CELLULE DATA
GRENOBLE ALPES

Les actus de la cellule

Séminaire CDGA, 25 mai 2022

Rappel : la CDGA

- Une **structure opérationnelle** pour :

  - Répondre **concrètement** à tous les questionnements des scientifiques
  - Fournir un **point d'entrée unique** aux communautés
  - Constituer et animer un réseau de **référents** autour des données pour chaque laboratoire
  - Mettre en place les **outils, services et infrastructures** répondant aux besoins des communautés
  - **Animer, former** (chercheurs, personnels techniques, doctorants...) sur les thématiques liées aux données de la recherche
  - Faire de la **veille juridique et technique**, et s'inscrire dans les initiatives nationales, européennes et internationales

Des compétences complémentaires : **techniques, science ouverte, juridique**.
21 membres (UGA, CNRS, CEA, INRIA) issus de :

- GRICAD
- BAPSO
- Labos (dont CEA)
- SED INRIA
- + la DPD (Déléguée à la Protection des Données) du site

Quelques actions de la CDGA

- Un **point d'entrée unique** :
  - Site web : https://scienceouverte.univ-grenoble-alpes.fr/
  - Support : sos-data@univ-grenoble-alpes.fr

- Des **formations et séminaires**
  - Formation CED
  - Séminaires : stockage, PGD

- Une **enquête** sur les besoins et usages autour des données de la recherche sur le site → **voir plus loin**

- La participation active au projet **Recherche Data Gouv** et une **collection UGA** hébergée sur cette plateforme → **voir plus loin**

- La mise en place d'un **réseau de référents données** dans les laboratoires → **liste juste après**

- Un **accompagnement au quotidien** sur toutes les questions autour des données : relecture / aide à la rédaction de PGD, accompagnement sur le stockage, sur les architectures techniques, questions juridiques, …

Liste des laboratoires avec référent

G-SCOP	EDYTEM	ISTERRE	LARHRA	Institut Fourier
CERMAV	TIMC-IMAG	AAU / ENSAG	LEGI	IPAG
ICMG	SyMME	SENS	IRIG / BGE (CEA)	INRIA
LECA	UMPA / ENS Lyon	LBFA	LIG	ISTERRE
LITT&ArTS	HP2	Pacte	MSH Alpes	IRIG / MEM (CEA)
LiPhy	GIPSA-Lab	LIDILEM	Institut Néel	IRIG (CEA)
GIN	INRAE	IGE	IGE	
AE&CC / ENSAG	Pacte	INRIA	Pacte	
LJK	OSUG	LMGP	DCM	
GRESEC	IAB	SIMAP	HP2	

Si votre labo n’apparaît pas : contactez nous !!

Les actions à venir

- **Séminaires**
  - 10 juin : tutoriel sur DMP Opidor
  - A venir : séminaire axé direction de thèse : comment accompagner les doctorants ?
- **Open Science Days@UGA**
  - Conférence du 13 au 15 décembre
  - Souhait de rendre cet évènement récurrent annuel
  - Thématique 2022 : **Codes et logiciels**
- **Référencement des développements** des laboratoires
  - Travail en cours sur des labos pilotes
  - Travail avec Software Heritage et HAL

Enquête sur les besoins et usages – site Grenoble Alpes

CELLULE DATA
GRENOBLE ALPES

Rapport sur l'enquête sur les usages et les besoins pour la gestion des données de la recherche sur le site de l'Université Grenoble Alpes

Contexte et objectifs

- Convergence de grands enjeux liés :
  - à l'**augmentation exponentielle** des données dans tous les domaines scientifiques
  - à l'amplification du mouvement d'**ouverture de la science**
- Modifications profondes des **processus** de recherche, et des **méthodologies** de travail des communautés scientifiques
- **Bonne gestion des données de la recherche** : une nécessité pour faire face à ces défis :
  - facilite le travail de recherche, individuel et collectif,
  - sécurise les fruits de la recherche,
  - simplifie leur diffusion

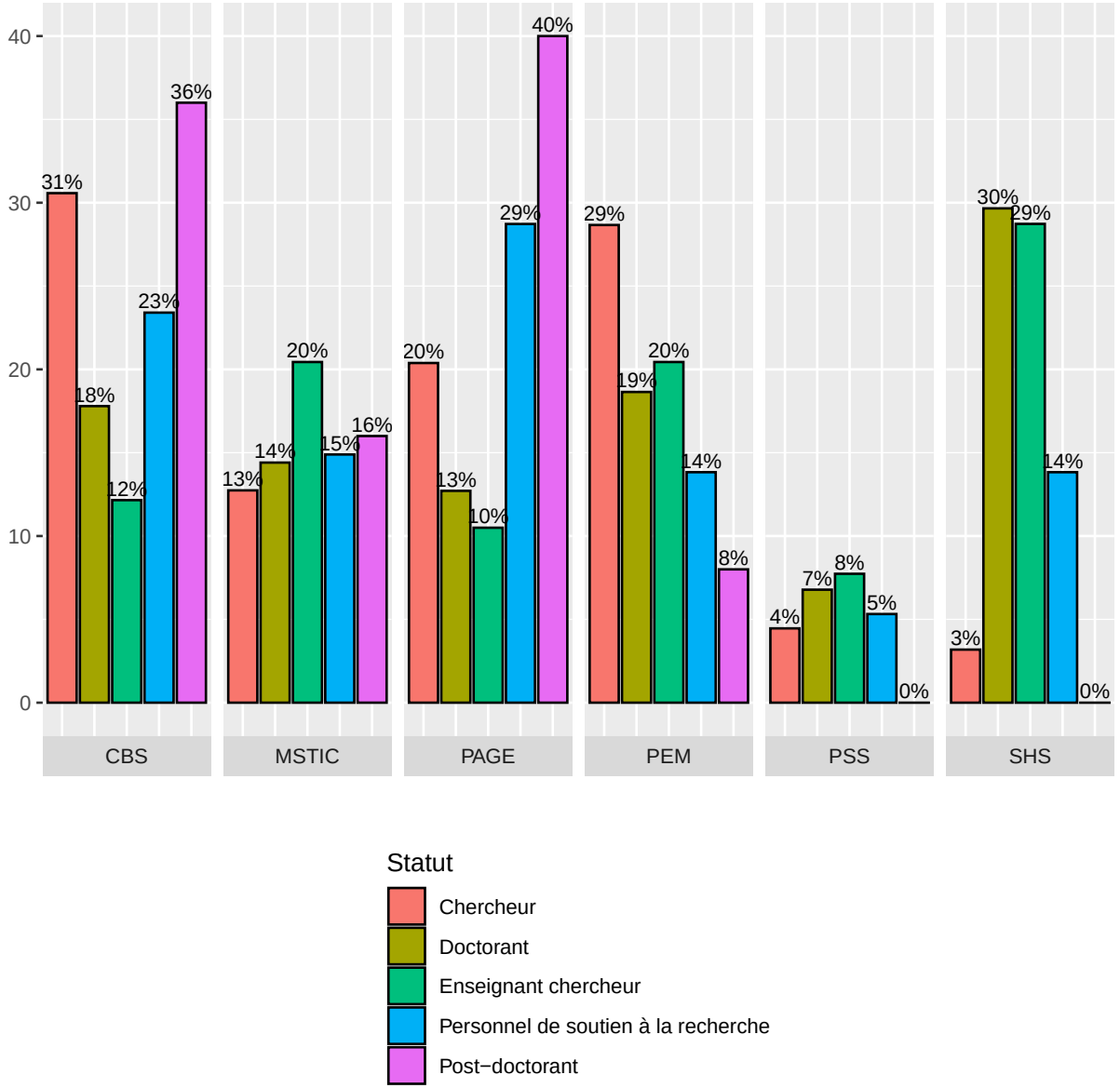
Objectifs : Réaliser un **état des lieux des pratiques** autour des données de la recherche auprès de toutes les communautés scientifiques du site de Grenoble de façon à **adapter l'accompagnement** dans cette dynamique de changement extrêmement rapide

