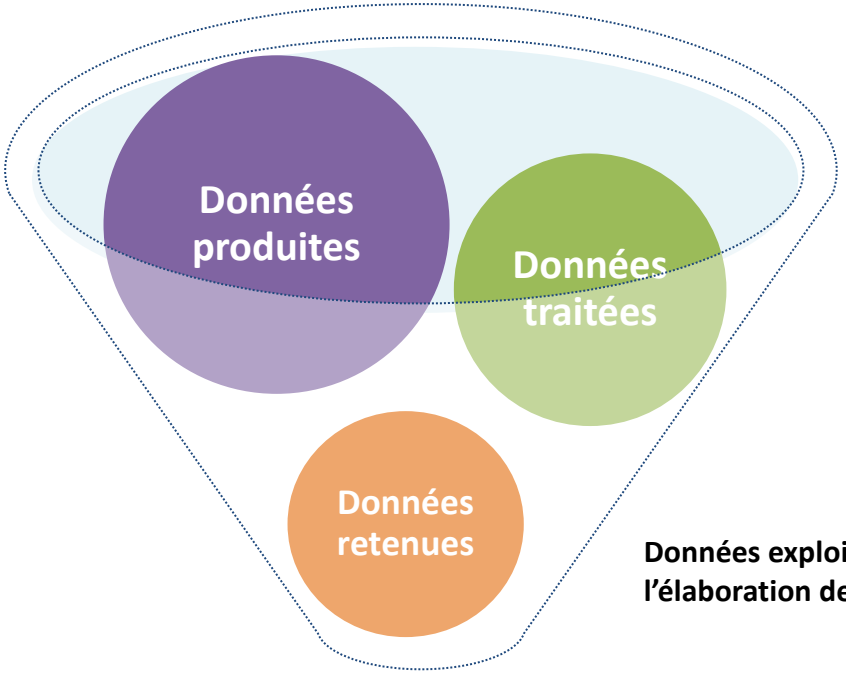
…

Le cycle de vie de la donnée aujourd'hui

Que deviennent les données aujourd'hui ?

Données primaires produites ou collectées au cours de la recherche

INVISIBLE



Données analysées, agrégées, compilées, transformées

Données exploitées pour l'élaboration des résultats

Perte d'information importante au cours du processus de publication

VISIBLE

Données publiées

Données intégrées à titre d'illustrations ou de résultats dans une publication sur support papier ou en ligne

Plusieurs constats

- Une part importante des données produites par la recherche reste invisible
- Les modes de publication des données ne permettent pas toujours leur réutilisation
- Les données ne font pas l'objet d'une stratégie de valorisation propre
- Les données sont rarement identifiées comme des éléments distincts du reste de la publication
- Leur statut juridique est souvent flou aux yeux des producteurs et des utilisateurs
- Leur sort après leur phase d'utilisation courante est bien souvent négligé

Pourtant une bonne gestion des données peut devenir un atout pour la recherche !

=> Le *Data management plan* : anticiper le sort des données

" A DMP **describes the data management life cycle for all data** sets that will be **collected, processed or generated** by the research project. It is a document outlining **how research data will be handled** during the research project, and even after the project is completed, **describing what data will be collected, processed or generated and following what methodology and standards, whether and how data will be shared and/or made open, and how it will be curated and preserved.**" (*définition H2020)*

En bref… le DMP décrit le cycle de gestion de toutes les données qui seront collectées, traitées ou générées par un projet de recherche. Il a pour vocation d'anticiper les problèmes de gestion qui peuvent survenir au cours d'une recherche et les conditions d'une conservation et d'une diffusion future des données.

