

Introduction



Les données au cœur de l'activité de recherche

Une production et une gestion des données en pleine mutation

Des changements de paradigmes importants, au-delà des données de la recherche

Des expertises à la fois très spécifiques et très transverses



**MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE**

*Liberté
Égalité
Fraternité*



Etude des métiers et activités de la donnée

Anne Laurent - Isabelle Blanc





Constats

Les métiers de la donnée sont parmi les plus recherchés
Un manque cruel de formations initiales et continues sur une partie des activités liées aux données

- Activités de gestion-diffusion des données : nettoyage, préparation, description (métiers), partage, circulation, ouverture
- Des activités au sein et orientées vers les métiers qui produisent et utilisent les données
- Ces activités trop souvent apparentées à celles des spécialistes de l'analyse des données

Enjeux pour l'ESR

Développer et reconnaître les **compétences en matière de gestion, préservation, circulation et ouverture des données** pour les chercheurs, les enseignants-chercheurs, les personnels d'appui à la recherche et administratifs tout au long de leur parcours

Concevoir nativement des données de recherche réutilisables

Engagement politique

Au-delà de l'ESR et en tant que ministère en charge de l'Enseignement supérieur :

- Identifier pour la formation initiale : les unités d'enseignement, les filières métiers à développer, renforcer...
- Développer les formations continues et décrire des parcours professionnels
- Soutenir et amplifier les formations continues et initiales existantes



Démarche

Une étude en plusieurs temps :

Premier focus sur les activités autour de la donnée de recherche

- 1. Synthèse bibliographique** sur les activités et métiers clés de la gestion, préservation, diffusion de la donnée (data steward, data librarian, data curator...) => étude bibliographique
- 2. Analyse quantitative et qualitative** de l'existant des pratiques et des trajectoires dans les environnements de l'ESR ayant mis en place des modes de fonctionnements et d'organisation performants
 - Identifier les pratiques de la donnée des répondants
 - Identifier le statut professionnel des répondants

Ensuite inventaire des formations initiales et continues

Elargissement des observations et propositions en interministériels

Modalités

Chargée de l'étude : Anne Laurent, professeur de l'université de Montpellier, Directrice de l'institut de science des données de Montpellier

Groupe de travail de l'étude réunissant des représentants de différentes communautés métiers de l'ESR impliquées dans la gestion des données de la recherche

Prestataire pour accompagner la conduite de l'étude :
Dataactivist

Pilotage étude : MESR Isabelle Blanc

Planning prévisionnel :
Novembre 2022-Décembre 2023



Périmètre

- Portrait de quelques métiers clés selon ce qu'en dit la littérature internationale
- Pas de recherche bibliographique exhaustive : priorité donnée aux revues de littératures et publications post-2018
- Prise en compte de référentiels existants

Premiers constats

- Pas d'étonnement
- Beaucoup de littérature sur des métiers et activités bien décrits par ailleurs et ne constituant pas le cœur de notre étude (data scientist/analyst/engineer)
- Une distinction à faire entre métiers/activités/compétences

Méthode

- Liste des activités
- Placement sur le cycle de vie de
 - La nature des données (en cours de traitement, validées, citables)
 - L'endroit d'exercice des activités (labos/platformes, structures centrales, ...)
- Besoin de formulations non ambiguës



Données de la recherche OCDE 2007

Les « données de la recherche » sont définies comme des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche

Métadonnées

National Information Standards Organization (NISO), 2004

« Une métadonnée est une information structurée qui décrit, explique, localise une ressource informationnelle ou la rend facile à trouver, utiliser ou gérer »

La NISO distingue 3 principaux types de métadonnées : descriptives, structurelles et administratives.

Borgman, C.L., 2020

Métadonnées scientifiques : description protocole obtention, traitement, matériel, logiciels...

Métadonnées documentaires : administratives, d'utilisation, de préservation, d'accès



**MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE**

*Liberté
Égalité
Fraternité*

CE QUE DIT LA REVUE DE LITTÉRATURE

Unité de recherche
Infrastructure de recherche
Service central
Unité/infra de recherche

Coordonner les activités de gestion et de diffusion des données

Garantir la qualité des données et leur conformité de diffusion par rapport aux cadres législatifs

Superviser l'ensemble du cycle de vie

Décrire et mettre en œuvre le protocole d'obtention des **données initiales - brutes**

Découvrir les données

Nettoyer, affiner, convertir et décrire les données brutes (par des métadonnées scientifiques)

Structurer les données et les outils pour la collecte, le stockage et le traitement

RÉUTILISATION

**PLANIFICATION
PLAN DE GESTION
DE DONNÉES**

**COLLECTE
CRÉATION
STOCKAGE**

Analyser/Représenter/Modéliser/Prédire (stat, algo, machine learning) => **données de sortie de l'analyse**

Accompagner la découverte des données

Données citables

Données en cours de traitement

**ANALYSE
TRAITEMENT
CALCUL**

**PUBLICATION
POUR
OUVERTURE
OU PARTAGE
(accès restreint)**

Publier en ouverture ou en partage restreint

Données validées

Unité/infra de recherche

**PRÉPARATION
DU DÉPÔT**
Structuration
et description

Créer, structurer, décrire les **jeux données validées** (ou aide à création, structuration, description)

DÉPÔT

Assurer la mise à jour des versions de jeux de données

Renseigner les métadonnées documentaires (contributeurs, établissements, licences, financeurs projets, ...)

