CELLULE DATA
GRENOBLE ALPES

La Cellule Data Grenoble Alpes

ANF RNBM, 5 octobre 2021

Sommaire

- Gestion des données de recherche : constats, contraintes, enjeux et réalité de terrain
- Contexte du site Grenoble Alpes : CDGA, une cellule d’accompagnement sur les données de la recherche
- Quelques exemples concrets de réalisations de la cellule
- Conclusion

Gestion des données de la recherche : constats, contraintes, enjeux et réalité de terrain

Quelques constats

Au quotidien pour les structures et les personnels d'appui à la recherche :

- Des interrogations autour du stockage
- Des questions autour du partage des données
- Des demandes pour la diffusion des données …

De plus en plus **nombreuses**.

- Dans un environnement national, européen et international extrêmement **dynamique**.
- Avec des **obligations** qui évoluent pour les scientifiques : Plan de Gestion de Données, Science ouverte, données ouvertes …
- Et un volume de données à gérer en constante augmentation

Et une diversification des usages du numérique

- Disponibilité de **très gros volumes de données** issues de sources **très variées**

- Les **méthodes de travail**, tout comme les **technologies d'analyse des données**, ont fortement **évolué** ces dix dernières années : travail collaboratif et nécessité de partage, émergence de l'open data, outils d'indexation et de recherche …

- **Diversification des besoins** : traitement de données, intelligence artificielle, calcul intensif, stockage, diffusion et valorisation de données, questions juridiques autour des données …

- **Accompagnement à adapter**, communautés "neuves" dans le domaine, connaissance et appropriation technique très hétérogènes des technologies et des infrastructures d'une communauté à l'autre, d'un individu à l'autre

- Un soutien technique dans les laboratoires **très inégal** d'une structure à l'autre

Dans un cadre de plus en plus contraint

- **RGPD** (2016) : règlement européen sur la protection des données personnelles

- **Obligation de Plan de Gestion de Données** exigée par de plus en plus de financeurs de la recherche (H2020, ANR, Horizon Europe, ...)

- Loi pour une république numérique (2016), Plan National pour la Science Ouverte (2018, 2021), H2020 (depuis 2017) et Horizon Europe (à partir de 2021), Feuille de route de la science ouverte du CNRS (2018), Plan données de la recherche CNRS (2020) : l'**ouverture des données** de recherche financées sur fonds publics devient la norme

Plan de Gestion des Données

Ce n’est pas :

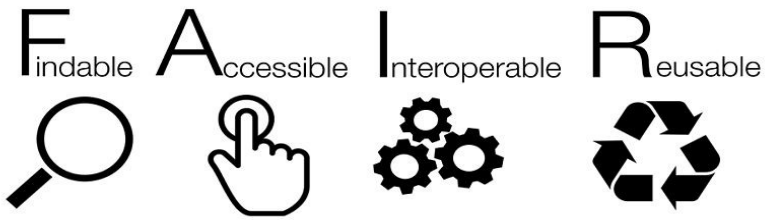
- un nième document administratif ... mais une **aide concrète à la gestion des données** tout au long du projet

- Document qui décrit la façon dont les données seront : **obtenues, traitées, organisées, stockées, sécurisées, préservées, partagées**,... au cours et à l’issue d’un projet.

- Incite à la mise en place de **bonnes pratiques de gestion** à toutes les étapes du cycle de vie des données

Pour répondre aux principes FAIR de l’open data

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

Ouverture des données de la recherche

Par défaut, l'**ouverture des données de la recherche est obligatoire selon le principe "aussi ouvert que possible, aussi fermé que nécessaire"**.

