CELLULE DATA
GRENOBLE ALPES

Accompagnement à la gestion des données

Séminaire Recherche Reproductible, 9 avril 2021

Violaine Louvet, GRICAD/CNRS

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
- Gestion des données et recherche reproductible
- 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 : DMP, open data, RGPD, …
- Et un volume de données à gérer en constante augmentation

Et une diversification des usages du numérique

- La 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 …

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

- **L'accompagnement à adapter**, des communautés "neuves" dans le domaine, une connaissance et une 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

- **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), H2020 (depuis 2017) et Horizon Europe (à partir de 2021), Feuille de route de la science ouverte du CNRS (2019), 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 :

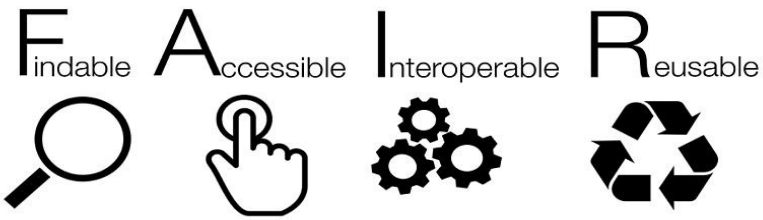
- un nième document administratif ... mais une **aide concrète à la gestion des données** tout au long du projet

- 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 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 Délégué à la Protection des Données (DPO, Data Protection Officer) 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
  - 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 « scienceouverte »**, point d'entrée unique
- Animation de la **liste de discussion**
- Réflexions sur les fonctionnalités nécessaires pour un **entrepôt de données** pour le site, lien avec les initiatives nationales.
- 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** ...