Déposer les jeux de données Choisir (ou aider au choix) de l'entrepôt

Unité/infra de recherche
Unité de recherche
Infrastructure de recherche
Service central

Accompagner la découverte
des données
Publier en ouverture ou
en partage restreint

Assurer la mise à jour
des versions de jeux de données
Déposer les jeux de données

Coordonner les activités de gestion et
de diffusion des données

Garantir la qualité des données et leur conformité de
diffusion par rapport aux cadres législatifs

Superviser l'ensemble du
cycle de vie

Découvrir
les données

Décrire et mettre en œuvre le protocole d'obtention
des données brutes

Nettoyer, affiner, convertir et décrire les données
brutes (par des métadonnées scientifiques)

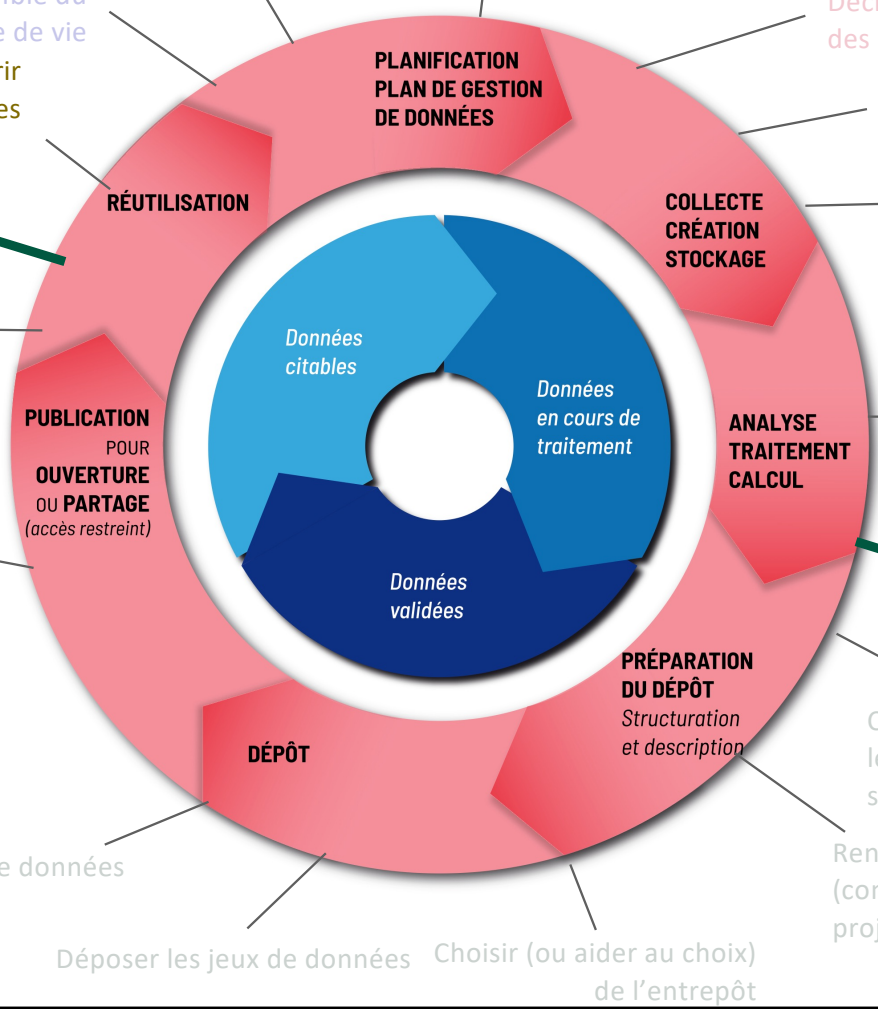
Structurer les données et les outils pour la
collecte, le stockage et le traitement

**Analyser/Représenter/Modéliser/Prédire (stat,
algo, machine learning) => données de sortie
de l'analyse**

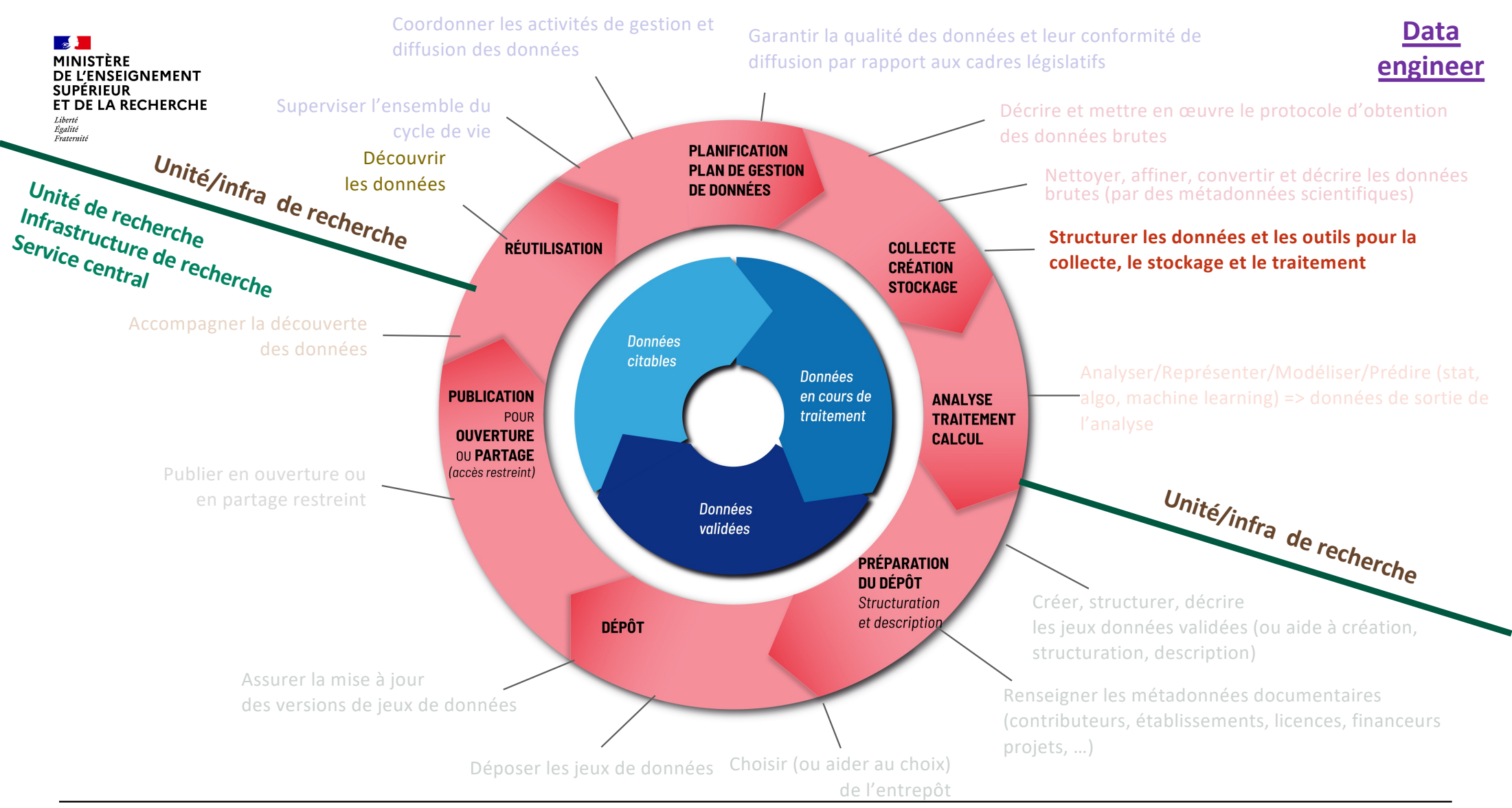
Créer, structurer, décrire
les jeux de données validées (ou aide à création,
structuration, description)

