CELLULE DATA
GRENOBLE ALPES

La Cellule Data Grenoble Alpes

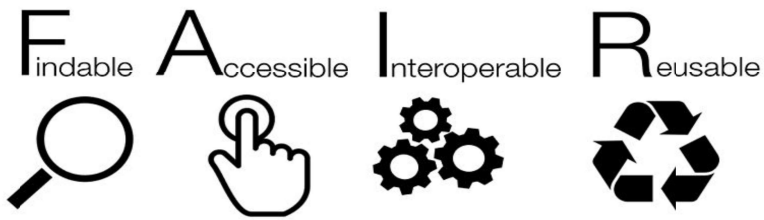
GRICAD, Journée utilisateurs, 30 novembre 2021

Sommaire

- CDGA, une cellule d’accompagnement sur les données de la recherche
- Quelques exemples concrets de réalisations de la cellule
- Les perspectives à l’UGA
- Conclusion

Pour commencer (petit rappel) ..

Principes FAIR



Findable / *Trouvable*

Données **faciles à trouver**.

- possédant un identifiant unique et pérenne
- décrites par des métadonnées riches
- enregistrées ou indexées dans une source interrogeable

Interoperable / *Interopérable*

Facile à combiner avec d'autres jeux de données, par les humains **et** les systèmes informatiques

- formats libres et ouverts
- mise à disposition du code source si le logiciel de traitement existe
- standards de métadonnées et vocabulaire standardisés

Accessible / *Accessible*

Données ou au moins méta-données facilement accessibles.

- entrepôt de confiance, pérenne, certifié
- définir les conditions d'accès et la licence de diffusion
- si embargo ou accès restreint : méta-données accessibles

Reusable / *Réutilisable*

Prêtes à être **réutilisables** pour une future recherche y compris via des méthodes informatiques

La réalité du terrain

“L’Europe me demande un DMP, qu’est-ce que c’est ?”

“Je n’arrive plus à ouvrir mon tableur avec ce fichier de données”

“Je veux utiliser AWS, comment je fais ?”

“Je stocke mes données sur un DD externe qui se trouve dans mon bureau”

“Je veux diffuser mes données sur le web”

“J’ai pas mal de données de types différents et je pense que ce serait très intéressant de pouvoir les croiser”

“Comment diffuser mon code avec mes données ?”

“On doit partager des données entre plusieurs collègues dont certains sont en Angleterre”

“J’ai loué de la volumétrie pour mon projet mais celui-ci est terminé et je n’ai plus d’argent”

“Le financeur de mon projet demande à ce que mes données soient diffusées”

“Mon équipe aimerait tester le Deep Learning sur nos données”

CDGA, une cellule d’accompagnement sur les données de la recherche

CELLULE DATA
GRENOBLE ALPES

Contexte du site Grenoble Alpes

- Des structures en soutien avec des expertises différentes :
  - GRICAD
  - BAPSO
  - MSH

- Des laboratoires plus concernés et plus impliqués

Composition de la cellule :

Des compétences complémentaires : **techniques, science ouverte, juridique**.

16 membres issus de :

- GRICAD
- BAPSO, DDOR (CNRS)
- MSH
- Labos
- + le DPO du site

Acculturation autour des données

- Les **coûts humains** liés aux données de recherche sont très importants, très sous-estimés et très peu pris en compte
- Ils ne pourront pas reposer sur un petit ensemble de personnes mais devront se répartir sur l'**ensemble du collectif**

Outre l'accompagnement pratique, l'un des objectifs de la cellule data est d'accompagner les changements **culturels, méthodologiques, professionnels**, liés aux données :

- de proposer des **actions de sensibilisation** sur ces questions
- d'organiser des **formations** au niveau du collège doctoral et à destination des chercheurs et ingénieurs
- de proposer des **séminaires, ateliers, retours d'expérience** ...