Exceptions :

- Données **personnelles** non anonymisées ou pseudonymisées
- Données **non achevées** (sauf données géographiques qui peuvent être inachevées, directive INSPIRE), certaines données environnementales (cf protection des espèces)
- **Photos** quand il y a une personne reconnaissable
- Données **scrapées** (car souvent interdit par les conditions d'utilisation) sauf :
  - si le préjudice du producteur est nul (risque encouru nul également)
  - le scraping d'une partie non substantielle des données reste possible

Enjeux d’une bonne gestion des données de la recherche

- En premier lieu **faciliter le travail de recherche** lui-même, mais les enjeux sont plus larges :

  - Assurer la **conservation** des données à moyen terme
  - Pouvoir **reproduire** les résultats scientifiques, y compris par l’équipe qui les a obtenus !
  - Pouvoir facilement les **réutiliser** pour produire de nouvelles recherches
  - **Valoriser** les résultats scientifiques, mais aussi les données et les codes qui ont permis de les obtenir et en augmenter la visibilité
  - Favoriser de **nouvelles collaborations**, de nouvelles approches

- Enjeux d’ordre sociétaux :

  - Assurer une **souveraineté** sur les données produites
  - Assurer l’**intégrité scientifique**
  - Garantir la **transparence**, et assurer la confiance des citoyens en la recherche

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"

Contexte du site Grenoble Alpes : 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

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

Dans un environnement dynamique

EOSC, European Open Science Cloud

- Assurer **l'interopérabilité et le partage** des données au niveau européen
- En s'appuyant sur les infrastructures et services existants

RDA, Research Data Alliance

- Démarche **bottom up**.
- Accélérer et faciliter le partage et l'échange des données scientifiques
- De nombreux groupes de travail et d'intérêts
- Des recommandations et produits de référence
- Un **noeud RDA France**

Assurer la présence grenobloise dans ces projets, faire le lien avec les communautés scientifiques 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** ...