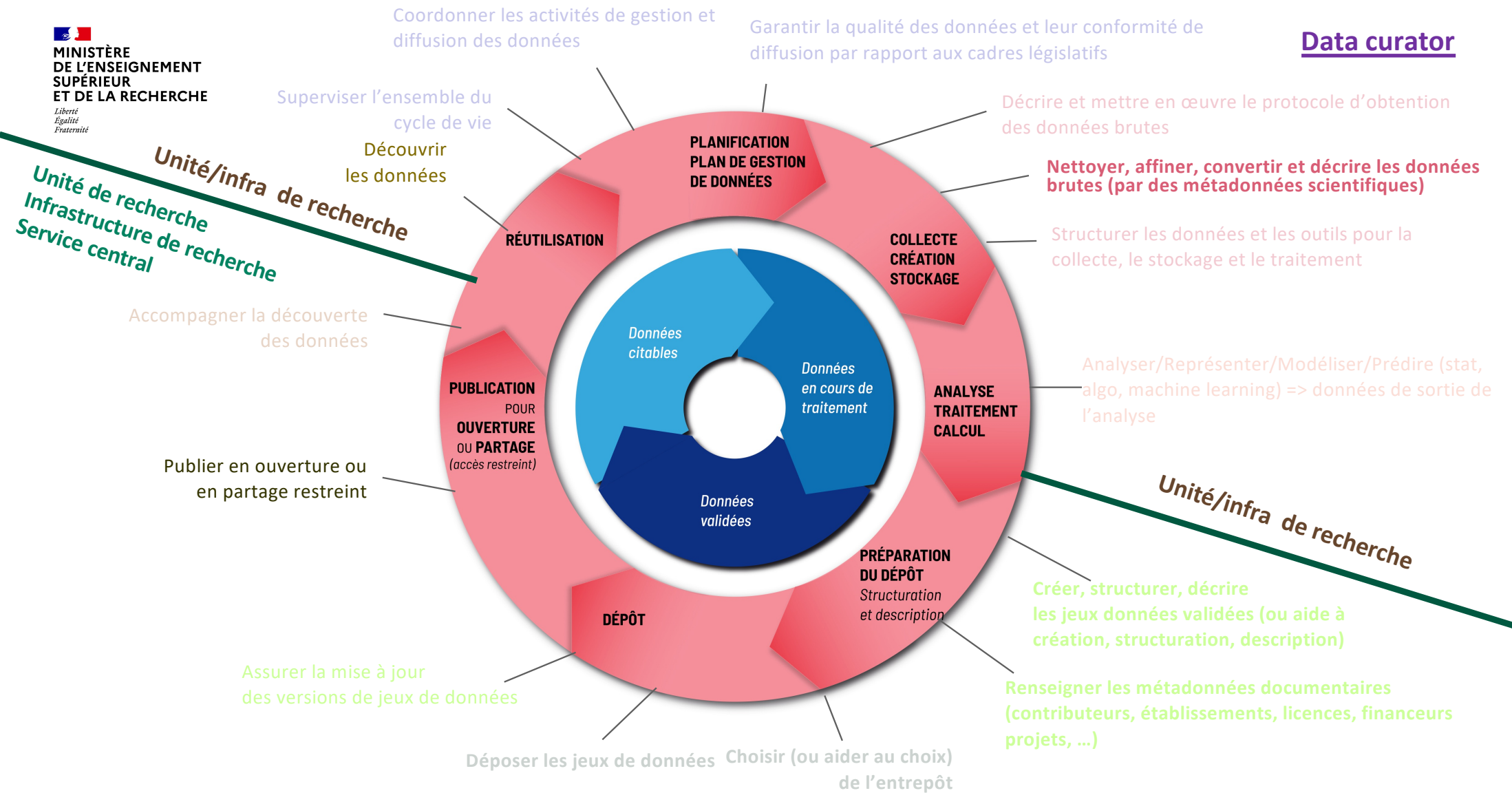
Renseigner les métadonnées documentaires
(contributeurs, établissements, licences, financeurs
projets, ...)