Le contexte international d'ouverture des données

Les initiatives internationales pour l'ouverture des données

- **Politiques d'ouverture des données**

  - Université d'Edimbourg (Royaume-Uni)
  - Université d'Oxford (Royaume-Uni)
  - Université de Göttingen (Allemagne)
  - Université de Leiden (Pays-Bas)

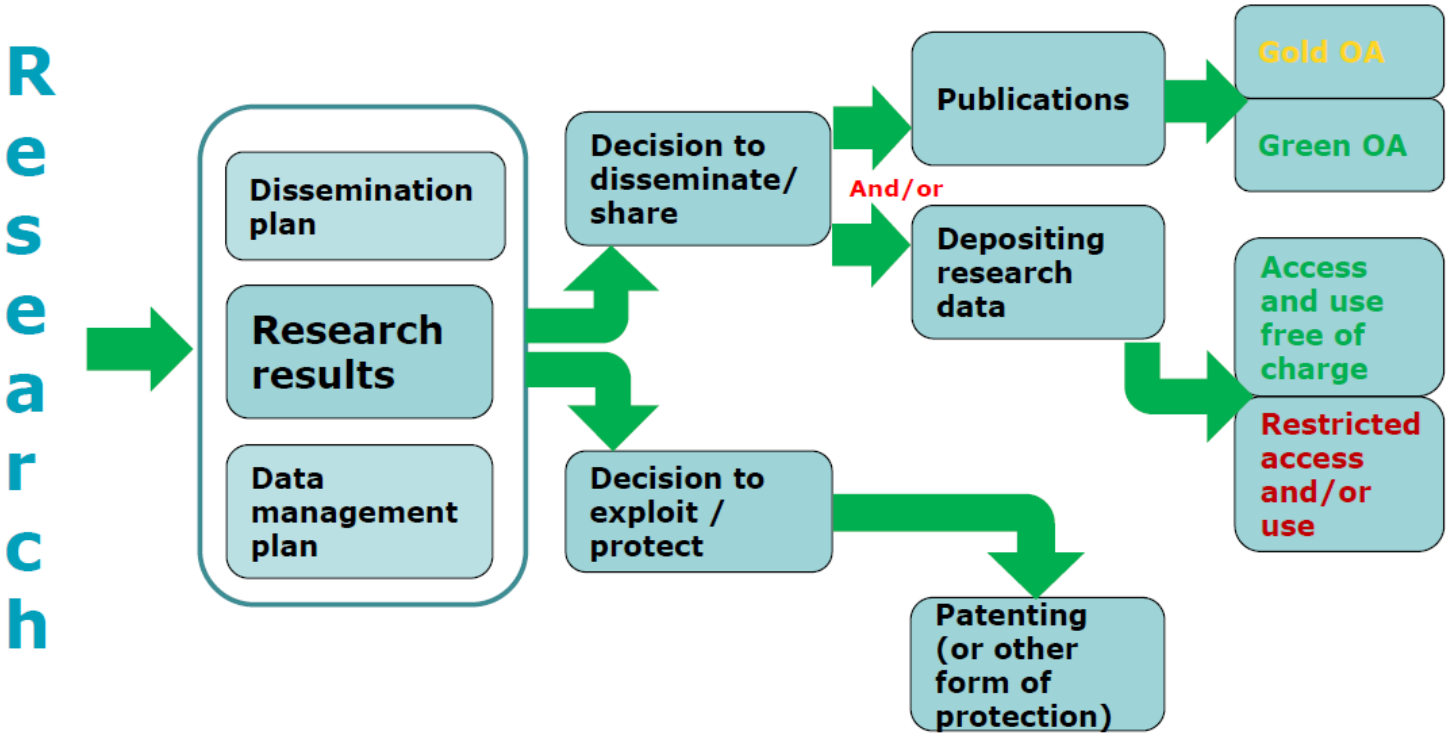
- ***Data management plans*** **:**
  - U.S. National Science Foundation (Etats-Unis)
  - U.S. Department of Energy (Etats-Unis)
  - U.K. Research Councils (Royaume-Uni)
  - The Netherlands Organisation for Scientific Research (Pays-Bas)

Le contexte européen : Horizon 2020

L'*Open Access* dans Horizon 2020

European Commission

OA in context: Dissemination & exploitation of research results



Digital Agenda

L'*Open research data pilot*

Dans le cadre du programme de financement Horizon 2020, la Commission européenne lance un projet pilote qui vise à **favoriser la réutilisation des données issues des projets** de recherche.

Ce pilote concerne :

- Les données et métadonnées associées nécessaires à la validation des résultats présentés dans les publications ;
- Les autres données et métadonnées associées décrites dans le plan de gestion de données.