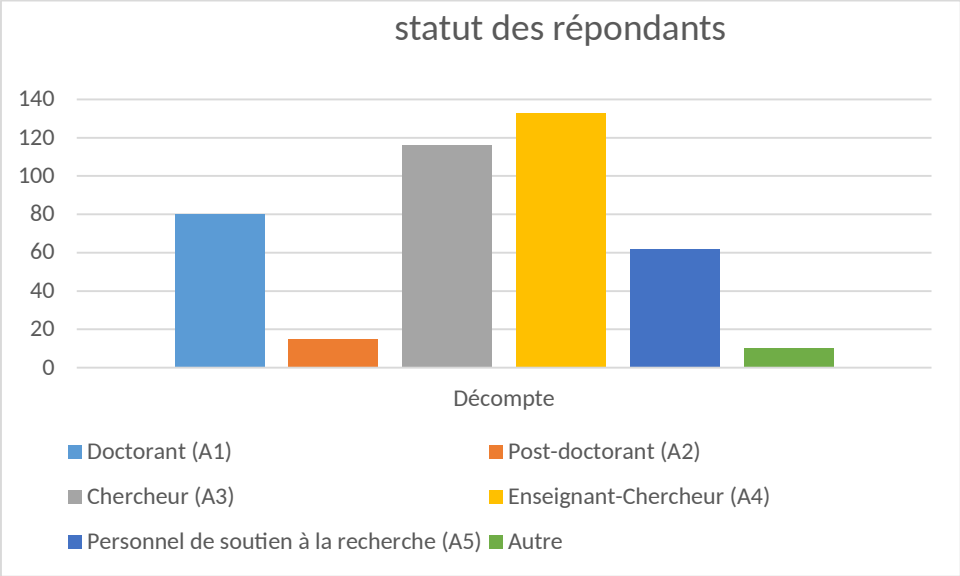
Quelques actions en cours

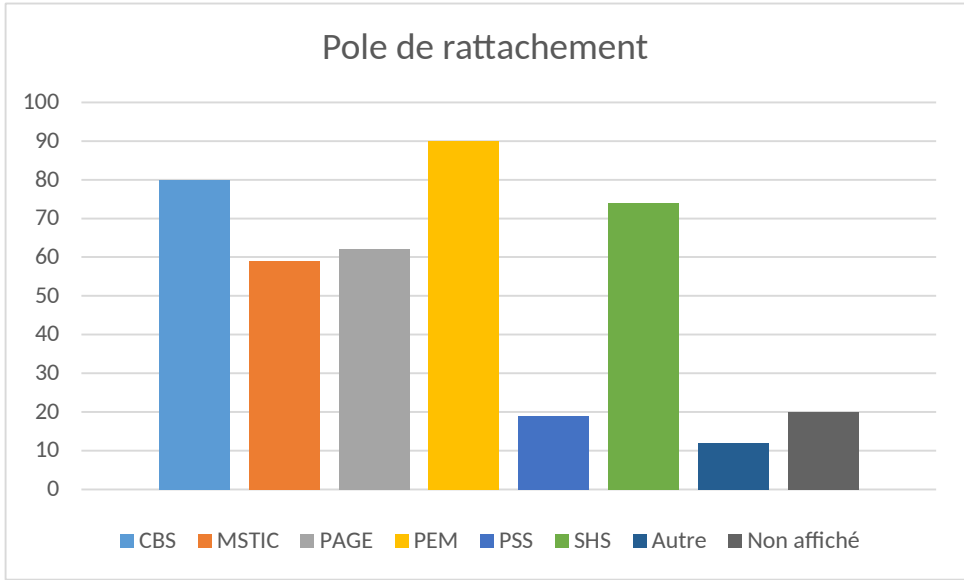
- **Enquête** autour des **besoins** des communautés du site
- Publication du **Baromètre Science Ouverte** du site intégrant les publications
- Actions de **formation** (collège doctoral, autre …) et d'**animation** (séminaires, ateliers …)
- Mise en place d'un **site web « science ouverte »**, point d'entrée unique
- Animation de la **liste de discussion**
- Mise en place d'un **entrepôt de données** pour le site, participation forte à l'initiative nationale **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