Méthodologie

- Questionnaire proposé du **2 novembre 2020 au 13 décembre 2020**
- Diffusion très large sur le site
- **877 réponses au total** dont 421 réponses complètes et 456 réponses plus ou moins incomplètes. Au total **578 réponses** complètement exploitables et des éléments supplémentaires des autres réponses.
- Questionnaire réalisé en collaboration avec des étudiants de M2 MDLE information et documentation
- Scripts Python et R pour analyse et croisement des données

Profils des répondants

- Bonne **représentativité** de la population recherche du site
  - En terme de couverture disciplinaire
  - En terme de fonction
  - En terme de ratio C/EC



Description des jeux de données et leurs volumes

- **Objectifs** : bonne compréhension des données qui sont manipulées par les communautés scientifiques du site, ainsi que des pratiques en matière de stockage.

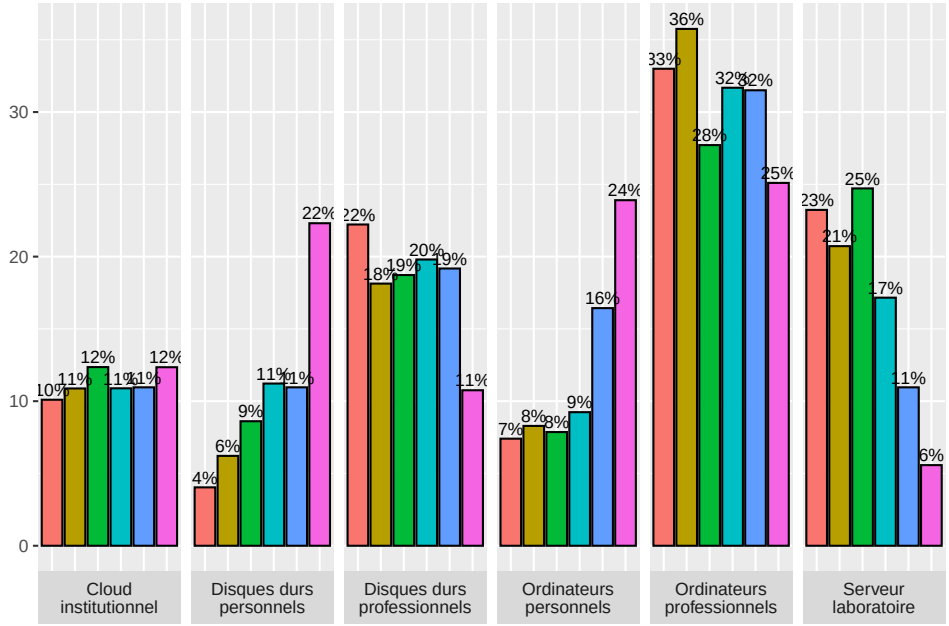
- **Grandes tendances** :

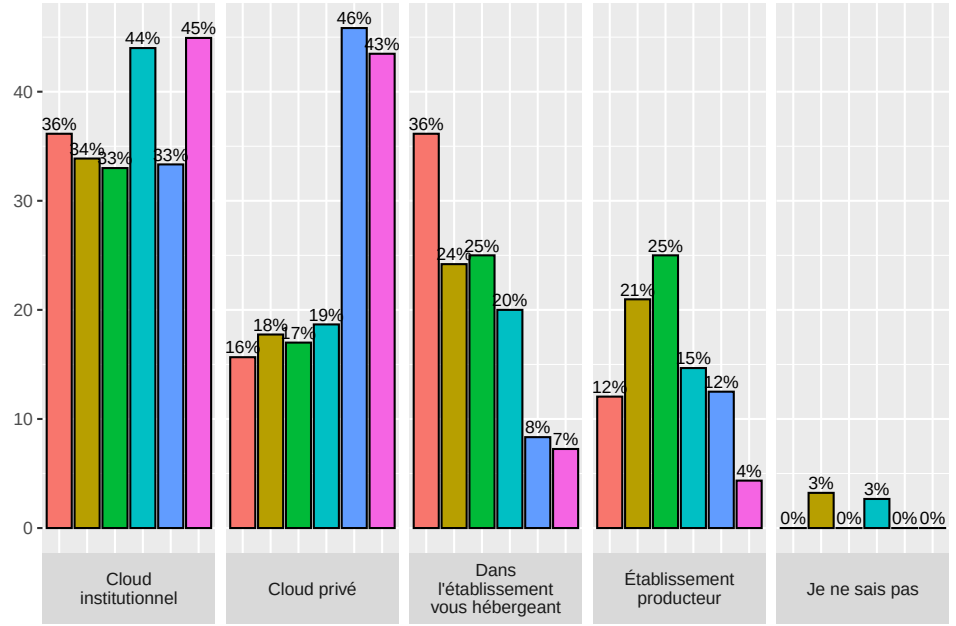
  - importance des données textuelles pour le pôle SHS,
  - données des réseaux sociaux majoritairement dans le pôle PSS,
  - présence des codes sources dans le pôles MSTIC, PAGE et PEM,
  - données de simulation plutôt également dans les pôles PEM, MSTIC et PAGE,
  - workflows très présents dans les pôles CBS et PAGE.