CDGA : répondre à la réalité du terrain

- 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
  - Assurer la présence grenobloise dans les **projets nationaux et européens**, faire le lien avec les communautés scientifiques du site

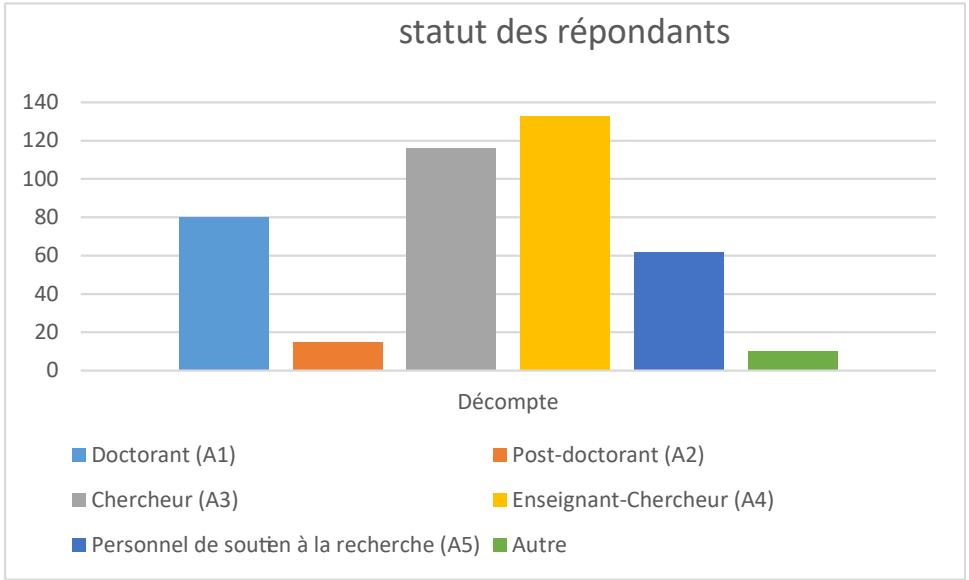
Quelques actions en cours

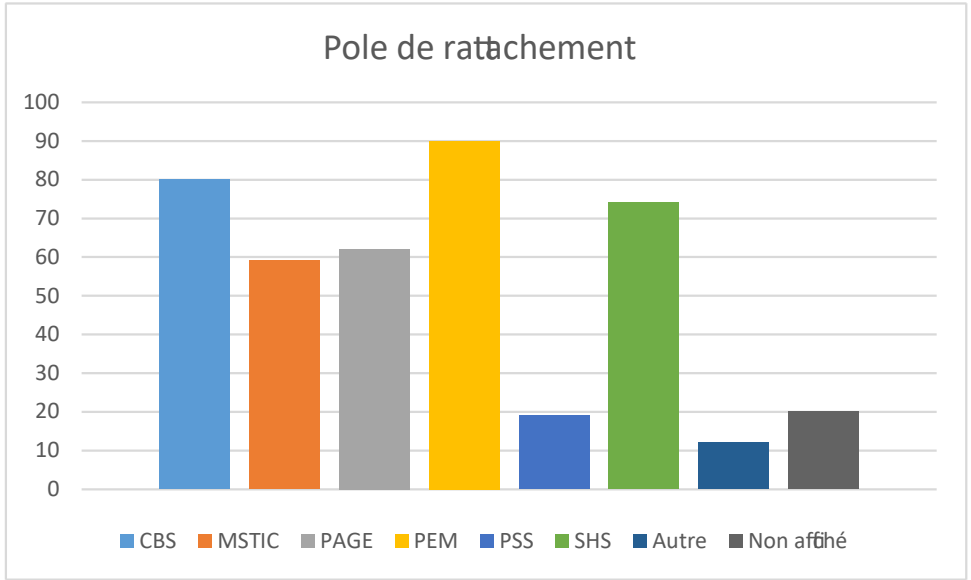
- **Enquête** autour des **besoins** des communautés du site
- Publication du **Baromètre Science Ouverte** du site intégrant les publications
- Actions de **formation** (collège doctoral, autre …) et d'**animation** (séminaires, ateliers …)
- Mise en place d'un **site web « science ouverte »**, point d'entrée unique
- Animation de la **liste de discussion**
- Participation forte à la création de l'entrepôt national **recherche.data.gouv**.
- Animation et renforcement du réseau de **référents labos données de recherche**
- Mise en place d'un GT autour des **impacts environnementaux** des données
- Suivi des initiatives nationales, européennes et internationales : le **CoSO**, la **RDA**, les activités autour d'**EOSC** ...