Le périmètre de l'*Open Research Data Pilot*

EXCELLENCE SCIENTIFIQUE

- Conseil européen de la recherche (ERC)
- Actions Marie Sklodowska-Curie
- Technologies futures et émergentes (FET)
- Infrastructures de recherche :

  - Développer de nouvelles infrastructures ...
  - Intégrer et ouvrir les IR d'intérêt européen
  - Infrastructures électroniques
  - Soutien à l'innovation...

PRIMAUTE INDUSTRIELLE

- Technologies de l'information et de la communication (TIC)
- Technologies clés génériques (KET)
- Espace
- Innovation dans les PME
- Accès au financement à risque

DEFIS SOCIETAUX

- Santé, bien-être, vieillissement
- Sécurité alimentaire, bioéconomie…
- Energies sûres, propres, efficaces :
  - Villes et communautés intelligentes
- Transports intelligents, verts, intégrés
- Climat, environnement, matières premières
- Sociétés inclusives et novatrices et capables de réflexion
- Sociétés sûres

Diffusion de l'excellence et élargissement de la participation

Science pour et avec la société

Institut Européen d'Innovation et Technologie (IET)

Centre commun de recherche / *Joint Research Center (JCR)*

Exigences pour les bénéficiaires

Les bénéficiaires doivent :

- **Déposer dans un entrepôt de données :**
  - Les données et métadonnées nécessaires à la validation des résultats présentés dans les publications
  - Les autres données et métadonnées mentionnées dans le plan de gestion de données (article 29.3 de la convention de subvention)
- **Fournir, par le biais de l'entrepôt de données, des informations sur les outils et le matériel nécessaires à la validation des résultats**
- **Produire dans les 6 premiers mois du projet de recherche un plan de gestion des données** (*data management plan*). Il est appelé à être mis à jour au minimum à mi-parcours et au rapport final. Il permet d'anticiper les modalités ultérieures de dissémination des données générées au cours de la recherche.

Les exceptions de diffusion

Il existe des exceptions au principe de diffusion des données. Le chercheur peut donc justifier de la non-diffusion pour diverses raisons :

- En cas d'incompatibilité avec l'exploitation industrielle et commerciale
- En cas d'incompatibilité avec des questions de sécurité et de confidentialité
- Pour protéger des données personnelles
- Si la diffusion des données risque de compromettre l'objectif du projet
- Si le projet ne collecte ou ne génère aucune donnée
- Pour toute autre raison légitime…

Mais ces exceptions ne dispensent pas de l'élaboration d'un plan de gestion des données.

Les solutions d'appui à H2020

Dissémination

Entrepôt de données scientifiques issu du projet européen ***Open AIRE +*** et développé par le CERN.

Si vous n'avez pas trouvé d'archive thématique appropriée pour vos données, Zenodo peut être utilisé pour le dépôt de vos données scientifiques. Il affecte des DOIs aux objets et expose les métadonnées de description au moissonnage à travers le protocole OAI-PMH, le protocole d'interopérabilité des archives ouvertes (Prodinra, HAL…).