- Autour des volumes manipulés, **différences notables mais attendues** entre les disciplines :

  - gros volumes de données dans les pôles PAGE, CBS et dans une moindre mesure PEM
  - les personnes qui n'ont pas réellement d'idée de la volumétrie de leurs données appartiennent majoritairement au pôle SHS
  - La grande majorité des répondants se positionnent sur des volumes de l'ordre de 1 à quelques centaines de Go.

Problématique du stockage





- Utilisation de **plateformes institutionnelles**, en particulier au plus près de là où sont produites les données beaucoup dans le pôle PAGE
- Utilisation de **matériels personnels et de clouds privés** est très majoritaire dans le pôle SHS, et dans le pôle PSS pour la partie clouds privés
- Utilisation de **disques durs externes professionnelles**, pratique très ancrée dans certains pôles comme MSTIC ou PEM
- Utilisation de **plateformes d'établissement ou de serveurs de laboratoire** assez développée dans les pôle CBS et PAGE.

Synthèse sur le stockage

- Pratiques **disciplinaires**
- Problématique des **grosses volumétries** poussent vers des plateformes **institutionnelles ou dédiées**
- Certains pôles **peu utilisateurs d'espaces mutualisés** portés par les établissements
- Utilisation assez importante de **matériel personnel** chez les jeunes

Recommandations :

- Proposer un **accompagnement spécifique** pour certaines communautés, en particulier les aider à surmonter le **gap technique** pour l'utilisation de plateformes de grande envergure
- Adapter l'aide aux **jeunes chercheurs** : formation, communication, information vers les directions de thèse, accompagnement spécifique
- Poursuivre les efforts de communication autour des **services académiques**, faciliter leur appropriation

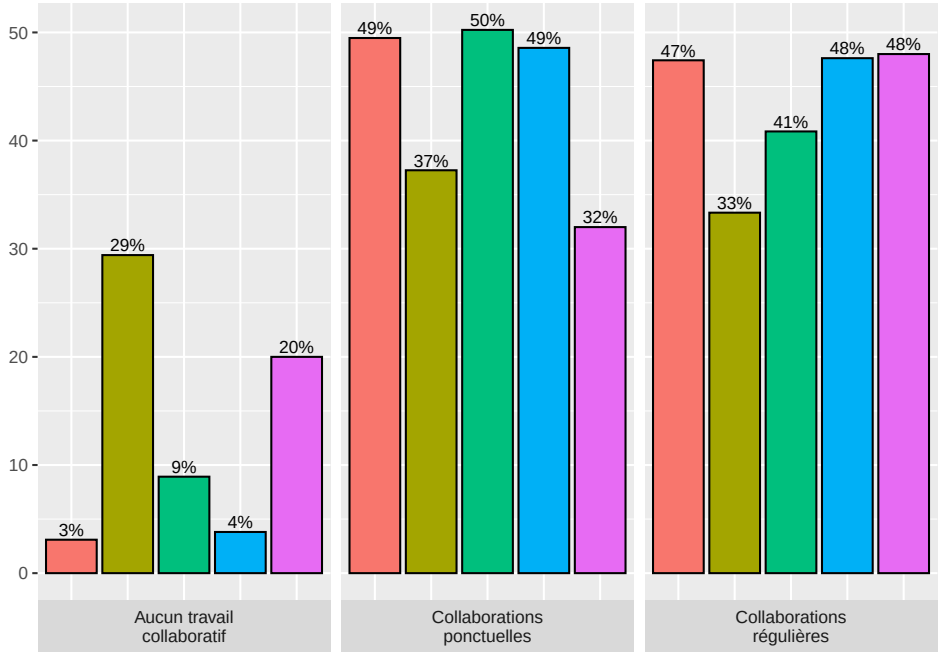
Données sensibles

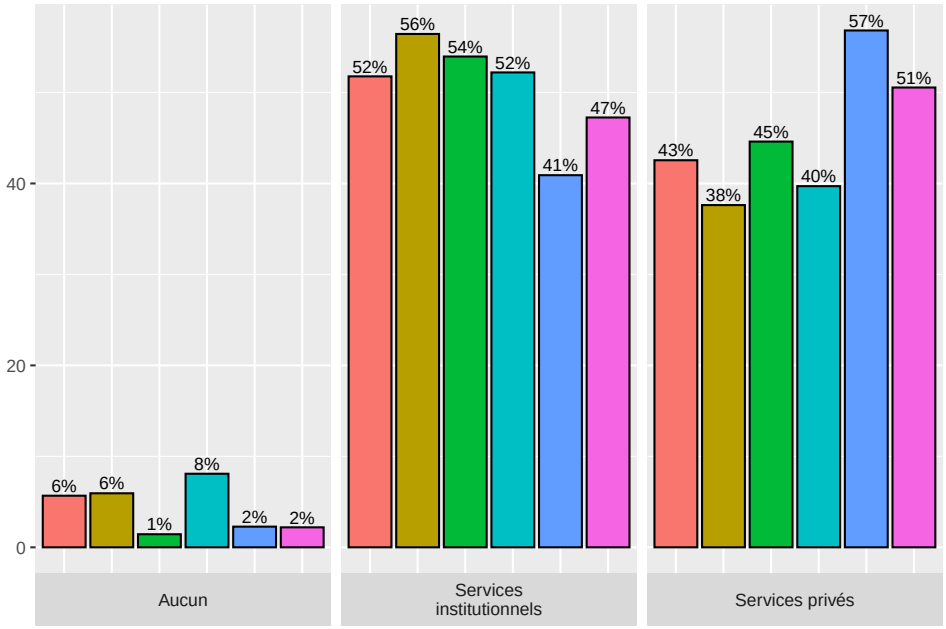
- 17% des participants concernés par des **données personnelles**, 15% par des **données de santé**, et 3% par d'autres formes de données sensibles (en particulier **données confidentielles** liées à des contrats industriels).