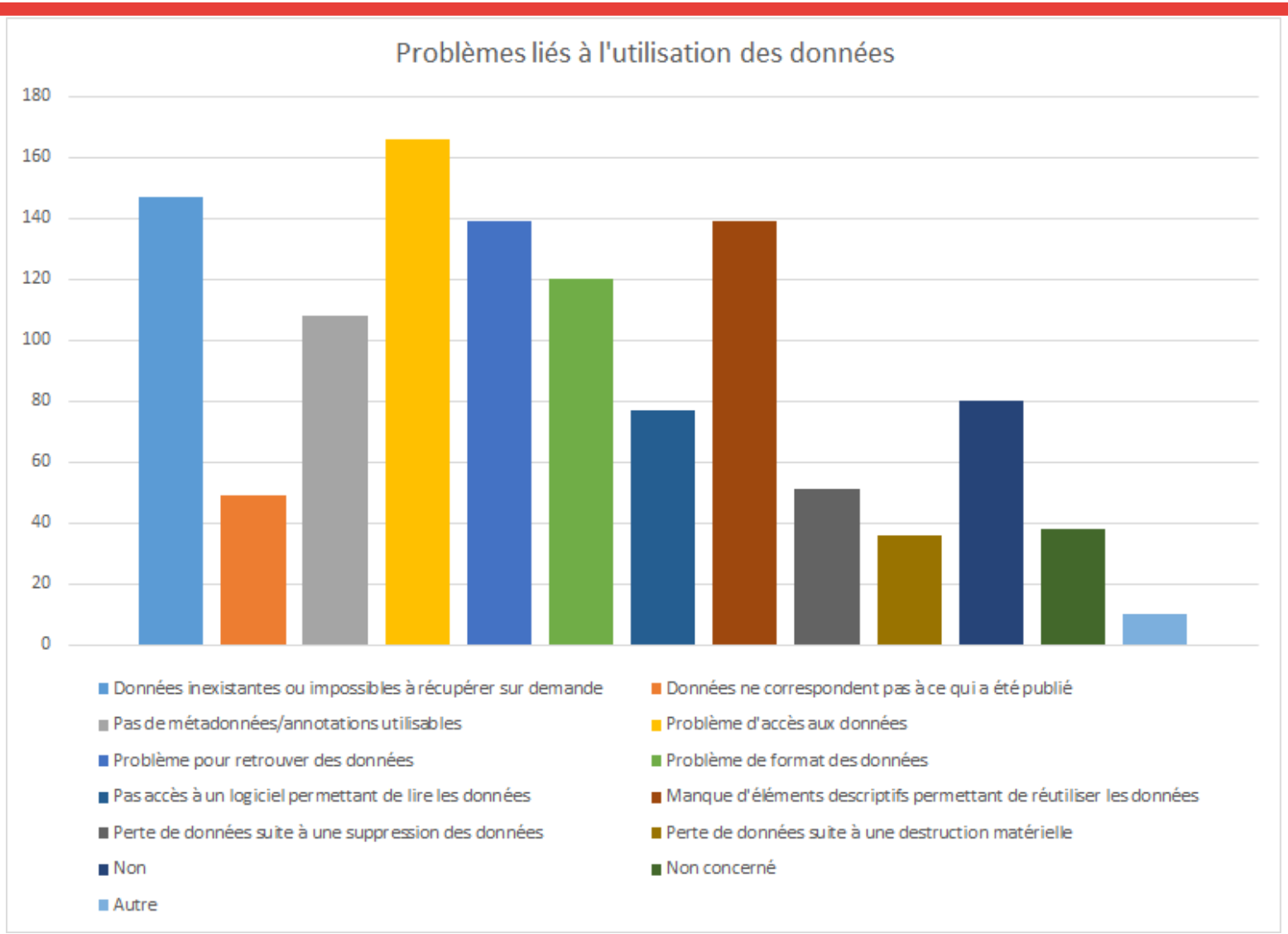
- **Objectif** :
  - recueillir les besoins des enseignants-chercheurs et des ingénieurs, afin de pouvoir mieux y répondre en termes d'accompagnement, de formation, de services, de ressources etc.
  - établir un état des lieux : quelles données, quelles pratiques des chercheurs, etc