⟶ **Valorisation sur OpenAIRE**

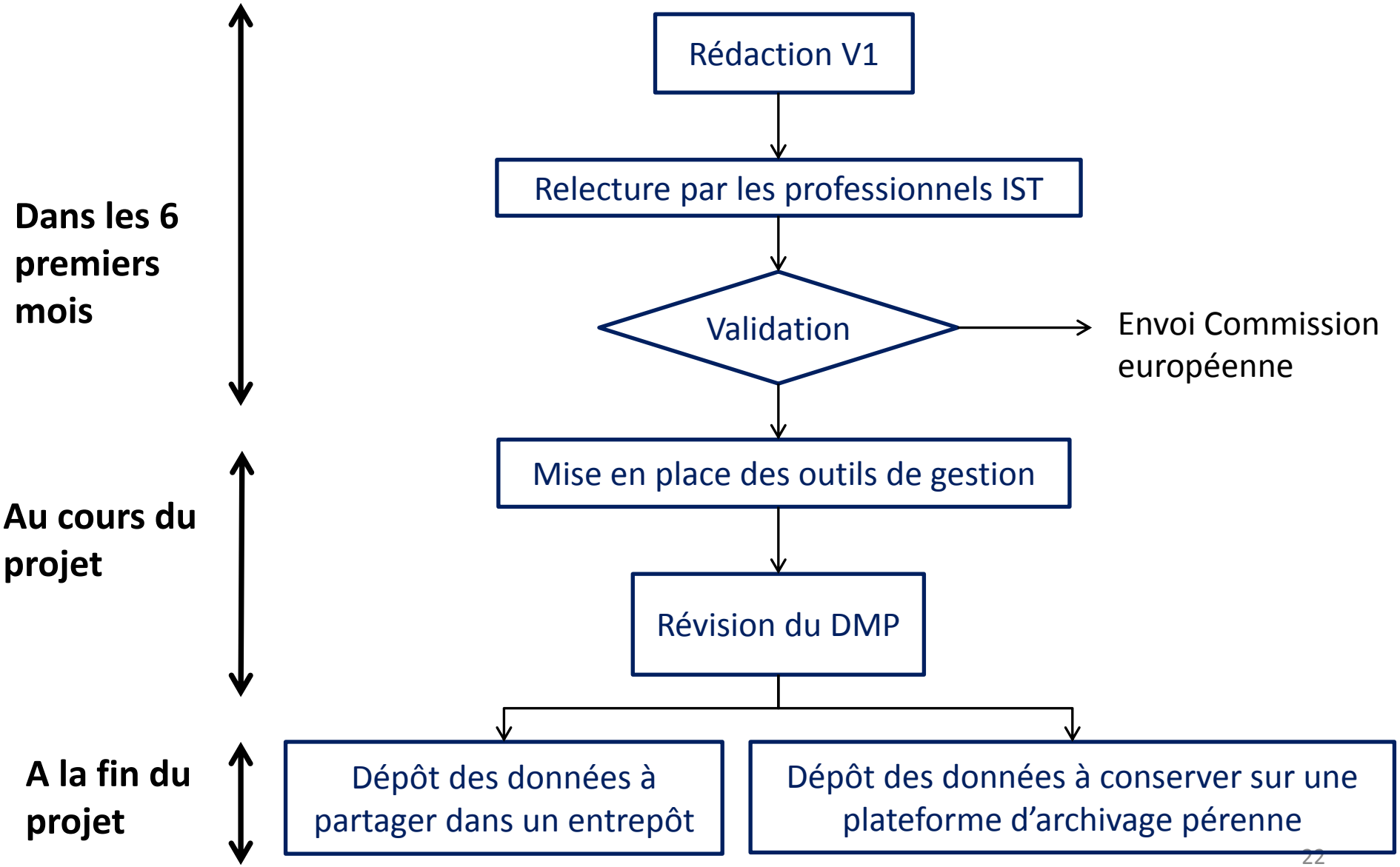
Préservation

Le projet **EUDAT** est une initiative européenne cofinancée par le 7e PCRD. Il vise à répondre aux besoins futurs des chercheurs en matière d'accès et de préservation des données scientifiques : réplication sécurisée des données ; transfert des données vers des centres de traitement ; service d'archivage de petits fichiers ; catalogue de métadonnées ; infrastructure d'authentification et d'autorisation (AAI). Le **CINES** participe au projet européen comme centre de ressources.

L'élaboration d'un plan de gestion de données

Le DMP est un document prospectif et évolutif destiné à être enrichi à mesure de l'avancement du projet.

Les étapes de rédaction et de validation



Les principaux champs du DMP

- ✓ **Informations sur le projet** : responsabilité du projet et des données, ressources et moyens nécessaires
- ✓ **Description des jeux de données :** identification et description
- ✓ **Métadonnées** : documentation et organisation des données
- ✓ **Stockage, accès et sécurité au cours du projet**
- ✓ **Dissémination des résultats** : publication**,** partage, protection des données sensibles
- ✓ **Archivage** : sélection et conservation à long terme

Le DMP est un document prospectif et évolutif destiné à être enrichi à mesure de l'avancement du projet.