Quelques retours de l'enquête

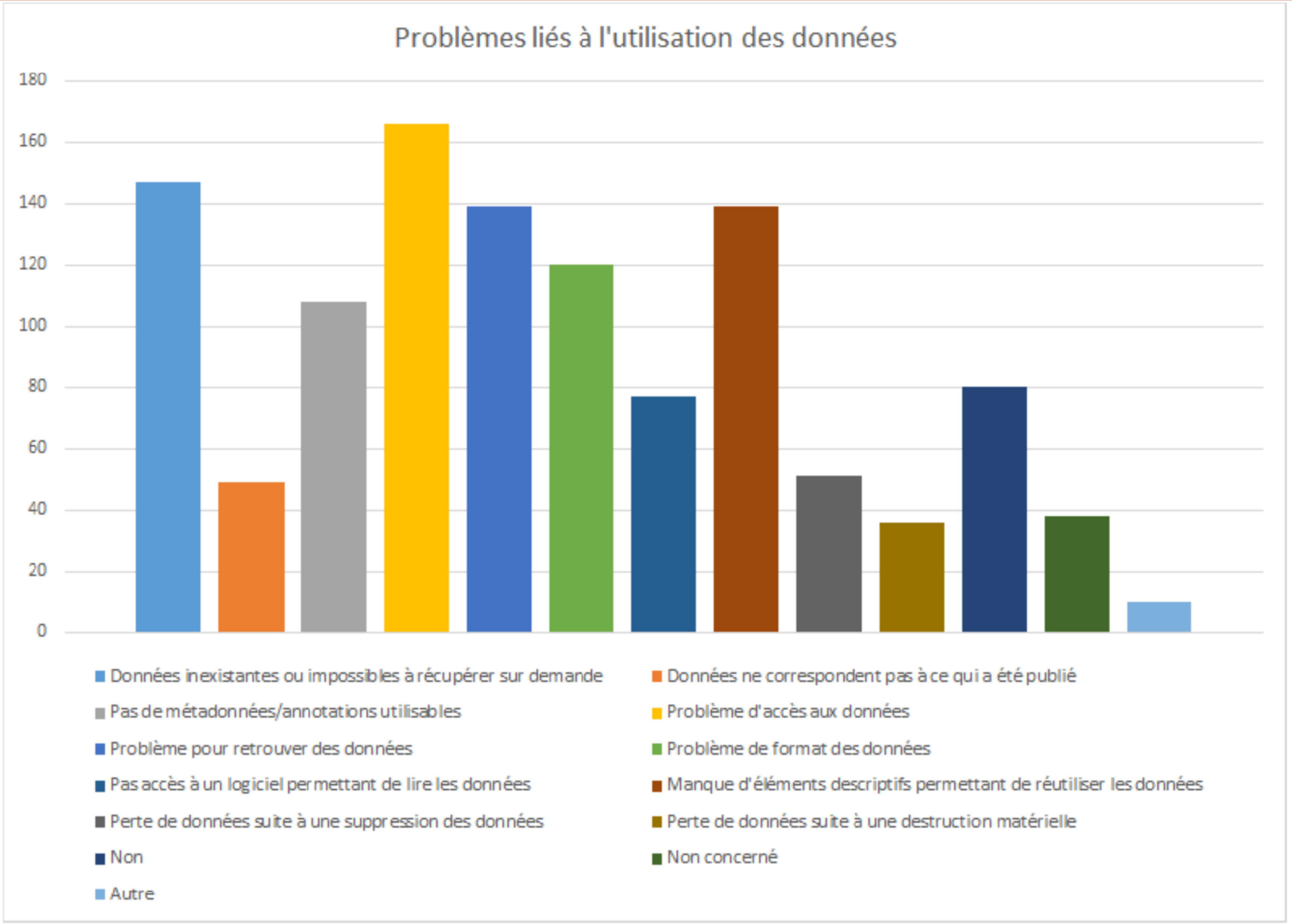
- **Objectif** :
  - recueillir les besoins des enseignants-chercheurs et des ingénieurs, afin de pouvoir mieux y répondre en termes d'accompagnement, de formation, de services, de ressources etc.
  - établir un état des lieux : quelles données, quelles pratiques des chercheurs, etc

- **Participants** :
  - Réponses complètes : 414 (extraction des graphes sur les réponses complètes)
  - Réponses incomplètes : 446 (dont beaucoup quasi complètes donc exploitables)
  - Total : **860 réponses**