Gestion des données et recherche reproductible

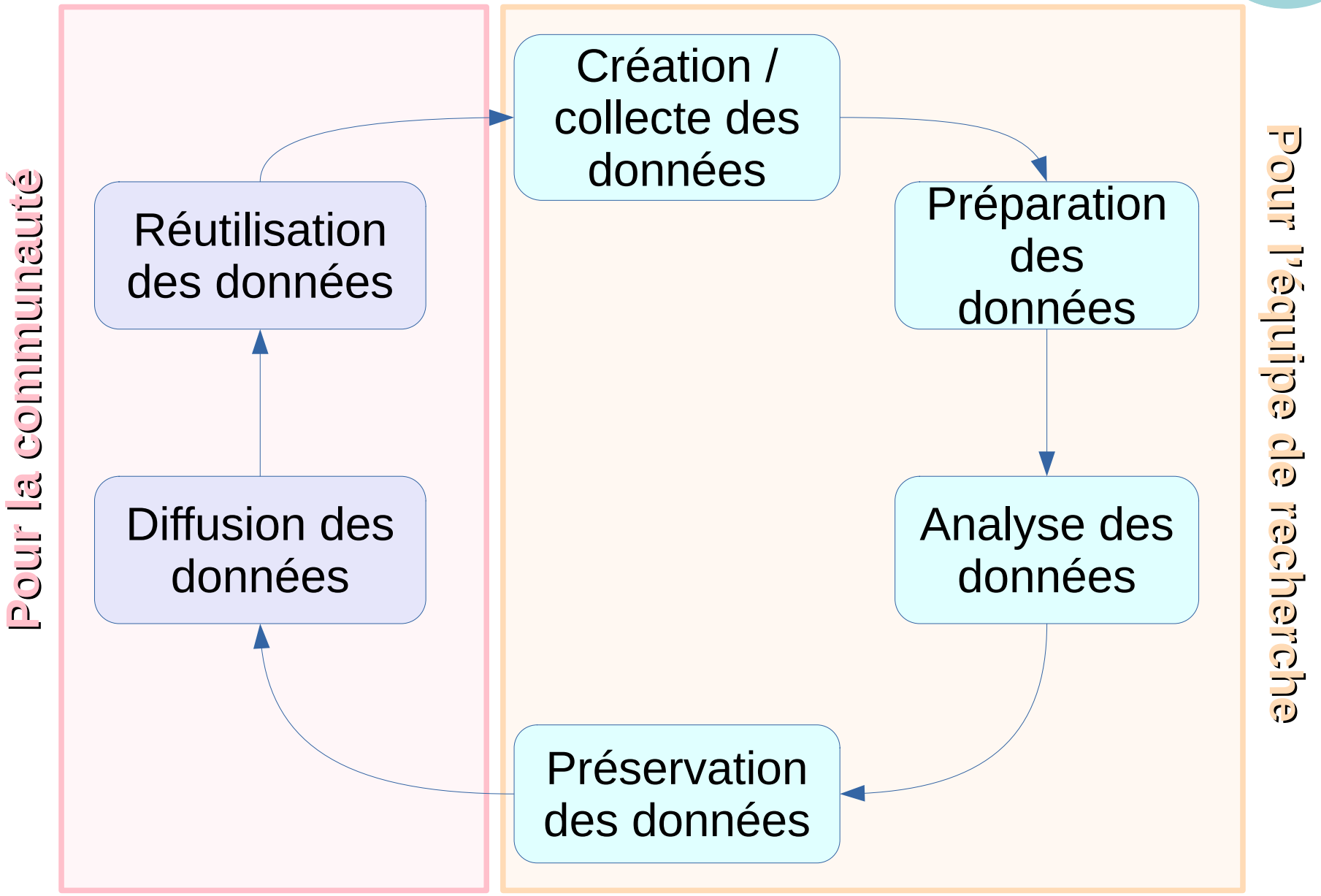
Données et recherche reproductible

- Données perdues, sur un DD qui vient de rendre l'âme ...
- Fichier de données illisible car format exotique, incompatible
- Fichier aux multiples versions _V0, _V9-1, _V2a … : laquelle est la bonne ?
- Où est donc ce fichier qui m'a servi à faire la courbe dans cette publi ?
- Sans parler de toutes les problématiques liées aux codes (qui sont aussi des données de recherche) !!

Les données sont souvent une cause majeure de non-reproductibilité

→ une bonne gestion des données est avant tout pour soi-même !

Bien gérer ses données : pour soi et pour les autres



Bien gérer ses données : le PGD

PGD = Plan de Gestion de Données (ou DMP, Data Management Plan) → un outil d'aide à la recherche reproductible

- 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 ?

PGD : poser les bases d'une recherche reproductible

- **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 propriétaire, 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

La cellule Data apporte une aide concrète et opérationnelle à la rédaction des PGD

Réfléchir à l’organisation de ses données

- Prendre du temps pour en gagner ensuite
- Etre pragmatique et réaliste : faire ce qu’il faut sans en faire trop !
- Pas de règle : la bonne organisation est celle qui vous convient
- Mais il faut prendre en compte le contexte : qui va utiliser les données ? Vous seul, l’ensemble de l’équipe, tous les partenaires du projet …
- N’oubliez pas de décrire l’organisation que vous aurez choisie, même si ce n’est que pour vous même !

En résumé, les différentes étapes :

- ➔ Déterminer la structure de l’organisation
- ➔ Définir la convention de nommage
- ➔ Spécifier la façon de gérer les versions
- ➔ Identifier les formats de fichiers des différents types de données

Se poser les bonnes questions sur le stockage

- 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**
- Anticiper le **partage** en hiérarchisant le contenu, les types, etc. et les autorisations d'accès
- Prévoir la **sécurisation** et les sauvegardes

Une formation stockage en préparation pour mai

Conserver et partager

- **Quelles données conserver, sur quelle durée ?**
  - Ne pas tomber dans le « piège » des big data et de la collecte et la conservation sans fin !
  - Evaluer la **valeur des données**
    - Sont-elles uniques ? Peut-on les reproduire et à quel coût ?
    - Sont-elles difficiles à obtenir ?
    - Faut-il conserver les données pré-traitées, les données analysées ?
  - Identifier les bonnes **durées de conservation** : 5 à 10 ans après la publication, plus de 10 ans si données difficiles à reproduire, à archiver pour des données non reproductibles et à forte valeur
- **Quelles données pour le partage ? Avec qui ? Sous quelles conditions ?**
  - Avec soi-même, l'équipe de recherche, toute la communauté
  - Tout de suite, après le projet, après un embargo, ...
  - « Aussi ouvert que possible, aussi fermé que nécessaire »
  - Avec une licence adaptée
  - Sur une plateforme ou un entrepôt approprié
  - Avec les métadonnées les plus riches possibles

Le coût des données

- Le stockage, et plus encore pour l'archivage, a un **coût financier et environnemental**.
- Ces deux indicateurs doivent aussi être utilisés pour déterminer à la fois la quantité de données à garder ainsi que la durée de cette conservation.

<u>Exemple de deux services proposés par l'UGA :</u>

- Calcul sur la plateforme DAHU : 1 heure.cœur émet **4,7 g CO_2e**

  → en 2019 : **8 millions heure.cœur** → **40 t CO_2e**

- Stockage sur la plateforme BETTIK : 1 Go.an émet **12 g CO_2e**

  → en 2019 : **1.4 millions Go.an** → **17 t CO_2e**

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 :
  - https://gricad.gricad-pages.univ-grenoble-alpes.fr/cellule-data-stewardship/web/

CELLULE DATA
GRENOBLE ALPES