Choisir (ou aider au choix)
de l'entrepôt



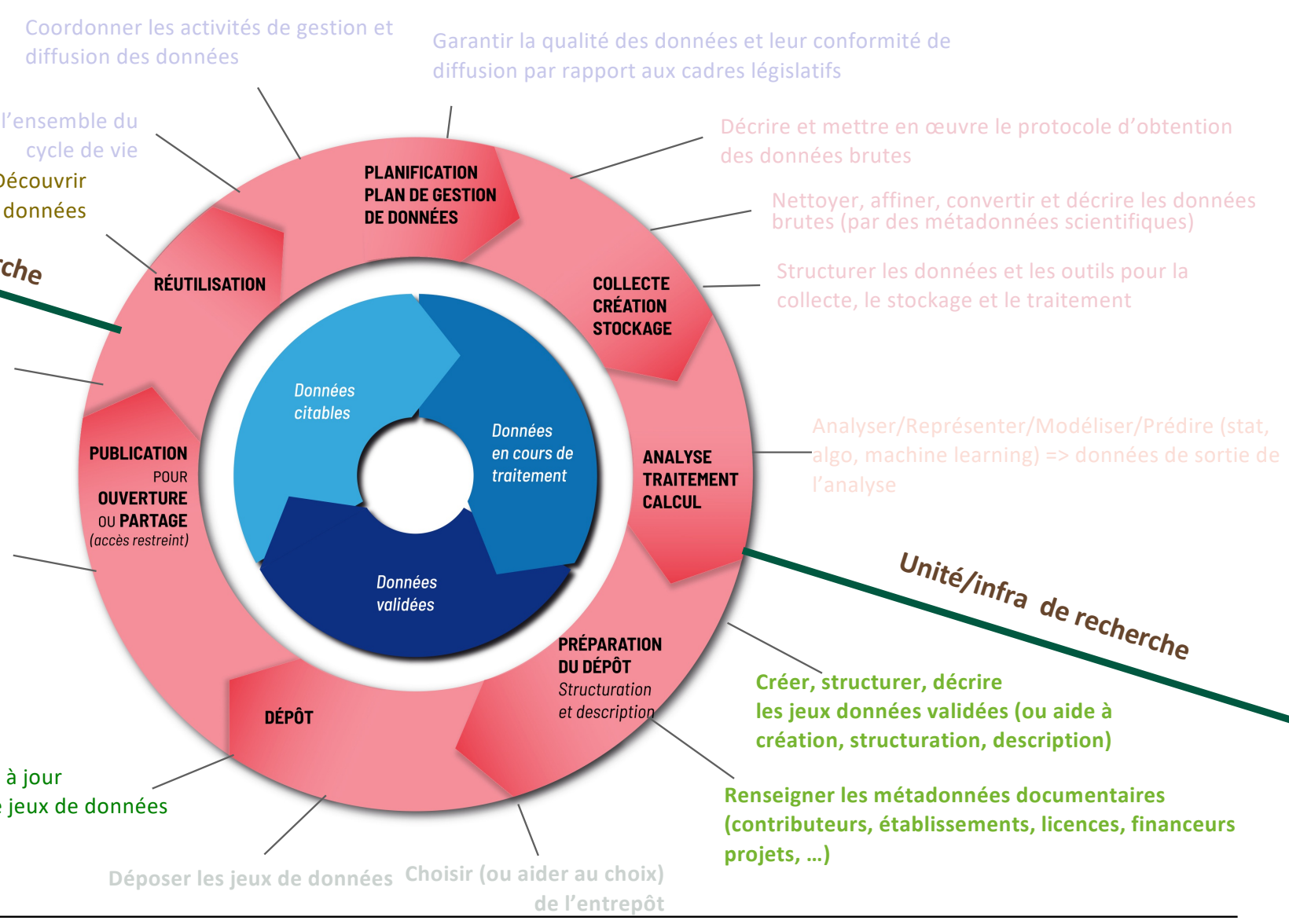
Unité/infra de recherche





Unité/infra de recherche
 Unité de recherche
 Infrastructure de recherche
 Service central

Méta-Data curator



Unité/infra de recherche
 Unité de recherche
 Infrastructure de recherche
 Service central

Coordonner les activités de gestion et de diffusion des données

Garantir la qualité des données et leur conformité de diffusion par rapport aux cadres législatifs

Superviser l'ensemble du cycle de vie

Décrire et mettre en œuvre le protocole d'obtention des données brutes

Découvrir les données

Nettoyer, affiner, convertir et décrire les données brutes (par des métadonnées scientifiques)

Structurer les données et les outils pour la collecte, le stockage et le traitement

RÉUTILISATION

**COLLECTE
CRÉATION
STOCKAGE**

Analyser/Représenter/Modéliser/Prédire (stat, algo, machine learning) => données de sortie de l'analyse

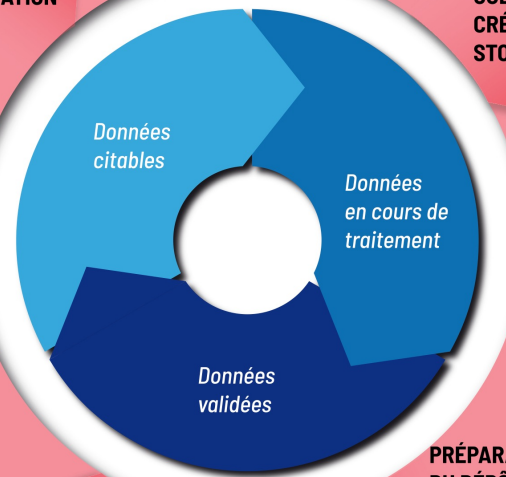
Accompagner la découverte des données

**PUBLICATION
POUR
OUVERTURE
OU PARTAGE
(accès restreint)**

**ANALYSE
TRAITEMENT
CALCUL**

Publier en ouverture ou en partage restreint

Unité/infra de recherche



Data librarian

**PRÉPARATION
DU DÉPÔT**
Structuration et description

Créer, structurer, décrire les jeux de données validées (ou aide à création, structuration, description)

Assurer la mise à jour des versions de jeux de données

DÉPÔT

Renseigner les métadonnées documentaires (contributeurs, établissements, licences, financeurs projets, ...)

Déposer les jeux de données Choisir (ou aider au choix) de l'entrepôt

Unité/infra de recherche
 Unité de recherche
 Infrastructure de recherche
 Service central

Unité/infra de recherche

Coordonner les activités de gestion et de diffusion des données

Garantir la qualité des données et leur conformité de diffusion par rapport aux cadres législatifs

Décrire et mettre en œuvre le protocole d'obtention des données brutes

Nettoyer, affiner, convertir et décrire les données brutes (par des métadonnées scientifiques)

Structurer les données et les outils pour la collecte, le stockage et le traitement

Analyse stat, algo, représentation des données, modélisation, machine learning, analyse prédictive => données de sortie de l'analyse

Créer, structurer, décrire les **jeux de données validées** (ou aide à création, structuration, description)

Renseigner les métadonnées documentaires (contributeurs, établissements, licences, financeurs projets, ...)