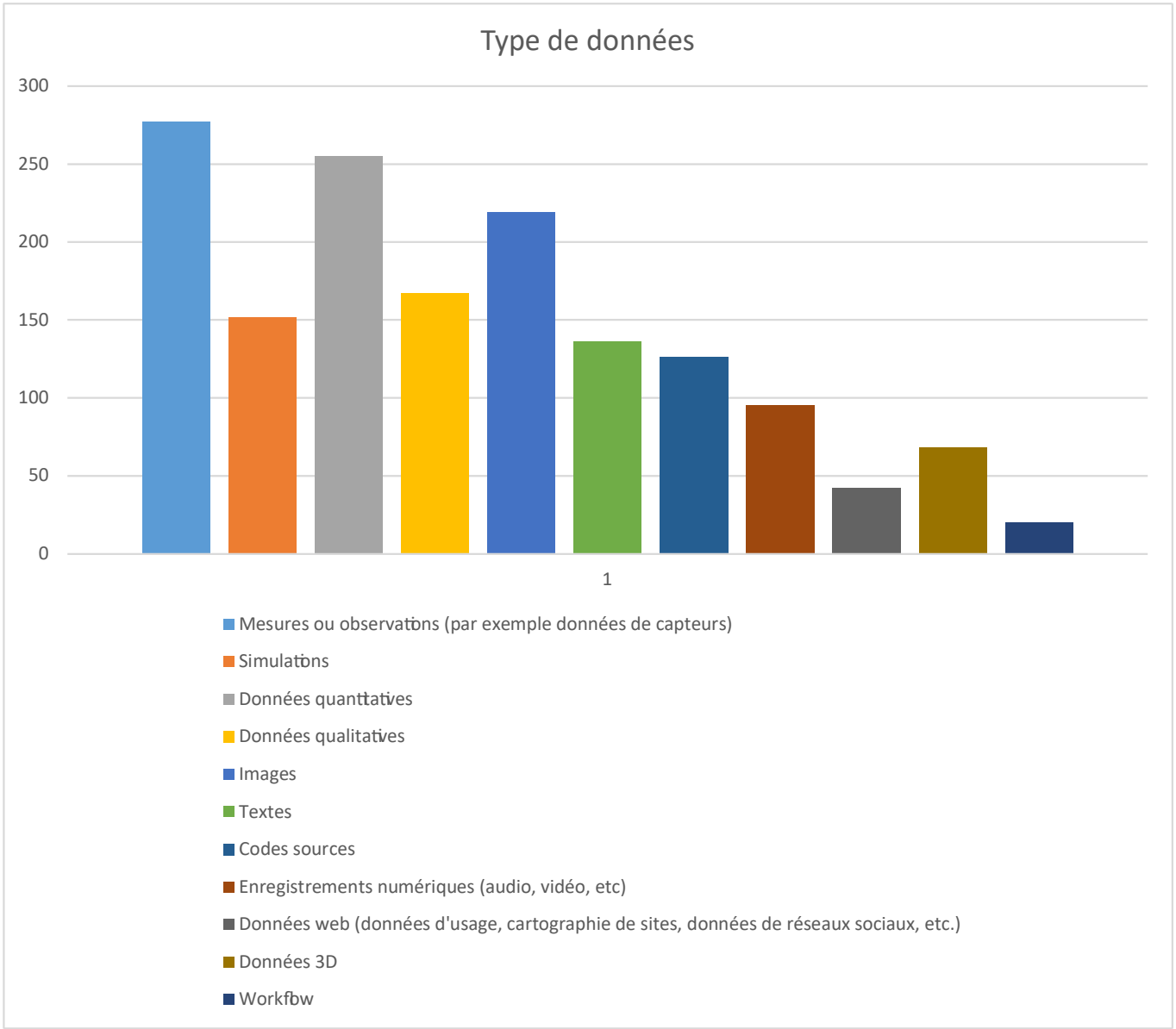


Quelques éléments de l'enquête

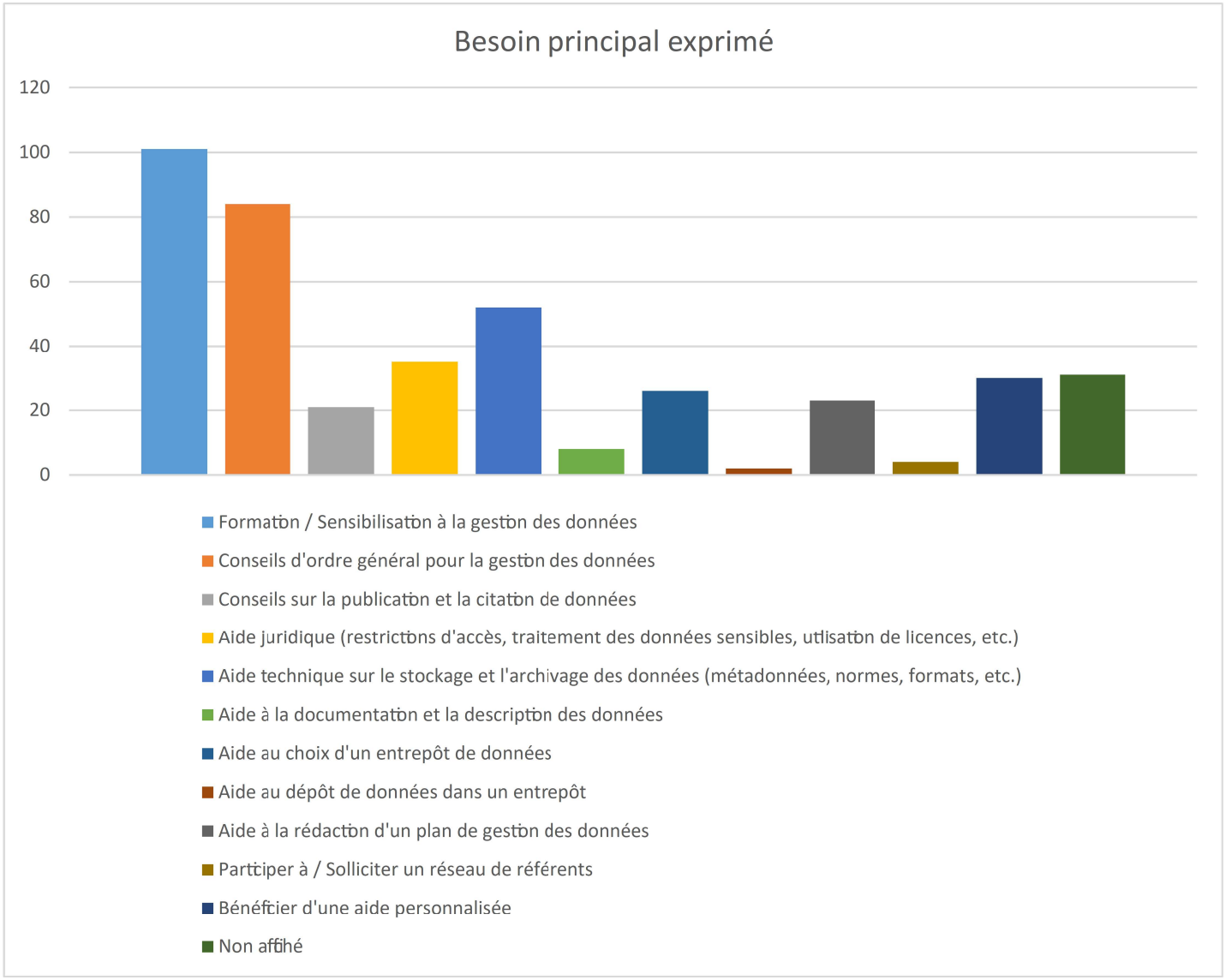


A noter : plus de 70 % des répondants (réponses complètes) ont déjà rencontrés des problèmes liés à l'utilisation des données

Quelques éléments de l'enquête



Quelques éléments de l’enquête



Rapport final en cours de rédaction pour une parution avant fin d’année ou tout début d’année prochaine

Quelques exemples concrets de réalisations de la cellule

1/ Rédiger un Plan de Gestion des Données
2/ Répondre aux questions réglementaires
3/ Où et comment stocker ?
4/ Comment et où traiter ses données ?
5/ Comment décrire ses données ?
6/ Comment et où diffuser ses données ?
7/Formations et séminaires

1/ Rédiger un Plan de Gestion des Données

Le PGD (ou DMP, Data Management Plan) peut être vu comme une contrainte administrative de plus **MAIS** il **permet de réfléchir à la gestion des données d'un projet en amont** :

- Quelles données vont être collectées : quel type de données, comment sont-elles collectées, où les stocker, comment on sécurise le stockage, quelle volumétrie, quels formats, quelle organisation ...
- Comment vont-elles être utilisées : comment on les partage, comment on les traite, où on les traite, ...
- Comment elles vont être préservées : à quel terme, quelles données, où, comment ...
- Comment elles vont être valorisées : comment les diffuser, sous quel format, sous quelle licence, quelles données, comment associer les codes, …
- Comment assurer le financement des ressources nécessaires ?

