

Direction de la Recherche et de l'Innovation

Le Plan de Gestion des Données en 4 questions

5 décembre





1^{ière} question

Qu'est qu'un PGD ?



Définitions et histoire

❖ Premiers PGDs :

- Années 1960 : PGD utilisé comme fiches synthétiques en sciences de l'ingénieur; Années 1970-80 : le PGD comme document cadre pour la gestion et la transformation de données en sciences de l'ingénieur
- Début des années 2000 : révolution numérique de la pratique scientifique, hausse des volumes de données digitales de la recherche.

Analyse 2003 au RU : « faute d'investissements dans les entrepôts de données, il y a risque d'un abandon et d'une perte de volumes énormes de données » notamment car les modalités de stockage et suivi étaient très limitées

- Début des années 2000 : le PGD est avant tout un objet pour faciliter l'appropriation de projets complexes dans des communautés spécifiques

❖ La révolution numérique et une appropriation institutionnelle :

- Deux leviers : la numérisation de la pratique de la recherche (e-recherche) et la pression économique
 - Massification du numérique : stockage hasardeux (disquettes, CD,... que reste il ?), réseaux informatiques pas toujours sauvegardés, etc...
 - Pression économique : position de l'OCDE en 2004 pour rendre ouvertes et accessibles les données de la recherche, à travers des dispositifs institutionnels formalisés.
- 2005-2007 : plusieurs sociétés savantes, organismes recommandent que le PGD soit utilisé pour penser la "préservation des données"
- Horizon 2020 (2013) fait « du libre accès aux données de recherche la règle générale », Horizon Europe a continué.
- MESRI, ANR, ANSES, etc... [déclaration](#) en 2020 qui appelle notamment à rédiger des PGD.

❖ Définition contemporaine :

- Document, attaché à un projet de recherche, à un moment de sa vie, qui précise les données collectées, traitées, réutilisées ainsi que leur gestion au cours du projet.
- Le PGD vise à cartographier le **cycle de vie des données** : collecte, analyse, usage, stockage, archivage et diffusion

Que contient un PGD ?

"Plans are worthless, but planning is everything" Dwight Eisenhower

- ❖ Plusieurs financeurs demandent un PGD initial et un PGD final :
 - un PGD n'est pas un contrat, c'est la matérialisation d'une série de questions méthodologiques sur le travail fait ou à venir
 - si les modalités du projet changent (membres, objectifs, financements), le PGD doit être repensé
- ❖ Le PGD a vocation à être un document court et opérationnel
 - renvoyer à des normes (ex ne pas réexpliquer pourquoi tel fichier est utilisé dans votre discipline)
 - il ne s'agit pas de justifier une collecte, mais de préciser l'intention (ce n'est pas nécessaire de justifier ses hypothèses)
- ❖ "Monsieur, je n'ai pas de données dans mon projet de recherche !"
 - La notion de données est polysémique et recouvre un vaste horizons d'objets : tableaux de chiffres, images, retranscriptions, un corpus de textes, etc...
 - Extraits de DARIAH (<https://campus.dariah.eu/>) – une plateforme EU pour les chercheurs en art et humanités employant des outils informatiques/des ressources numériques :
 - Au lieu de "données" on parle généralement de "sources primaires"
 - Une parties des "données" n'est pas un objet qui appartient à un projet de recherche mais est un objet "culturel" qui appartient à des institutions comme des musées, des bibliothèques



Que contient un PGD ?

❖ Il n'y a pas de modèle unique de DMP !

- Modèles spécifiques pour certains financeurs : ANR, ERC, H2020, INSERM,...
- Modèle produits par certaines universités (en lien avec un workflow de validation)
- Modèle « de norme » : celui de Science Europe, pensé pour être *machine actionnable*

❖ Éléments :

- Organisé par "produit de la recherche" (jeu de données, sous-ensemble de documents, série d'images, etc... - tout ce qui forme un ensemble cohérent)
- Pour chaque produit :
 - typologie des données
 - comment sont elles produites ? sont elles réutilisées ?
 - les métadonnées ?
 - format des données, notamment lorsqu'il s'agit de fichiers informatiques qui peuvent dépendre d'outils spécifiques
 - degré d'organisation des produits (index, nomenclatures, architecture de dossiers, etc...)
 - modalités de stockage, de sauvegarde, d'archivage
 - questions de droit : RGPD, éthique, ...
- Pour l'ensemble du projet :
 - finalité, description du projet
 - modalités financières
 - personnes mobilisées et référentes