Déposer les jeux de données Choisir (ou aider au choix) de l'entrepôt

Assurer la mise à jour des versions de jeux de données

Publier en ouverture ou en partage restreint

Accompagner la découverte des données

Découvrir les données

Superviser l'ensemble du cycle de vie

RÉUTILISATION

PLANIFICATION
PLAN DE GESTION
DE DONNÉES

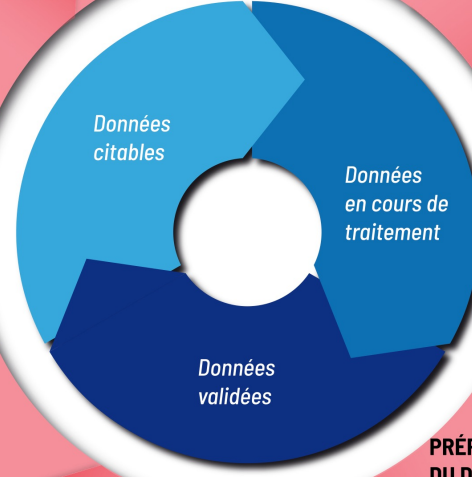
COLLECTE
CRÉATION
STOCKAGE

ANALYSE
TRAITEMENT
CALCUL

PRÉPARATION
DU DÉPÔT
Structuration
et description

DÉPÔT

PUBLICATION
POUR
OUVERTURE
OU PARTAGE
(accès restreint)





**MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE**

*Liberté
Égalité
Fraternité*

ENQUÊTE PAR QUESTIONNAIRE



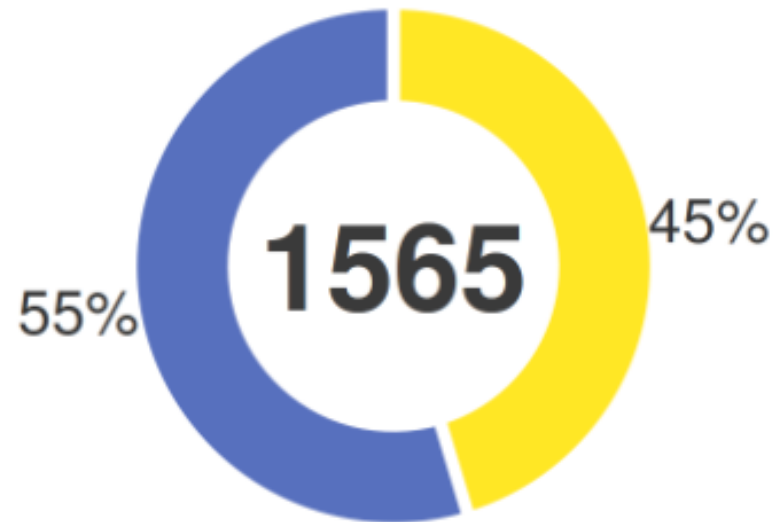
Objectifs



- Destinée à toutes les personnes ayant des activités en lien avec les données de la recherche
- Vise à identifier
 - Les pratiques de la donnée des répondants
 - Lieu de réalisation des activités
 - Liste des activités
 - Rôle dans les activités (« je réalise » / « je coordonne »)
 - Poids des activités
 - Le statut professionnel des répondants
 - Les parcours et les dynamiques



Premiers résultats

1565 répondants



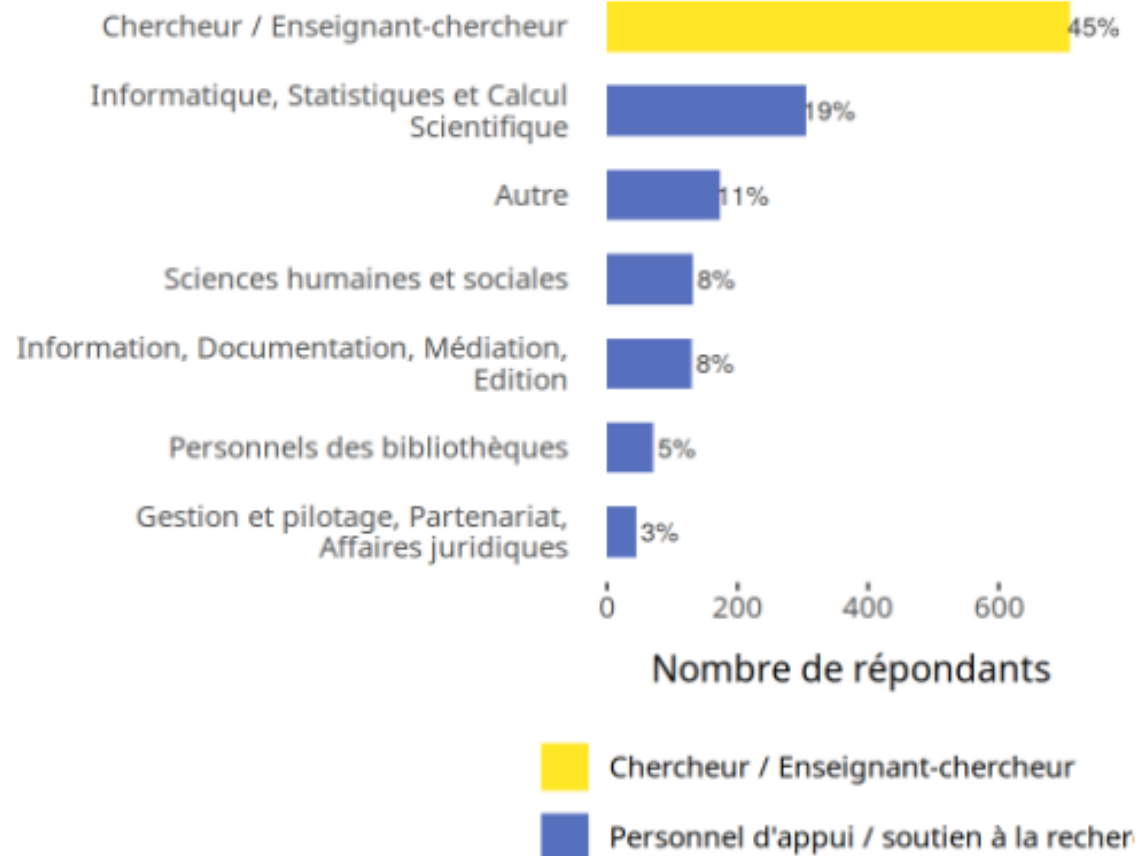
-  Chercheur / Enseignant-chercheur
-  Personnel d'appui / soutien à la recherche