A l'échelle du projet

Section [1- 3] - Informations sur le projet et responsabilité du plan de gestion

- ✓ **Objectifs** : informer sur le contexte du projet de recherche et définir le suivi des données
- ✓ **Principaux champs** :
  - Identifiant de l'appel à projets
  - Thématiques
  - Objectifs du projet
  - Propriété des données
  - Ressources nécessaires à la mise en œuvre du DMP

A l'échelle du Jeux de données *(Dataset)*

[Section 4.1] - Description des données

- ✓ **Objectifs** : présenter le type de données qui seront produites et reçues dans le cadre du projet
- ✓ **Principaux champs** :
  - Identifiant du jeu de données
  - Nature des données
  - Méthode de production des données
  - Formats des données

Cette partie peut être dupliquée pour chaque jeu de données.

[Section 4.2] - Stockage, accès et sécurité des données

- ✓ **Objectifs** : définir les modalités d'hébergement, de sauvegarde et d'accès aux données pendant la phase active du projet
- ✓ **Principaux champs** :
  - Volumétrie prévisionnelle
  - Gestion des accès
  - Identification des risques et menaces

  (confidentialité, intégrité, traçabilité, disponibilité)
  - Prévention des risques

[Section 4.3] - Métadonnées : documentation et organisation des données

- ✓ **Objectifs** : préciser la manière dont seront décrites et organisées les données produites ou reçues au cours du projet
- ✓ **Principaux champs :**
  - Standards et formats disciplinaires
  - Responsabilité des métadonnées
  - Règles de nommage des jeux de données
  - Arborescence de classement
  - Autres informations complémentaires

[Section 4.4] Dissémination

- ✓ **Objectifs** : préciser les modalités et les éventuelles précautions éthiques, juridiques et techniques selon lesquelles les données seront diffusées.
- ✓ **Principaux champs** :
  - Principe général de diffusion
  - Protection des données sensibles (*justification*)
  - Licence
  - Dépôt et mise à disposition des données
  - Publications associées
  - Potentiel de réutilisation
  - Embargo