Les PGD et les principes FAIR

- ❖ Résumé d'un tableau comparatif (Guide pratique pour une harmonisation internationale de la gestion des données de recherche - V2 sur le site "Ouvrir la Science")
 - Facile à trouver
 - Métadonnées exhaustives
 - Jeux de données avec un identifiant unique et pérenne
 - Accessible
 - Les jeux de données sont accessibles à travers un outil connu, ouvert
 - Les modalités d'accès aux jeux de données sont paramétrables (possibilité de restreindre l'accès ou demander une authentification)
 - La description des jeux de données (métadonnées) sont publiques et ouverts même si les données elles mêmes ne le sont pas
 - Interopérable
 - Les métadonnées et données sont structurées, normées selon les standards de la discipline
 - Les jeux de données renvoient vers d'autres jeux de données pertinents
 - Réutilisable
 - Les données sont enregistrés dans des formats qui peuvent être utilisés facilement par d'autres chercheurs
 - Les métadonnées précisent les modalités d'enregistrement
 - Précision sur la licence de réutilisation (CC)





2^{ième} question

Pourquoi est qu'on fait des PGD ?



Motivations macro pour les PGD

❖ Une pression venue d'en haut :

- Demande des financeurs : l'UE, l'ANR, ... souhaitent inscrire leur action de soutien de la recherche dans les pratiques de la SO
- Cadre réglementaire qui s'est précisé à ce sujet :
 - Article L112-1 du Code de la Recherche (modifié en 2013) :
La recherche publique a pour objectifs : [...]
 - b) La valorisation des résultats de la recherche au service de la société, [...]*
 - c) Le partage et la diffusion des connaissances scientifiques en donnant priorité aux formats libres d'accès [...]*
 - e) L'organisation de l'accès libre aux données scientifiques.*
 - Loi n° 2015-1779 du 28 décembre 2015 relative à la gratuité et aux modalités de la réutilisation des informations du secteur public : facilitation de la réutilisation des données produites par la fonction publique puis **Loi pour une République numérique (2016)**.
 - Décret n° 2021-1572 du 3 décembre 2021 relatif au respect des exigences de l'intégrité scientifique :
Les établissements publics [...] promeuvent la diffusion des publications en accès ouvert et la mise à disposition des méthodes et protocoles, des données et des codes sources associés aux résultats de la recherche afin d'en garantir la traçabilité et la reproductibilité. Ils incitent à la publication des résultats de recherche dits négatifs
Les établissements publics [...] définissent une politique de conservation, de communication et de réutilisation des résultats bruts des travaux scientifiques menés en son sein. A cet effet, ils veillent à la mise en œuvre par leur personnel de plans de gestion de données et contribue aux infrastructures qui permettent la conservation, la communication et la réutilisation des données et des codes sources.
- Précisions INRAE (Principes de gouvernance des données du 10 /12/2020)
 - "Une bonne gestion des données consiste à [...] réaliser des plans de gestion des données"
 - "Le PGD, [...], est pris en charge, pour l'essentiel, par le porteur du projet scientifique dès le montage de celui-ci, avec l'appui local des réseaux fonctionnels concernés (IST, CPI, DPO si données personnelles, etc.). Le porteur du projet scientifique organise et s'assure de la mise en œuvre du PGD au sein de l'équipe"

❖ Nécessités au sein de certains domaines disciplinaires

- Etudes sur l'humain/cliniques : nécessité d'anticiper largement en avance les projets
- Humanités numériques : bcp d'enjeux liés aux données recueillies et leur traitement



Motivations individuelles pour les PGD

- ❖ Finalité des PGDs : (garder en tête FAIR)
 - Tracer la vie des données produites : qu'est que j'en fais, qu'est qu'elles deviennent ?
 - Permettre aux parties prenantes (financeurs, employeurs,...) de veiller au respect des principes de la SO
 - Faciliter la réutilisation des données (penser d'autres emplois)
 - Faciliter la reconnaissance du travail de chacun (continuité avec les cahiers de labos)
 - Faciliter le traitement de réclamation en cas de questions liées à l'éthique ou l'intégrité
 - Permettre une cartographie de jeux de données locaux (ex : au labo)
 - Permettre une meilleure reproduction des travaux
 - Anticiper des questions liées à l'anonymisation des données
 - Penser la nécessité de solliciter d'autres parties pour la collecte, le stockage ou le partage de données
 - Systématiser l'évaluation de la nécessité de stocker des données au long terme : est que j'ai retiré toute la valeur de ma "donnée" brute ? est que le bénéfice possible de garder la donnée brute justifie le cout de stockage pérenne de celle-ci ?
- ❖ Votre recherche est (en partie au moins) financée par l'argent public : le PGD permet de rendre visible vos données produites au-delà de votre ordinateur ou de votre équipe.





3^{ième} question

Est que c'est vraiment utile un PGD ?



Oui – et alors ?

- ❖ Ces raisons là ne me motivent pas ...
- ❖ La plupart des questions que le PGD vous amènera à évaluer vous seraient venues "au fil de l'eau" du projet. Il s'agit ici de les penser dès le départ et d'agir en conséquences.