ÉTUDE SUR LES MÉTIERS DE LA DONNÉE Enquête par questionnaire



Premiers résultats

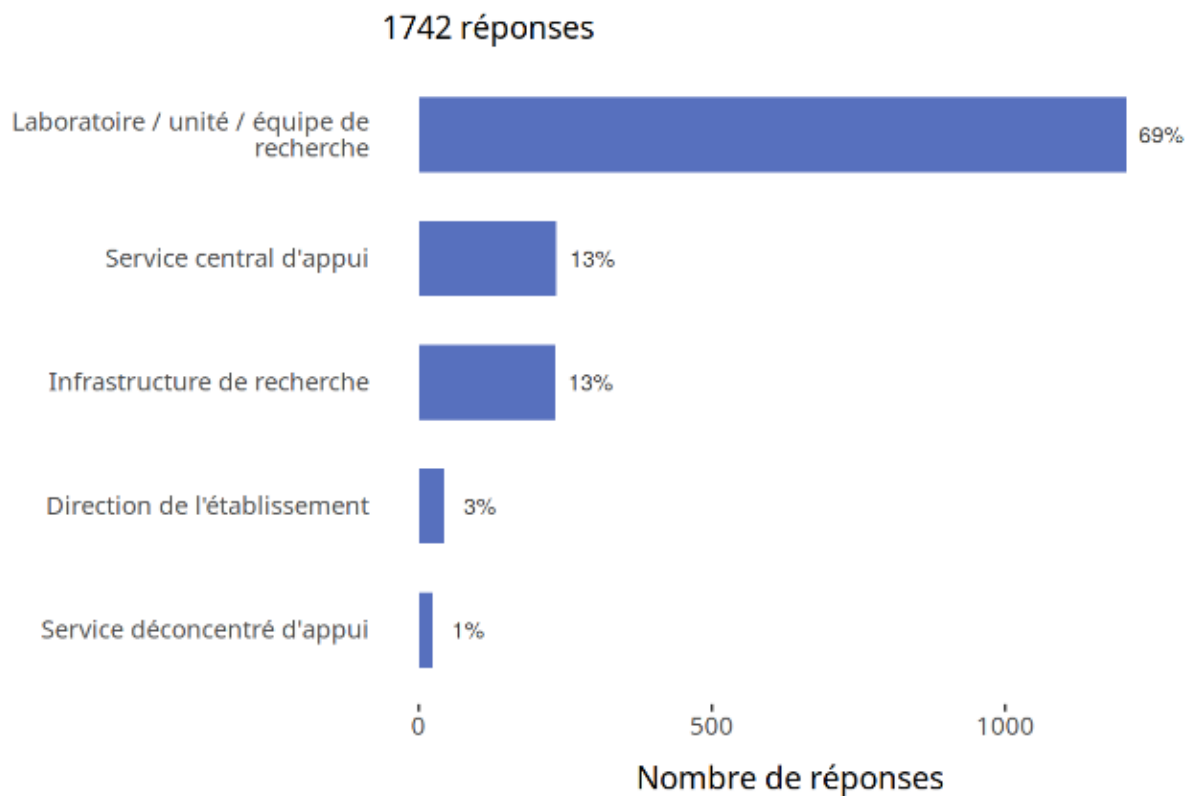
1565 répondants



ÉTUDE SUR LES MÉTIERS DE LA DONNÉE
Enquête par questionnaire

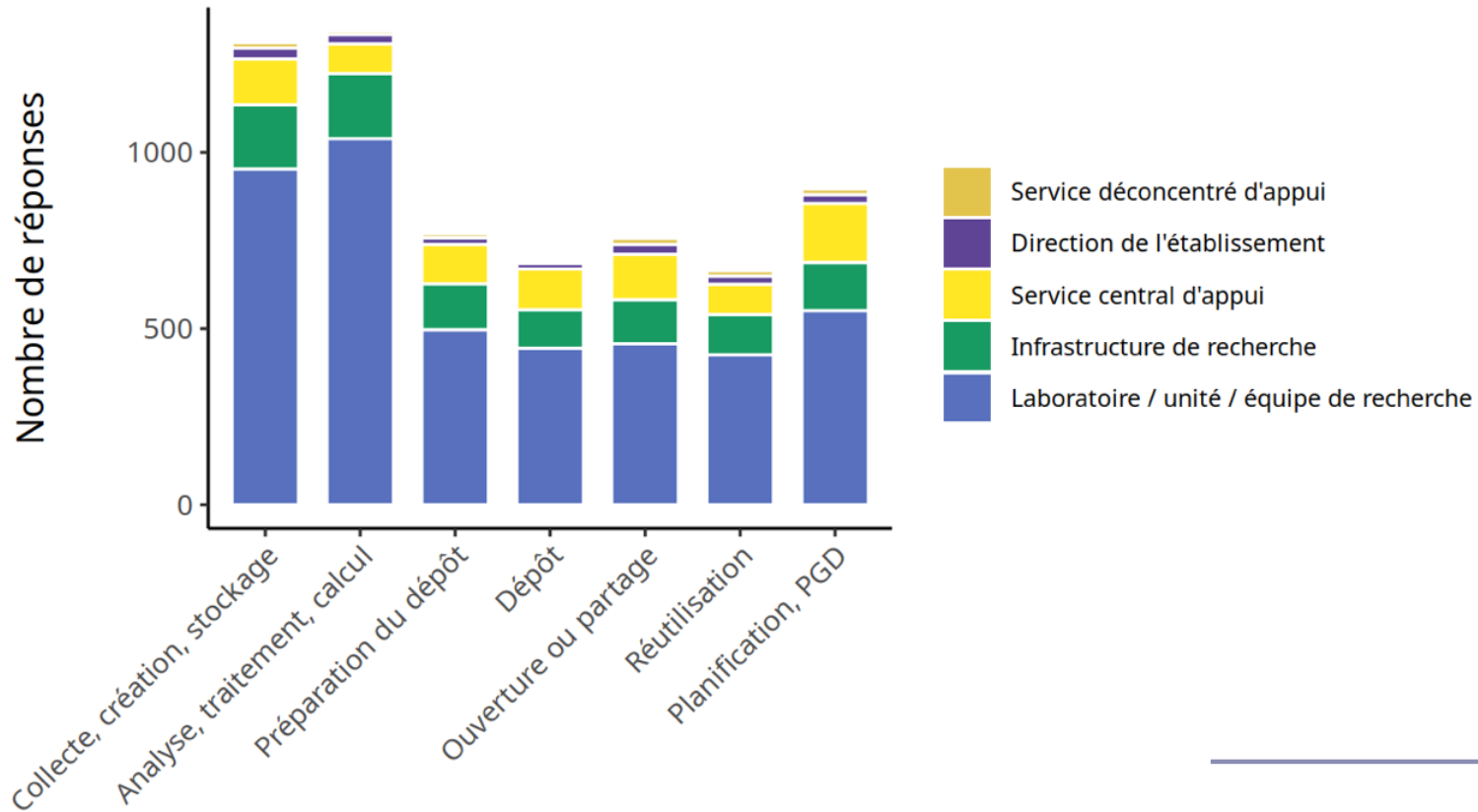


Où réalisez-vous votre activité en lien avec les données de recherche ?