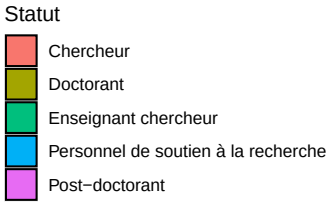
  - **PSS et SHS** manipulent en partie des données personnelles, et également de façon moins importantes des données de santé.
  - Utilisation de données de santé dans **CBS**.
  - **MSTIC** concerné également à la fois par des données de santé et des données personnelles ce qui illustre les **liens interdisciplinaires** existants entre ce pôle et les autres.

Synthèse : Importance de ces données pour certains pôles. Nécessité d'accompagner sur le stockage qui implique des **architectures adaptées et sécurisées**. Identifier sur le site les **expertises** sur les plans juridique et technique et de construire les services d'accompagnement les plus pertinents : aide à la rédaction des **Plans de Gestion de Données**, accompagnement pour la mise en place des architectures techniques, conseil pour la mise en œuvre de processus **d'anonymisation**.

Pratiques collaboratives







Synthèse : majorité de la recherche dans un **cadre collaboratif** avec quelques disparités liés à la **discipline** (pôle SHS) et à la **fonction** (doctorants). Utilisation plus ou moins importante d'**outils commerciaux**.

Pratiques collaboratives : recommandations

- **Outils institutionnels** sont insuffisamment connus et peut être mal adaptés à l'ensemble des usages.

- De vrais questions autour des **enjeux de souveraineté** et de protection des données et du patrimoine scientifique.

- Nécessité
  - de **faire connaître** les outils existants,
  - d'**aider** à leur usage,
  - d'identifier les **freins** à leur utilisation.

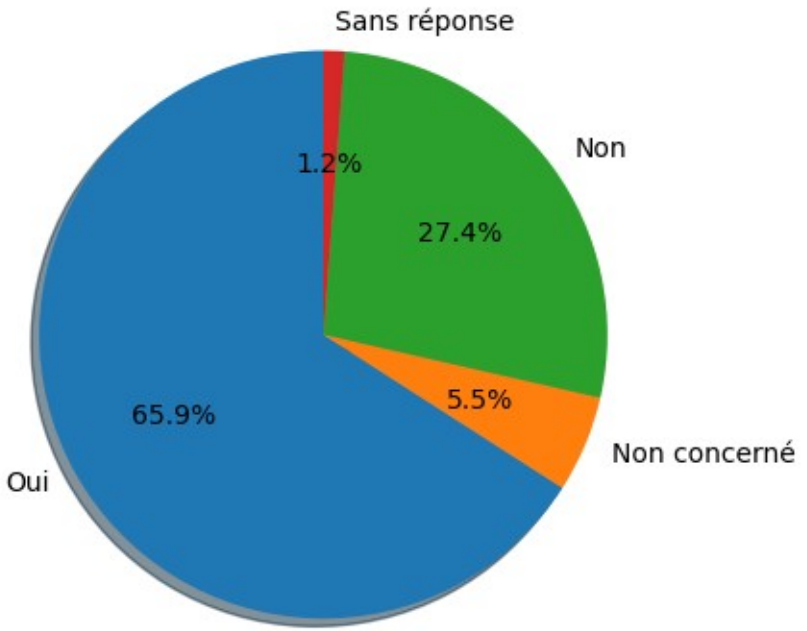
- Importance d'un **accompagnement de proximité** et de la mise en place d'échanges simples et faciles avec les utilisateurs

Réutilisation et diffusion des données

- **Très large proportion** des répondants a déjà **réutilisé** des données publiées par d'autres chercheurs.

- Pôles **PEM, PAGE et CBS** parmi les plus gros réutilisateurs

- **41%** des répondants ont déjà **diffusé leurs données**.