"Les DMP ne devraient pas apparaître comme une tâche administrative supplémentaire, mais un moyen utile d'aide à la planification et la réalisation d'un projet de projet de recherche "

Science Europe - Guide pratique pour une harmonisation internationale de la gestion des données de recherche - V2

- ❖ Il s'agit de formaliser vos pratiques :
 - anticiper la collecte de certaines métadonnées au cours du travail de recherche
 - structurer de façon efficiente, rigoureuse votre collecte de données et leur stockage
 - anticiper des questions liées (par exemple, espace de stockage)
 - éviter la duplication des données, la perte de données, les risques de piratage

Anticiper plutôt que réagir !



Le PGD comme levier du changement des pratiques

- ❖ La littérature grise (ensemble du savoir/des données qui n'existent pas formellement) : comment limiter la perte de sources/données éphémères ?
 - Lors de la rédaction d'un PGD, se poser la question des éléments de contexte qui peuvent être perdus s'ils ne sont pas enregistrés dans le projet
- ❖ Penser au long terme : on assiste un mouvement de "technologisation" des outils de la recherche (cahier de laboratoire électronique, e-recherche dont LIMS)
 - Développement d'outils de stockage locaux et d'entrepôts, leur utilisation nécessite qu'on décrive correctement les données – rédiger le PGD revient en décrire une grande partie.
 - Les chantiers de l'éthique et l'intégrité nous invitent à formaliser de plus en plus les pratiques de recherche
 - Standardisation de certaines pratiques de recherche à l'échelle internationale (ex enquêtes dans le Data Documentation Initiative)
- ❖ Normer ses jeux de données pour qu'ils puissent conformer aux standards en vigueur
 - Outils de partage (ex Dataverse)
 - Aligner sa description par les métadonnées sur des normes (ex les champs du référentiel Dublin Core)





4^{ième} question

**Quels outils et accompagnement sont
à disposition pour rédiger un PGD ?**



Un outil à connaître

- ❖ Une plateforme collaborative : OPIDOR <https://dmp.opidor.fr/>
 - D'autres plateformes existent : <https://argos.openaire.eu/splash/> et <https://ds-wizard.org/>
 - OPIDOR reste l'outil recommandé par le GT SO Données de Couperin
- ❖ Fonctionnalités utiles :
 - La possibilité de travailler à plusieurs (autres chercheurs, services d'appui)
 - Possibilité de copier un PGD (version initiale vers version finale)
 - Accessible en Français et en Anglais
 - Possibilité de récupérer des données depuis l'ANR
 - Modèles fournis nombreux : ANR, Horizon Europe
 - Possibilité de développer un espace institutionnel (personnalisation des conseils)



Outils informatiques

- ❖ Rappel important : « **les établissements [...]** veillent à la mise en œuvre par leur personnel de plans de gestion de données »

Deux idées :

- L'établissement *accompagne et contrôlent* la rédaction des DMP
- Ce sont les chercheurs qui doivent rédiger les DMP

Corollaire : on ne sous-traite pas son DMP à son doctorant !

- ❖ Ressources humaines :

- Personnel d'accompagnement : Guichet de la Donnée, personnel d'appui MSHS
- Agrégation des expériences : mise en commun de DMP au sein d'une équipe de recherche (ex CeRCA)



Quelques ressources pour les PGD

- ❖ Des ressources (qu'on va cartographier au cours des mois à venir) nombreuses :
 - <https://www.couperin.org/science-ouverte>
 - <https://scienceouverte.couperin.org/>
 - <https://www.ouvrirlascience.fr>
 - <http://weburfist.univ-bordeaux.fr/> (partenariat qui se construit)
 - <https://www.openaire.eu/>

- ❖ Des entrepôts de données :
 - PROGEDO : réseau universitaire sur les données (stockage, usage, curation) spécialisées en SHS
 - Recherche Data Gouv : qui fera (d'abord) portail d'identification de jeux de données stockés ailleurs



Éléments à retenir

- ❖ Le PGD est un outil de formalisation de bonnes pratiques scientifique autour du cycle de vie des données
- ❖ Les PGD ont vocation à devenir la norme pour tous les projets de recherche
- ❖ On dit que les données/publications doivent être "aussi ouvertes que possible, aussi fermées que nécessaire", on peut dire que le PGD doit être une étape de réflexion "aussi utile que possible, aussi exhaustive que nécessaire"
- ❖ Ne pas réinventer la roue :
 - Vos collègues de bureau ont peut-être rédigé des PGD, peut-être que leurs descriptions des jeux de données sont similaires aux vôtres ?
 - Vous avez des collègues BIATSS qui peuvent vous accompagner – sollicitez les ! Ils sont pertinents notamment sur certains aspects techniques précis (ex : les entrepôts de données)