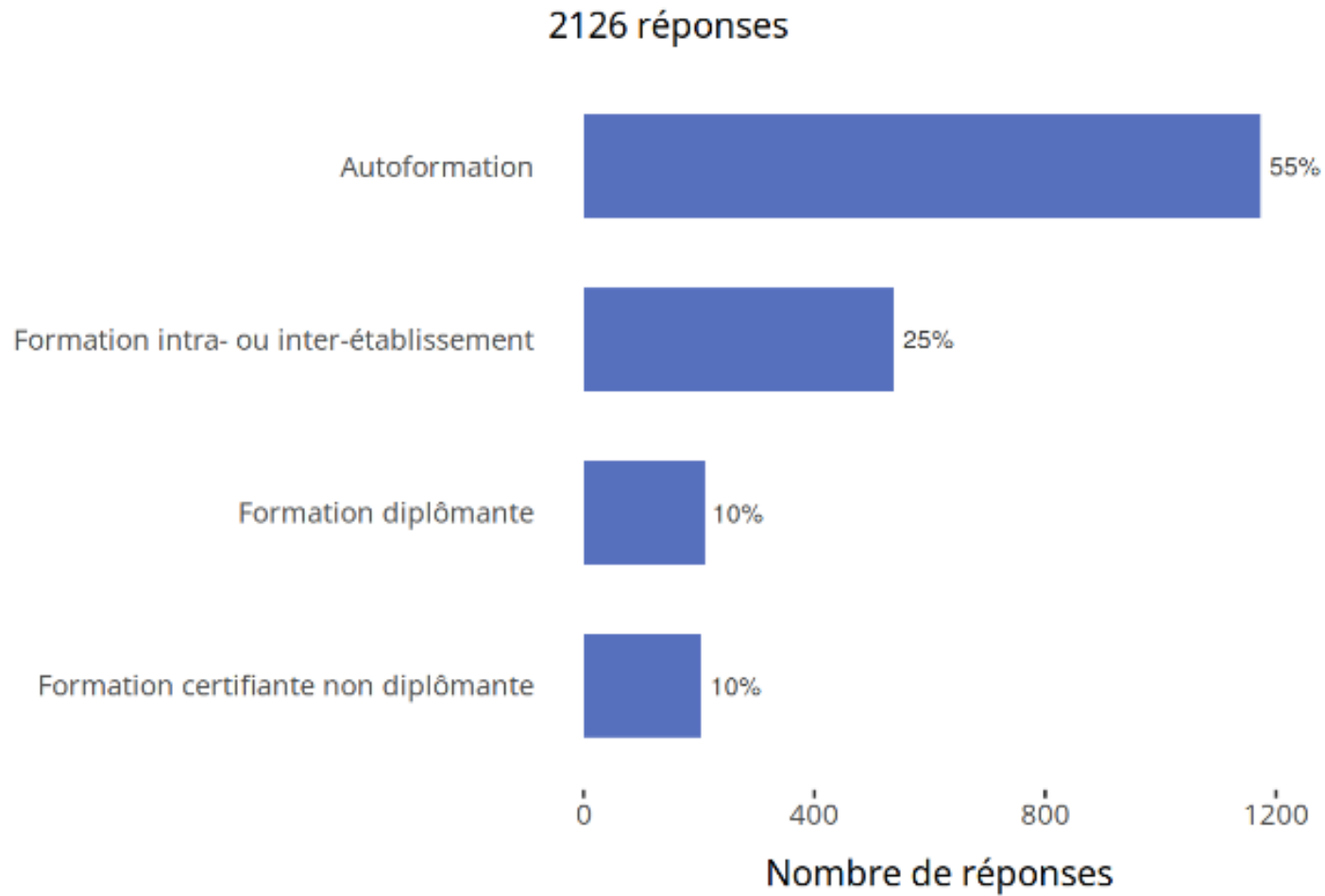


**Lieux d'activités selon l'intervention
dans le cycle de vie de la donnée**
6434 réponses



ÉTUDE SUR LES MÉTIERS DE LA DONNÉE
Enquête par questionnaire

**Comment vous êtes-vous
formé à la gestion des
données ?**





**MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE**

*Liberté
Égalité
Fraternité*

ENQUÊTE PAR ENTRETIENS



Objectifs

Approfondir la réalité des activités et des parcours

- Position dans l'organisation
- Tâches et compétences liées aux données
- Parcours professionnel (y compris parcours de formation/auto-formation)

Modalités

- Entretien avec un panel large de 30 acteurs représentatifs, notamment les volontaires de l'étude par questionnaire



Retours et Verbatims

Illustration des cas généraux et particuliers relevés par l'enquête par questionnaire
Précisions sur certains points, notamment la formation

Des difficultés...

*"Chaque chercheur est très attaché à ses données. Donc **convaincre les collègues de diffuser les données, ce n'est pas facile**. Alors certains collègues prennent le train, mais d'autres pas du tout. Donc j'essaye... J'interviens en AG par exemple. Dès que je peux je prends la parole." - personnel d'appui (gestion de bases de données), université, laboratoire de géographie de l'environnement*

***"Une des difficultés au sein de mon projet est de le coordonner. C'est pas simple à motiver les troupes, ça demande une énergie folle."** - personnel d'appui (gestion de bases de données), université, laboratoire de géographie de l'environnement*

*"Dans l'équipe [en charge des données de l'UMR, d'environ 8 personnes] **on a des techniciens, des ingénieurs de recherche, des scientifiques**, donc avec des profils différents qui ont l'habitude de ce système [de partage des données] et qui peuvent l'expliquer de façon non numérique, non informatique à des gens dont la problématique est de résoudre un problème et pas de comprendre la technique." - personnel d'appui (stockage et partage de données, développement logiciel), organisme de recherche, laboratoire environnement/agronomie/écologie*



**MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE**

Liberté
Égalité
Fraternité

De très grands besoins de formation...