1/ Rédiger un PGD
Intégrer les bonnes pratiques

Objectif de la cellule : aider à la mise en œuvre de bonnes pratiques tout au long du projet

- **Identifier les données du projet,** leur type, leur volumétrie, les contraintes associées ...
- **Documenter les données** : description du projet, du processus de collecte, des matériels et logiciels utilisés, de la structuration de la base de données, du processus de nettoyage, …
- **Utiliser des métadonnées** standards et spécifiques à sa communauté
- **Utiliser des formats de fichiers ouverts**, non propriétaires, documentés, reconnus dans sa communauté (https://facile.cines.fr/)
- **Utiliser des conventions de nommage et d'organisation** des fichiers et répertoires, préciser les versions et dates (ou utiliser un gestionnaire de version)
- **Définir les conditions juridiques d'utilisation** de ces données
- **Définir les modalités de diffusion**, de stockage et d'archivage des données

1/ Rédiger un PGD
En pratique

- Aide à l'utilisation de l'outil **DMP Opidor**
  - Adapté aux différents DMP (ANR, Europe)
  - Fonctionnalités utiles, recommandations
  - Travail en cours pour intégrer des recommandations propres au site de Grenoble Alpes
- Aide à la **rédaction, relecture et commentaires**
- En particulier, élaboration d'un document qui regroupe tous les éléments techniques concernant la **plateforme de stockage SUMMER** à intégrer dans le DMP
- Accompagnement des porteurs de projet (ANR, Européen)

2/ Répondre aux questions réglementaires

S’appuyer sur les **expertises présentes** sur le site pour répondre aux questionnements des scientifiques :

- DPO (Data Protection Officer) des établissements
- Services de valorisation
- Expertises dans les laboratoires

Faire de la **veille juridique et réglementaire** :

- Research Data Alliance
- COmité pour la Science Ouverte
- CNIL…

Aider à la **rédaction des documents** type analyse d’impact … sur les parties techniques

3/ Où et comment stocker ?
Se poser les bonnes questions

- Déterminer les **lieux et supports de stockage** selon le volume, la fréquence de consultation, le besoin en traitement et analyse, le besoin de partage, la durée de stockage …
- Identifier les **coûts**
- Qualifier les **données sensibles** et leur niveau de protection nécessaire
- Déterminer les **durées de stockage, y compris la fin de vie des données**
- Anticiper le **partage** en hiérarchisant le contenu, les types, etc. et les autorisations d'accès
- Prévoir la **sécurisation** et les sauvegardes

Plusieurs formations sur le sujet, un accompagnement des communautés au quotidien

3/ Où et comment stocker ?
Les infrastructures disponibles

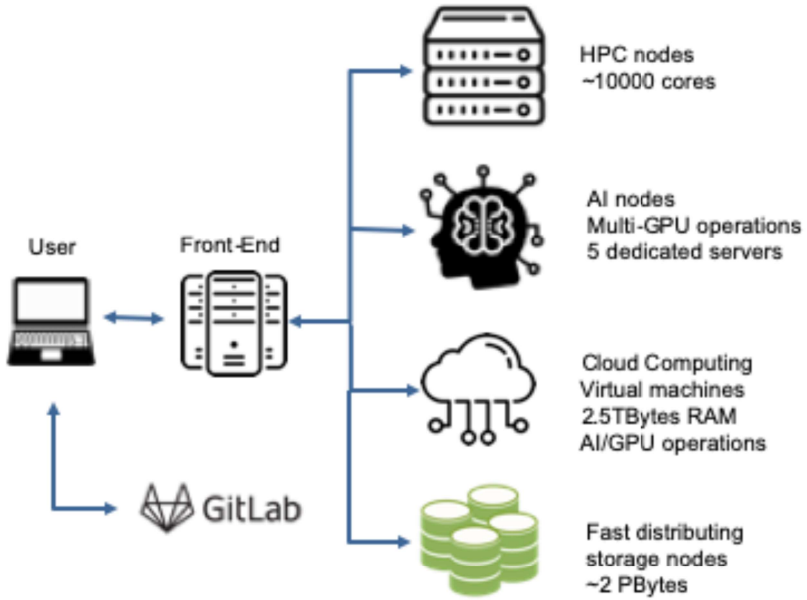
Orienter vers et accompagner sur les infrastructures adaptées aux besoins

- Différentes **plateformes** : SUMMER, Mantis, Bettik
- Différentes **technologies** : NetApp, IRODS, BeeGFS
- Différents **usages** : sécurisé, distribué, performant
- Différents **type d'accès** : local, global, spécifique calcul
- Des **infras nationales** : Huma-Num (SHS), CINES, centres de données dédiés