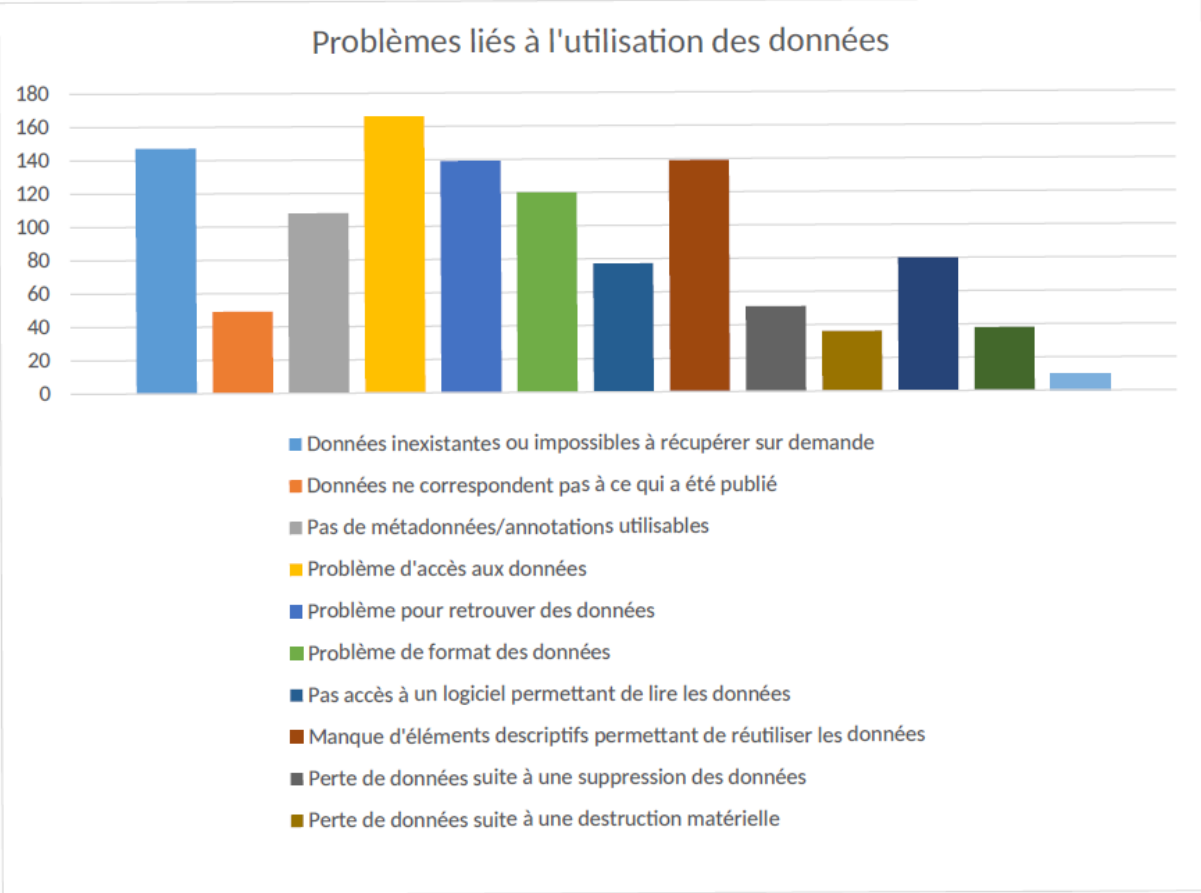


Réutilisation et diffusion des données

- **80%** des répondants indiquent avoir rencontré des **difficultés** dans l'utilisation des données

- **54%** des répondants déclarent **ne pas avoir diffusé leurs données**

- **Doctorants** peu nombreux à diffuser leurs données



Synthèse :

- Importance des **principes FAIR**
- Nécessité de proposer une **aide et des outils adaptés** (entrepôt de données, conseils sur les standards de métadonnées, sur les formats etc.)
- Renforcer les **formations** sur ces questions, en particulier au niveau des écoles doctorales.

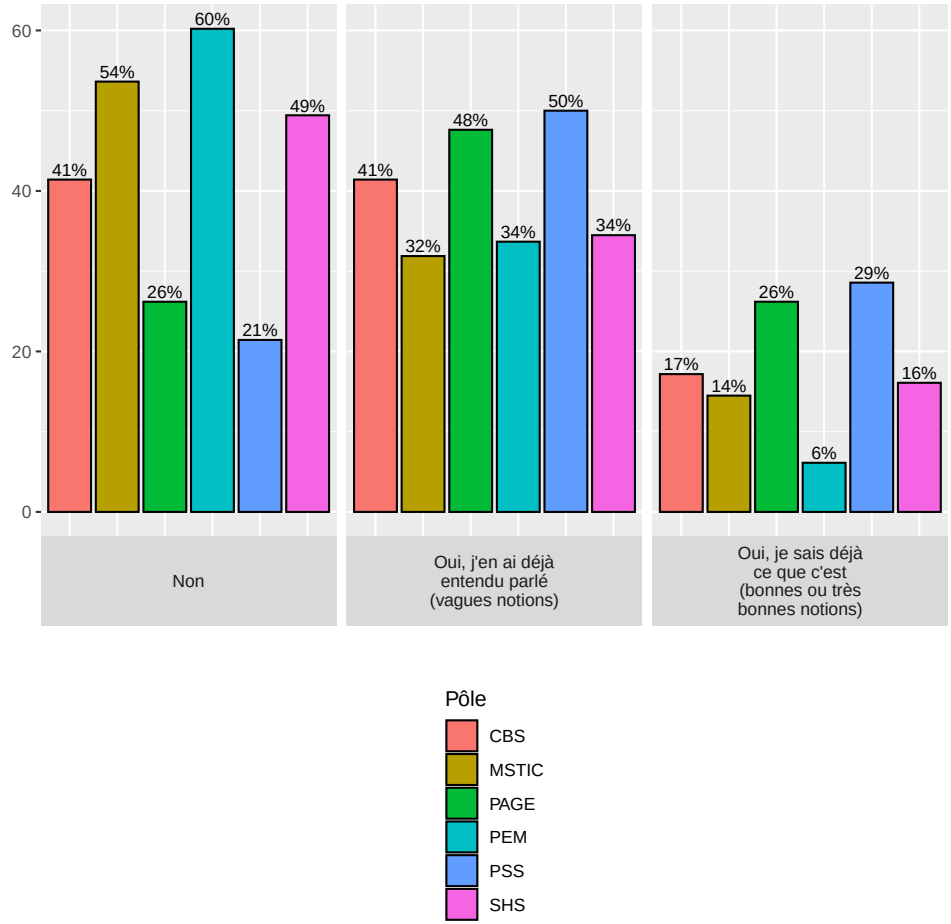
Archivage

- Notion d'**archivage** vraisemblablement **peu comprise** dans l'enquête.
- Termes **stockage, stockage sécurisé et archivage** concepts pas clairs pour la majorité des communautés scientifiques.
- Nécessité de mettre en place des actions de **sensibilisation et d'information**.
- Mais **peu de solutions** actuellement aux besoins d'archivage de données, et solutions **pas réellement accessibles** au chercheur dans son laboratoire, et pas ou peu d'expertise d'archiviste pour accompagner les communautés de recherche.
- Nécessité d'**initier un travail** sur le besoin réel, identifier les compétences sur le site et au niveau national, référencer les possibilités et mettre en place au niveau de l'établissement une réelle politique pour l'archivage des données de recherche.

Plans de gestion des données

- **53%** ont déjà entendu parler des PGD mais seul **15%** parmi eux savent précisément ce que cela recouvre.

- Pour ceux qui ont été confrontés à un PGD, **30 % ont eu des difficultés** :
  - un certain désarroi face à quelque chose qui n'est pas compris,
  - le manque d'exemples disponibles,
  - la problématique des questions éthiques et juridiques pour lesquelles il faut des expertises spécifiques,
  - un manque d'expérience sur le sujet etc.