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



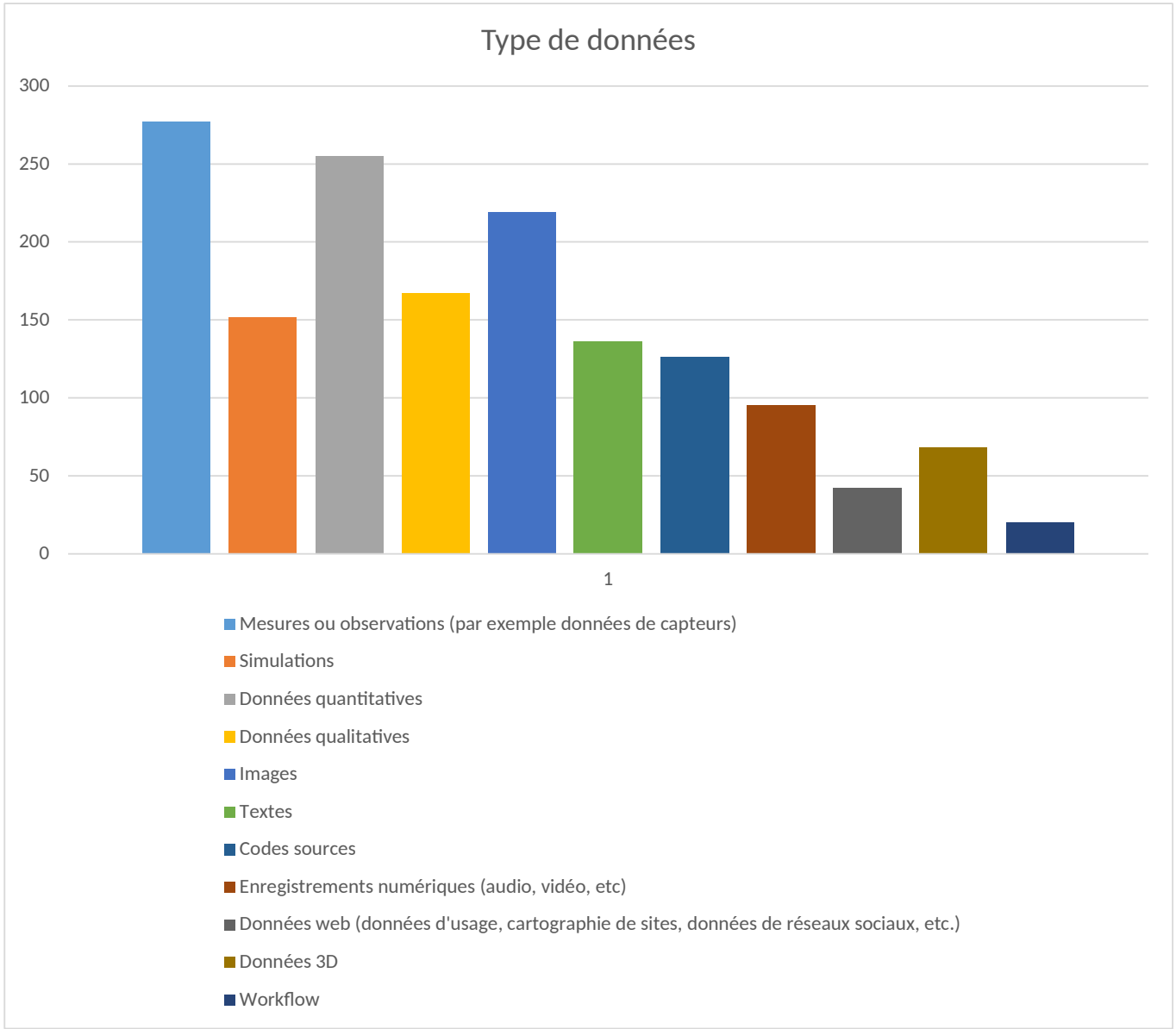


Quelques éléments de l'enquête

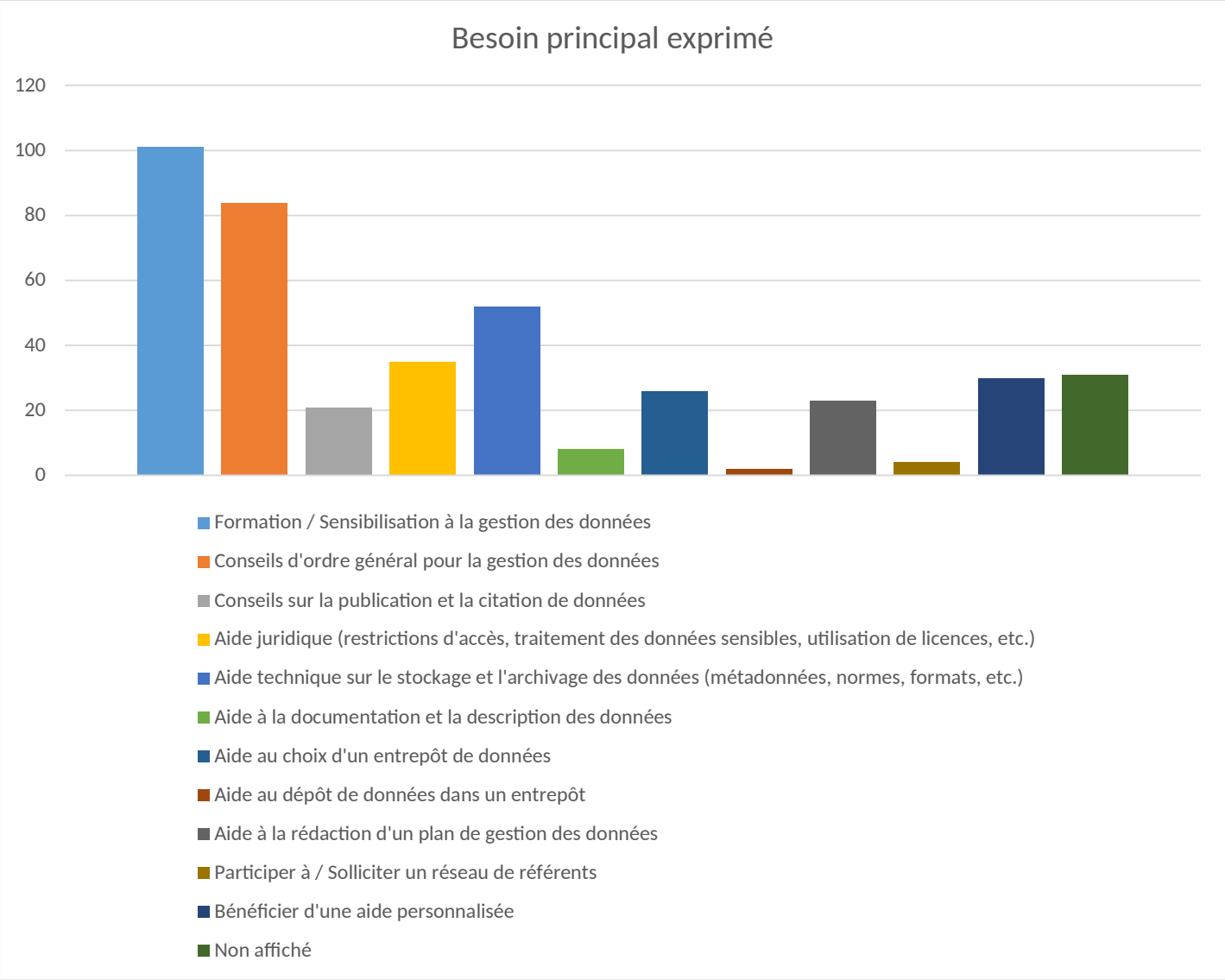


A noter : plus de 70 % des répondants (réponses complètes) ont déjà rencontrés des problèmes liés à l'utilisation des données

Quelques éléments de l’enquête



Quelques éléments de l’enquête



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 ?

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

- **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

→ Une **formation** du Collège des Ecoles Doctorales sur la gestion des données de la recherche

2/ Répondre aux questions réglementaires

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

- Groupe de travail RGPD, 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...

2/ Répondre aux questions réglementaires
Exemple de problématique

Dans le cadre d'un projet multi-partenaires avec d'autres sites : collecte de données (à partir d'un dispositif expérimental) et de données personnelles identifiantes.

Travail sur :

- La détermination des **données à conserver** dans les jeux de données pour préserver leur pertinence sans permettre l'identification des personnes
- La mise en place d'un **processus de pseudonymisation**
- La sécurisation de la **table de correspondance**
- La **sécurisation des données** pseudonymisées

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

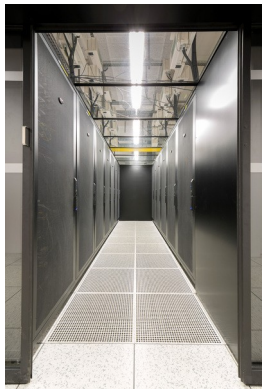
- Déterminer les **lieux et supports de stockage** selon le volume, la fréquence de consultation, le besoin en traitement et analyse, le besoin de partage, la durée de stockage …
- Identifier les **coûts**
- Qualifier les **données sensibles** et leur niveau de protection nécessaire
- Déterminer les **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



3/ Où et comment stocker ?
Exemple sur le partage des données

Dans le cadre de nombreux projets de recherche :

