ADN Accompagnement à la donnée à et avec Nanterre

Ouvrir l'accès aux données de la recherche ?

De la gestion à l'ouverture

Octobre 2024

Cécile Delay-Artous

Service commun de la documentation de l'université Paris Nanterre

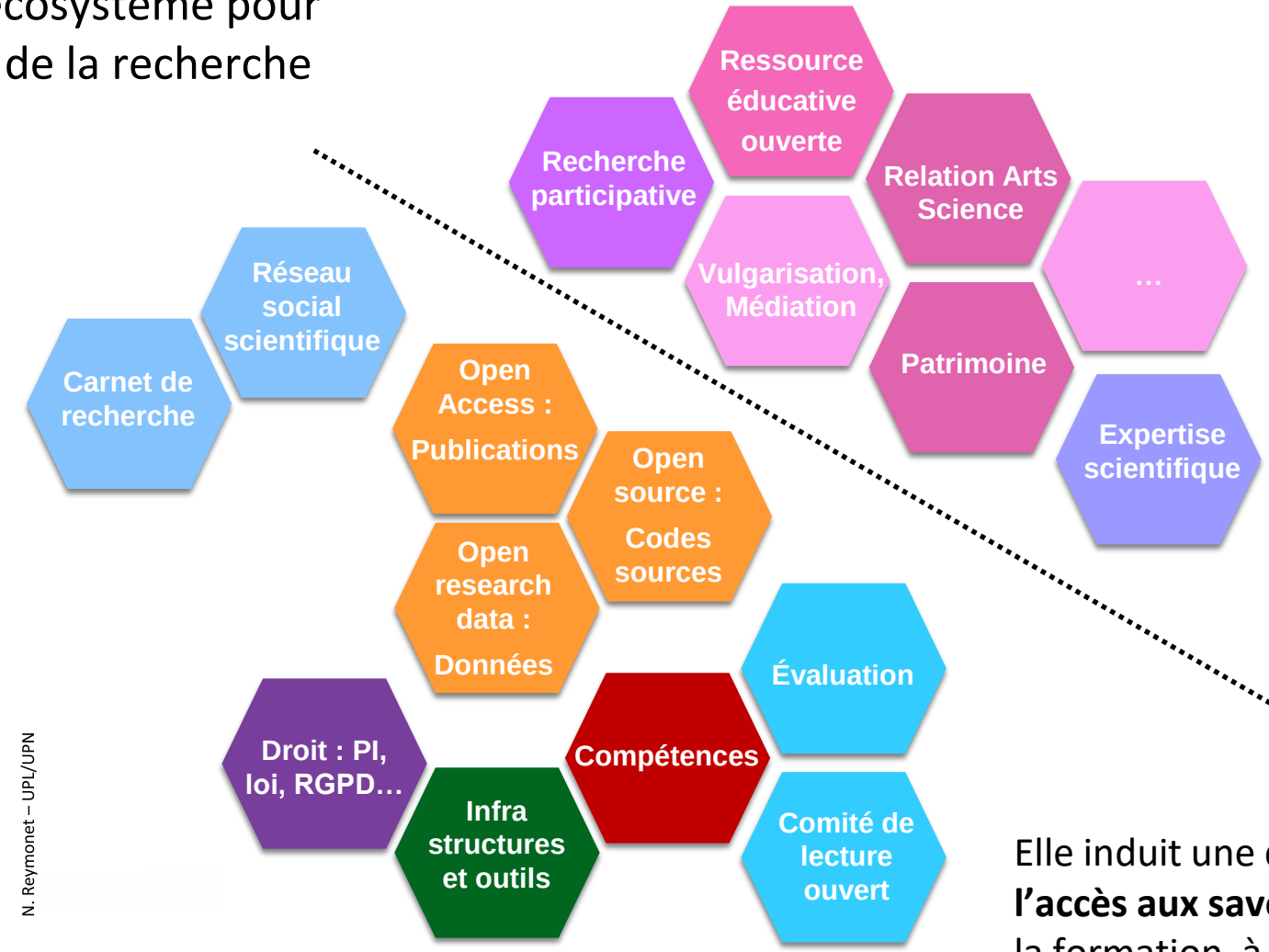
MSH MONDES · Université Paris Nanterre · LA CONTEMPORAINE

Un contexte, quelques définitions

Et beaucoup de questions.