*"[dans le cadre d'un réseau national de spécialistes de la collecte et traitement des données environnementales] on va faire évoluer le programme de la formation, qui ne s'adresse pas à des spécialistes de l'informatique et de l'électronique mais plutôt à des biologistes qui font de l'instrumentation. Donc on va essayer de leur simplifier, de leur donner des outils et des méthodes pour qu'ils puissent arriver à faire de la mesure **sans devenir des spécialistes de la mesure, mais qu'ils arrivent à répondre à leurs besoins rapidement.**"* – personnel d'appui (collecte et prétraitement de données d'observation), organisme de recherche, laboratoire environnement/agronomie/écologie

*"**Il y a des compétences plus avancées que je n'ai pas.** Par exemple, je pense à tout ce qui concerne l'encodage en XML mais aussi pour tout ce qui est des données archéologiques, des compétences en SIG. [...] **Au besoin, on fait appel à des ingénieurs, à d'autres collaborateurs qui ont ces compétences-là.**"* – chercheuse (collecte, mise en qualité et partage des données), organisme de recherche, laboratoire d'archéologie

*"On a de **grosses difficultés pour toucher les doctorants.** Le fait d'être obligé de passer par le catalogue des écoles doctorales ça complique. [...] Et les doctorants ça les intéresse pas de suivre une formation qui n'est pas validée. C'est un frein important."* – personnel d'appui (bases de données et web sémantique), Université/CNRS, laboratoire d'archéologie

*"Toutes les promesses de l'intelligence artificielle au sens large vont amener des demandes sur le traitement d'image et tout ça. **Il y a un besoin de formation pour un besoin d'efficacité.** Certains projets qui nous sollicitent sur ça, quand je vois le temps que ça va nous prendre... Rien que pour savoir vers où s'orienter. Mais c'est du délire. Donc avoir des formations pour savoir vers où aller ou comment gérer ce genre de problématique. Moi me semble urgent et indispensable."* – personnel d'appui (stockage et partage de données, développement logiciel), organisme de recherche, laboratoire environnement/agronomie/écologie



**MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE**

*Liberté
Égalité
Fraternité*

Mais des évolutions positives !

*“Le discours [sur les données de la recherche] a changé et c'est **bénéfique, on peut s'appuyer sur le ministère et l'ANR** pour obtenir des arbitrages. Le support hiérarchique est essentiel, si on ne l'a pas ça peut être un frein énorme.”* – personne d'appui (méthodologie d'analyse et ouverture des données), université, laboratoire de sciences de l'univers

*“L'aspect appui à la recherche est **très intéressant** parce qu'**on sort vraiment de l'idée qu'on peut se faire des métiers de bibliothécaires**. C'est vraiment assez atypique et original. Et comme on a vraiment beaucoup de demandes sur les données de la recherche, **on a l'impression d'être vraiment utile**. Enfin, moi j'ai l'impression de faire un travail utile, parce je vois qu'il y a un vrai besoin et en général [les chercheurs] sont satisfaits, ils sont contents parce qu'on a réussi à les aider. Ils me remercient et ça fait du bien.”* – personnel d'appui (accompagnement à la gestion des données de recherche), université, service de bibliothèques



**MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE**

*Liberté
Égalité
Fraternité*

AVANCEMENT ET LIVRABLE(S)



Avancement

- Synthèse bibliographique sous la forme d'un rendu mettant en lumière les activités et leur déclinaison le long du cycle de vie de la donnée ✓
- Approfondissement de l'étude statistique du questionnaire ✓
- Analyse et synthèse des entretiens ✓ *
- Description de 4 ensembles d'activités sur le cycle de vie, lieu d'exercice et compétences 🕒
- Travail sur les formations et les parcours 🕒

* Données et scripts disponibles

Un groupe de travail très actif
Enrichissement du travail au fil des mois



Livrables à venir

Travail de valorisation des travaux menés sous trois formats

- **Fiches synthétiques** des étapes du cycle de vie, des activités et des intervenants à chaque étape, avec leurs profils et compétences propres
- **Fiches métiers** permettant une meilleure connaissance et reconnaissance des activités et des parcours de formation
- **Vidéos** permettant de présenter les différents acteurs et de mieux comprendre les activités de chaque intervenant dans les étapes du cycle de vie

MERCI A TOUS LES PARTICIPANTS DU GROUPE DE TRAVAIL

Frédéric De Lamotte	Collège données, INRAE, IFB, centre ref thématique
Céline Hernandez	Collège données, CNRS
Christine Hadrossek	Collège données, DDOR, CNRS
Joanna Janik	Collège compétences, CNRS
Johann Berti	Collège compétences, AMU, ADBU
Véronique Stoll	Collège données, Observatoire de Paris
Pierre-Yves Arnould	Collège données, CNRS, Univ. Lorraine

Yvan Le Bras	PNDB, MNHN, centre ref thématique
Stéphanie Cheviron	BU, Univ. Strasbourg
Mariannig Le Behec	GIS URFIST, Univ. Lyon Centre de ressources compétences Recherche Data Gouv
Gabriel Gallezot	GIS URFIST, Univ. Nice Centre de ressources compétences Recherche Data Gouv
Sylvie Rousset	DDOR, CNRS
Zoé Ancion	ANR
Fabien Borget	CEDRE, AMU
Sandrine Gropp	GT montée en compétences ateliers de la donnée, Univ. Montpellier
Juliette Dibie-Barthelemy	DRH, INRAE
Sylvie Damy	Univ. Franche Comté
Cyril Heude	Sciences Po
Soizick Lesteven	Centre de Données astronomiques de Strasbourg
Grégoire Rey	Inserm, France Cohortes



**MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE**

*Liberté
Égalité
Fraternité*

QUESTIONS