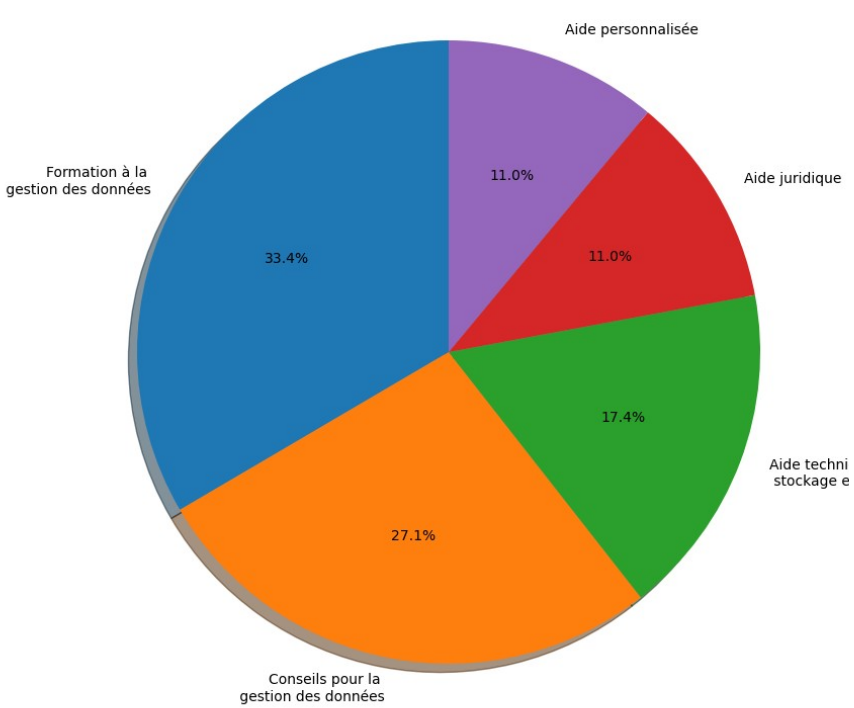


Synthèse :

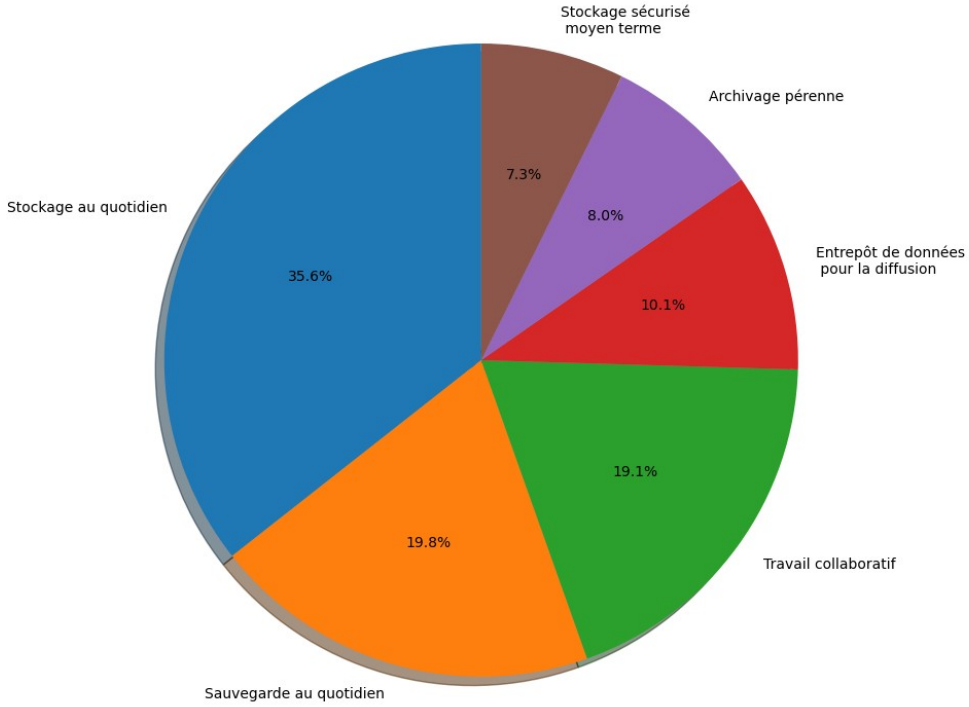
- PGD **largement méconnus**

- Importance des besoins de **formation et d'accompagnement**

Besoins exprimés



Accompagnement et formations



Services numériques

Focus doctorants

Des spécificités liées à :

- au **manque d'information** sur les différentes ressources auxquels les doctorants peuvent avoir accès, tant sur les possibilités d'accompagnement que sur les plateformes et services disponibles,

- au **fort besoin de formation** sur toutes les questions liées aux données.

→ Les services existants sur le site sont d'abord à destination des **membres permanents** des laboratoires, et difficilement utilisables par des doctorants.