4/ Comment et où traiter ses données ?
Infrastructures disponibles

Orienter vers et accompagner sur les <u>infrastructures adaptées aux besoins</u>

- **HPC** : Dahu
- **IA** : BigFoot
- **HTC, traitement de données** : Cigri
- **Cloud** : Nova
- **Notebooks** : Jupyter
- Des **infras nationales** : GENCI, France Grille et **européennes** : PRAC

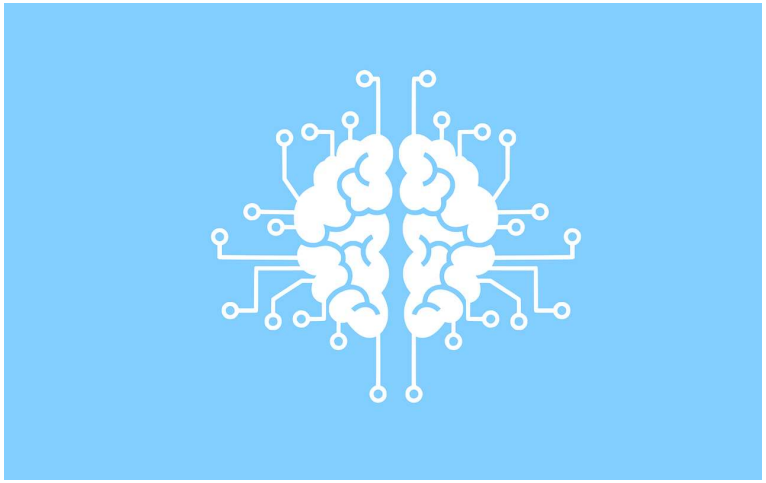


4/ Comment et où traiter ses données ?
Exemple autour de l'IA

- Identification de la **plateforme la plus adaptée** aux traitements à effectuer : BigFoot, Jean Zay ou Nova.

- Aide à l'**utilisation des logiciels** : installation des softs si non disponibles, mise à disposition

- Aide sur les problématiques de **stockage et de mouvements** des données
  - Par exemple : connexion directe entre l'espace de stockage Mantis et les machines de l'IDRIS



5/ Comment décrire ses données ?
Des standards disponibles

- Généralistes
  - Datacite, Dublin Core

- Disciplinaires (voir liste de la Research Data Alliance)
  - Sciences sociales : Data Documentation Initiative (DDI)
  - Ecologie : Ecological Metadata Language (EML)
  - etc

Métadonnées importantes à compléter
- Auteur, titre, sujets, date, format, licence d'usage, etc..

Bonnes pratiques
- Identifiants uniques (de type doi)
- Fichier Readme

Aide proposée : aide à la description des données

6/ Comment et où diffuser ses données ?
Différentes modalités de diffusion

- Déposer dans un **entrepôt de données**
  - Des entrepôts **généralistes** (Zenodo)
  - De nombreux entrepôts de données **thématiques**
  - Un **répertoire d'entrepôts** (voir par exemple un annuaire comme Re3data)
  - Des **moteurs de recherche** spécialisés comme Datacite Search
- Ecrire un data paper
  - Liste de data journals par l'université d'Edinburgh

Aide proposée : **identification d'entrepôts pertinents** selon la thématique de recherche et les contraintes et besoins du projet de recherche, accompagnement au dépôt

7/ Formations et séminaires

- Une **formation du Collège des Ecoles Doctorales** sur la gestion des données de la recherche
  - 2 jours prévus en mai 2022 (module déjà réalisé en 2021)
  - Approche concrète: cours, exercices, travaux en groupe
  - 2 modules complémentaires indépendants
    - Respect du RGPD
    - Diffusion des données
- Une formation sur le **stockage des données**
  - Co-organisée avec l'URFIST de Lyon, l'université de Lorraine, l'université de Strasbourg (supports et vidéos en ligne)
- Des **séminaires réguliers**
  - Présentation des services de la CDGA
  - Rédaction d'un plan de gestion des données (support et vidéo)

Perspectives

Où déposer ses données quand on est dans un laboratoire UGA ?