- Nécessité de **partager les données actives** entre tous les partenaires du projet
- Des **partenaires disséminés** sur tout le territoire national et à l'international
- La plupart du temps **inconnus des référentiels locaux**

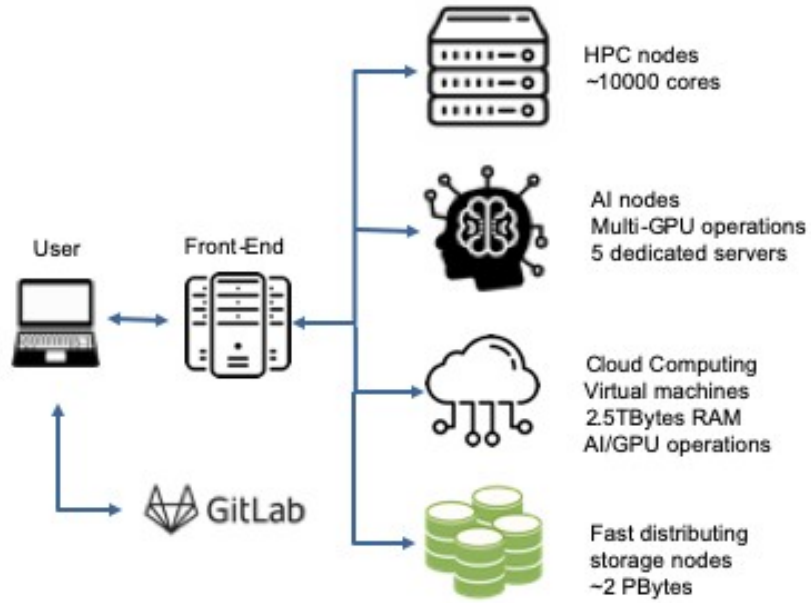
Mise en place d'espaces de stockage partagés :

- S'appuyant sur les **plateformes SUMMER et NOVA**
- Accessible par le **protocole ssh/scp** (multi-OS)
- Pour lesquels la **gestion des accès** se fait au cas par cas, de façon sécurisée
- Avec la possibilité de visibilité de ces espaces sur les **machines de calcul** pour traiter les données si besoin sans les recopier.

4/ Comment et où traiter ses données ?
Infrastructures disponibles

Orienter vers et accompagner sur les infrastructures adaptées aux besoins

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

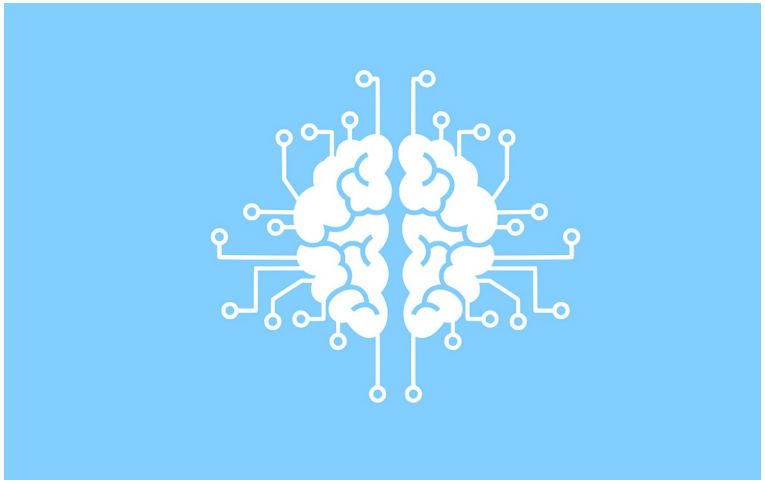


4/ Comment et où traiter ses données ?
Exemple autour de l'IA

- 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

Conclusions

Conclusions / contacts

- 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 (bientôt regroupé sur le site science ouverte de l'UGA) :
  - https://gricad.gricad-pages.univ-grenoble-alpes.fr/cellule-data-stewardship/web/

CELLULE DATA
GRENOBLE ALPES