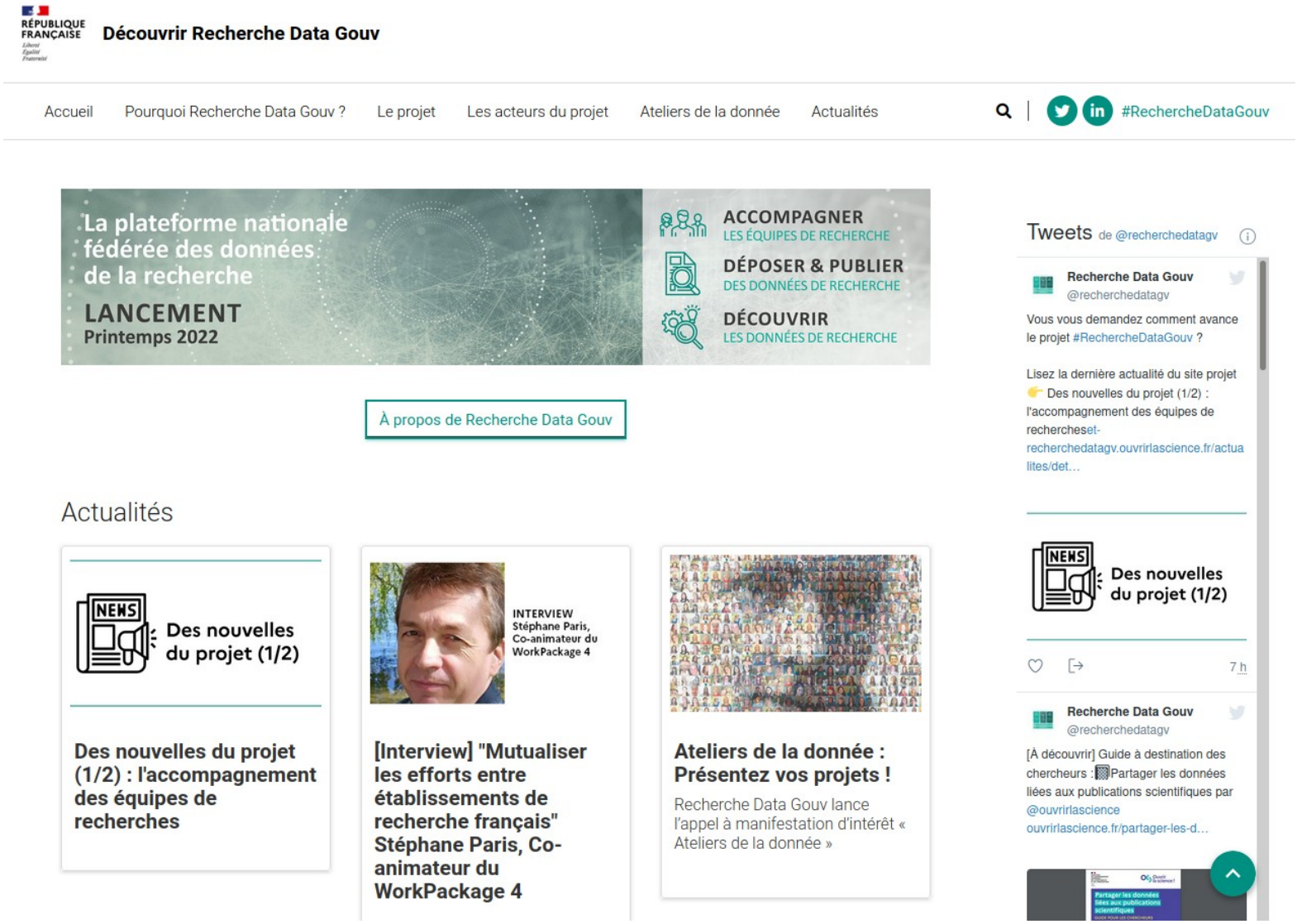
→ La mise en place d'un **accompagnement spécifique** et adapté à cette population est donc crucial.

Conclusions et perspectives suite à l'enquête

- Travail important encore à réaliser sur la **sensibilisation et l'information**
  - Services existants
  - Aide aux usages
  - Identifier les freins à leur utilisation
- **Formation et accompagnement**
  - Adapté aux communautés
  - Accompagnement spécifique pour les doctorants
  - Identifier les expertises (juridiques, archivage …)
- **Nouveaux outils ou services**
  - Identifier les expertises (juridiques, archivage …)
  - Mettre en place les services et outils pour lesquels les besoins sont clairement exprimés (par ex cahier de labo électronique)

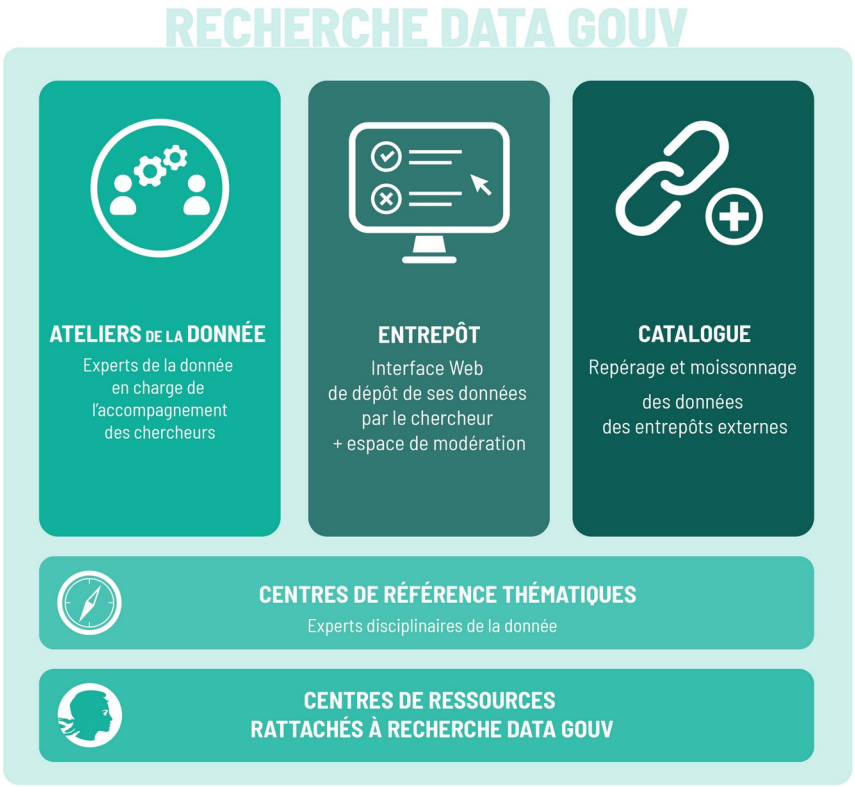
→ **Pas mal de choses initiées** depuis la fin de l'enquête par la CDGA

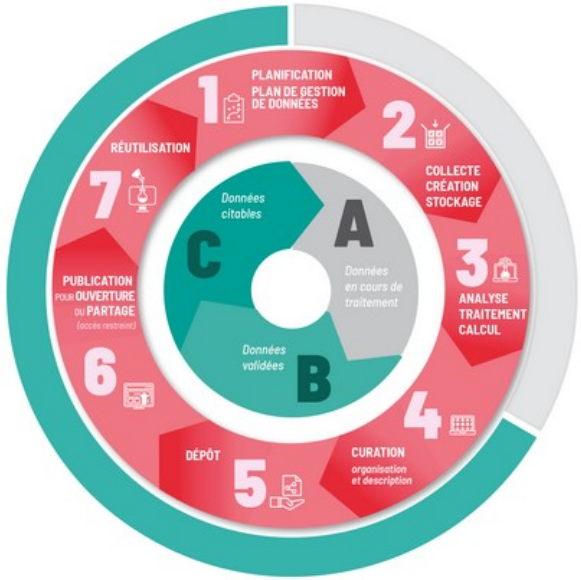
Recherche Data Gouv



Description de RDG

- **Objectif** : accompagner les besoins autour des données de la recherche.