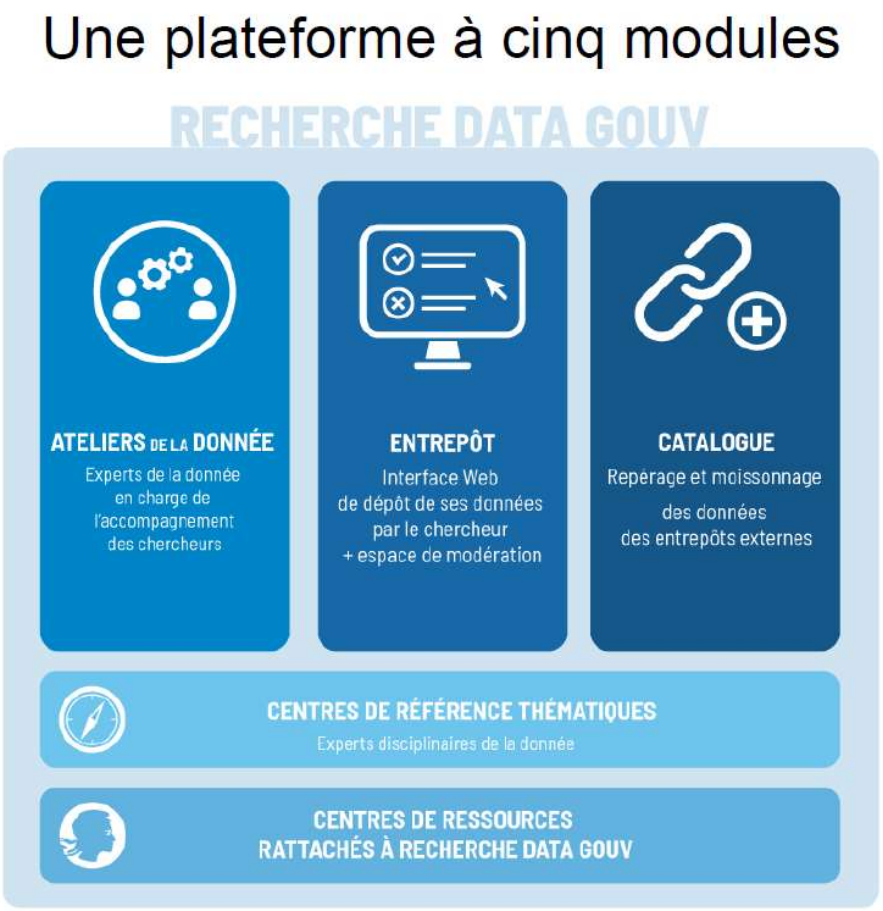
- Stratégie de l'établissement : proposer un **entrepôt institutionnel** pour répondre aux besoins des scientifiques qui n'ont pas de solution disciplinaire

  - Un calendrier de mise en œuvre qui avait été prévu pour une entrée en production au 1er semestre 2022
  - Donc comparable au calendrier du projet national recherche.data.gouv

- Le choix a été fait de ne rien déployer en production en local mais de **rejoindre la plateforme nationale** dès son démarrage.

  - Investissement humain important du site dans le projet RDG

Focus sur Recherche.data.gouv



- Modules **« entrepôt - catalogue »**
  - Projet confié par le MESRI à l'INRAE avec la participation d'établissements volontaires dont l'UGA
- Module **« ateliers de la donnée »**
  - AMI en cours. Objectif de labelliser des initiatives d'accompagnement des communautés recherche sur l'ensemble du cycle de vie des données
  - **Réponse en cours de construction à l'UGA basée sur les actions de la Cellule Data Grenoble Alpes**

Focus sur Recherche.data.gouv

- Première version de l'entrepôt en **mars 2022**
- Première vague de labellisation des ateliers de la donnée en **avril – mai 2022**
- Le dispositif d'accompagnement va aussi s'appuyer sur :
  - Des **centres de ressources thématiques**, spécialisés sur des aspects disciplinaires
  - Des **centres de ressources rattachés à Recherche Data Gouv** (techniques, mutualisation de formations, ...)

UGA et Recherche.data.gouv

- En cours : travail sur l'organisation pour l'**accompagnement des dépôts** sur Recherche Data Gouv
  - Un groupe de travail composé de membres de laboratoires de différentes disciplines
- En cours : **construction de la réponse** pour labelliser la CDGA comme atelier de la donnée
  - Discussions à venir avec les référents données des laboratoires pour assurer l'adéquation de la réponse avec les besoins
  - **Est-ce que votre labo a bien désigné un référent donnée ?**

Conclusion

Conclusions / contacts

Une adresse support unique : sos-data@univ-grenoble-alpes.fr

Un site de ressources :
https://scienceouverte.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

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

CELLULE DATA
GRENOBLE ALPES