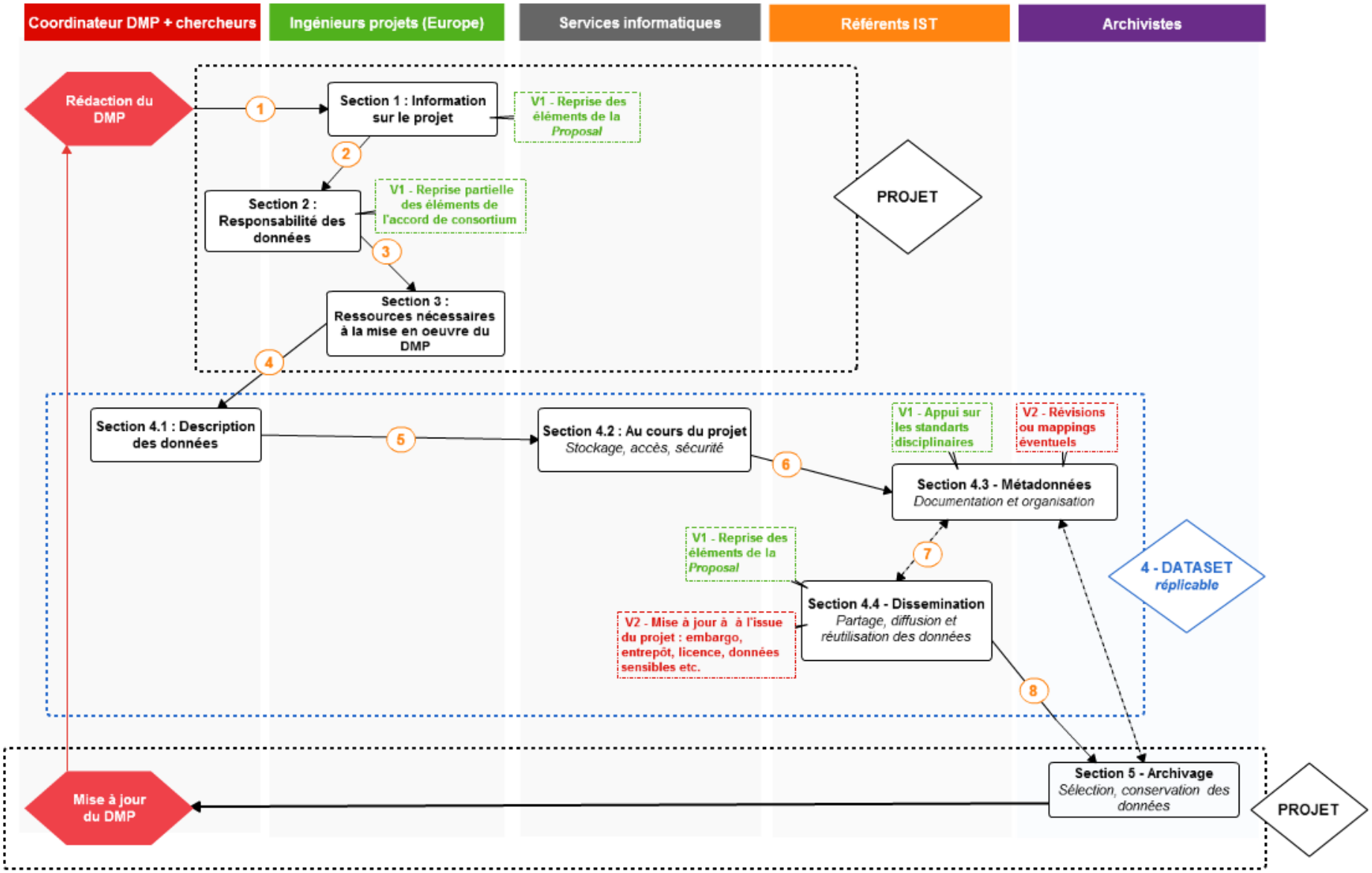
A l'échelle du projet

*[Section 5] - **S**élection et archivage*

- ✓ **Objectifs** : renseigner sur le sort des données à l'issue du projet et les dispositifs de conservation choisis.
- ✓ **Principaux champs** :
  - Sélection des données
  - Sort des données à l'issue du projet
  - Volume final des données
  - Durée de conservation
  - Plateforme d'archivage choisie

Cette section concerne l'ensemble des données produites ou collectées au cours du projet, qu'elles aient été diffusées ou non.

Les acteurs du plan de gestion de données



Les ressources utiles

- Commission européenne. *Recommandation de la commission du 17.7.2012 relative à l'accès aux informations scientifiques et à leur conservation* (17.7.2012) C(2012) 4890 final http://medoanet.sciencesconf.org/conference/medoanet/pages/recommendation_access_and_preservation_scientific_information_fr_copie.pdf
- European Commission . *Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020* http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf
- European Commission. *Guidelines on Data Management in Horizon 2020* http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
- Open Aire. https://www.openaire.eu/guide-for-project-officers-in-horizon-202/view-document
- OCDE. *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics* (2007) http://www.oecd.org/fr/science/sci-tech/38500823.pdf
- Open Access in Horizon 2020 - EC funded projects. *Briefing paper for Project Coordinators* . https://www.openaire.eu/guide-for-project-coordinators-in-horizon-2020/document-details

En conclusion : les données selon le pilote H2020

Détectables

Les données et métadonnées sont-elles détectables et identifiées par un mécanisme de type DOI?

Accessibles

Les données sont-elles accessibles et selon quelles modalités?

Interopérables

Les données permettent-elles l'échange avec des chercheurs, des institutions, des pays étrangers?

Fiables et intelligibles

Les données sont-elles fiables et intelligibles par des pairs ?

Réutilisables

Les données peuvent-elles être réutilisées par des tiers, même longtemps après?

Merci pour votre attention

Aurore Cartier – aurore.cartier@parisdescartes.fr
Magalie Moysan – magalie.moysan@univ-paris-diderot.fr
Nathalie Reymonet - nathalie.reymonet@univ-paris-diderot.fr