- **Volet accompagnement** : ateliers de la donnée, centres de ressources et de référence thématiques
- **Volet entrepôt** pour déposer et publier des données
- **Volet catalogue** pour signaler des données déposées dans des entrepôts externes

Le pilier entrepôt

- Annoncé dès 2018 par la Ministre de l'Enseignement Supérieur, de la Recherche et de l'Innovation : service générique d'accueil et de diffusion des données.
- **Recherche Data Gouv** : solution souveraine aux chercheurs pour déposer leur données de recherche.
- Sur la solution logicielle open source **Dataverse**
- Organisé en **espaces institutionnels** de publication et de signalement des données des établissements qui souhaitent participer au projet.
- **Gouvernance collective** qui réunit les représentants des propriétaires d'espaces institutionnels de données.
- **Participation importante** de membres de la CDGA au projet

Collection UGA et organisation sur le site

- L'UGA a fait le **choix de s'appuyer sur RDG** pour son entrepôt de données

- **Espace institutionnel UGA** : composé principalement des collections des laboratoires et autres entités qui demanderaient à avoir une collection à la racine.

- **Modération des dépôts** par une cellule centrale à laquelle les laboratoires sont appelés à participer :

  - les demandes issues d'un laboratoire seront traitées en priorité par les modérateurs affiliés à ce laboratoire, et à défaut par les autres modérateurs.
  - Les laboratoires le souhaitant pourront modérer eux-mêmes leur collection.
  - Sinon c'est la Cellule Data qui s'en occupera.

- A noter :

  - **Tous les types de données** sont acceptés
  - Taille max des jeux de données : XXX
  - Embargo de XXXXX
  - Accès restreint possible

- Choisir en priorité dans un **entrepôt thématique** reconnu par sa communauté

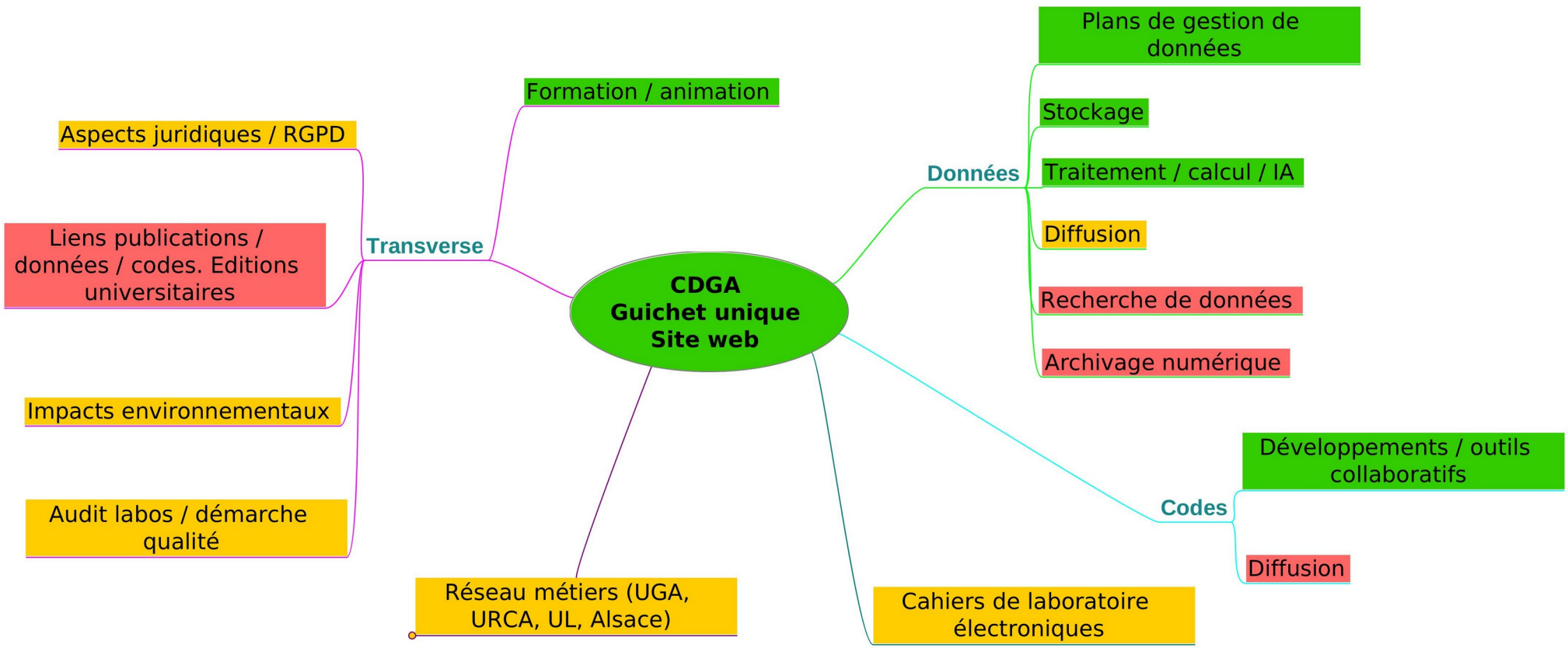
- **Dépôts anticipés** dès à présent possibles sur l'espace UGA : **nous contacter** !

Pilier Atelier de la donnée

- Objectif : labelliser les initiatives existantes ou nouvelles sur l'accompagnement des communautés scientifiques tout au long du cycle de vie des données.
- En proximité, regroupant des compétences complémentaires
- Différents niveaux de labellisation selon la maturité et le niveau de service apporté
- Réponse de la CDGA à l'appel à projets en février. En attente des résultats pour début juin.

Pilier Atelier de la donnée : services de la CDGA

- Vert : Service en place
- Orange : Service existant en cours de consolidation
- Rouge : Nouveau service à initier



Conclusions / contacts

- Besoin d'accompagnement ou de support :
  sos-data@univ-grenoble-alpes.fr

- Pour contacter la cellule :
  uga-cellule-data@univ-grenoble-alpes.fr

- Pour se tenir informer : abonnement sur la liste uga-research-data
  - https://listes.univ-grenoble-alpes.fr/sympa/info/uga-research-data

- Site web :
  - https://scienceouverte.univ-grenoble-alpes.fr/

CELLULE DATA
GRENOBLE ALPES