Un contexte

La science ouverte : un écosystème pour un partage des résultats de la recherche



N. Reymonet – UPL/UPN

Elle induit une **démocratisation de l'accès aux savoirs**, utile à la recherche, à la formation, à l'économie, à la société.

La science ouverte

La science ouverte est la **diffusion sans entrave** des résultats, des méthodes et des produits de la recherche scientifique. Elle s'appuie sur l'opportunité que représente la mutation numérique pour développer l'**accès ouvert aux publications** et - autant que possible - **aux données, aux codes sources et aux méthodes de la recherche**.

Le **mouvement de la Science ouverte** vise à construire un écosystème dans lequel la science sera plus cumulative, plus fortement étayée par des données, plus transparente, plus rapide et d'accès universel.

Comité pour la science ouverte (2021). Deuxième Plan national pour la science ouverte

Plus d'infos : consulter le portail Science ouverte de l'UPN

Qu’est-ce qu’une donnée, ou un jeu de données, ou des données, de ou de la recherche ?

Les données de recherche

La définition la plus couramment utilisée est celle de l'OCDE :

Les données de la recherche sont définies comme des **enregistrements factuels** (chiffres, textes, images et sons), qui sont utilisés comme **sources principales** pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider les résultats de la recherche.

Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics (2006)

Une définition plus large et plus opérante :
Il s'agit de l'**ensemble des informations collectées, observées ou créées sous forme numérique dans le cadre d'un projet de recherche**.

Deboin, M.-C. (2020). S'initier en ligne aux données de la recherche et à leur gestion. CIRAD.

Données de la recherche : une diversité d'objets

Textes

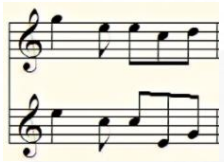
Archives

Manuscrits

```
foreach my $attribut (@fields)
    my $error_message
        = 'Fichier de configur
        . "<$attribut> non tro
        . $end_mess_error
        . ". Verifiez $CONFIG_
```

Code

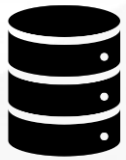
Enregistrements sonores



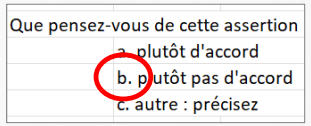
Partitions

Photographies

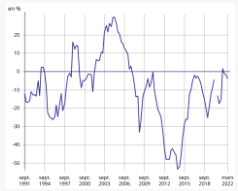
Cartes et plans



Spatiales ou temporelles



Résultats d'enquête



Séries statistiques

Vidéos

Images 3D

etc.

Données de recherche : une typologie

Types de données	Définitions	Valeur et spécificité	Exemples
Données d'observation *(Observational datas)*	Données obtenues en temps réel	Souvent uniques et irremplaçables	Données atmosphériques, d'enquêtes, échantillons, neuro-image
Données expérimentales *(Experimental datas)*	Données obtenues en laboratoire à partir d'équipements spécifiques	Reproductibles mais à des coûts dissuasifs	Séquence de génome, chromatographie, spectres RMN
Données de simulation *(Simulation datas)*	Données générées à partir de modèles test	Métadonnées et modèles ont une valeur supérieure aux résultats	Modèles climatiques, reconstitution en 3D de monuments historiques
Données dérivées ou compilées *(Derived or compiled data)*	Données obtenues par compilations ou traitement des données brutes	Reproductibles mais à des coûts dissuasifs	Texte et *data mining*, bases de données compilées
Données de référence ou données canoniques (*Reference or canonical datas)*	Collections statiques ou organiques de jeux de données validées	Données publiées ou qui ont fait l'objet d'une curation	Portail de l'Unesco sur les pays, banque de données sur le génome, structures chimiques

Quel rapport avec la science ouverte ?

Le cycle de vie des données scientifiques

Le cycle de vie des données de recherche, c'est l'ensemble des étapes de gestion, conservation, diffusion et réutilisation des données scientifiques liées aux activités de recherche. (CIRAD)

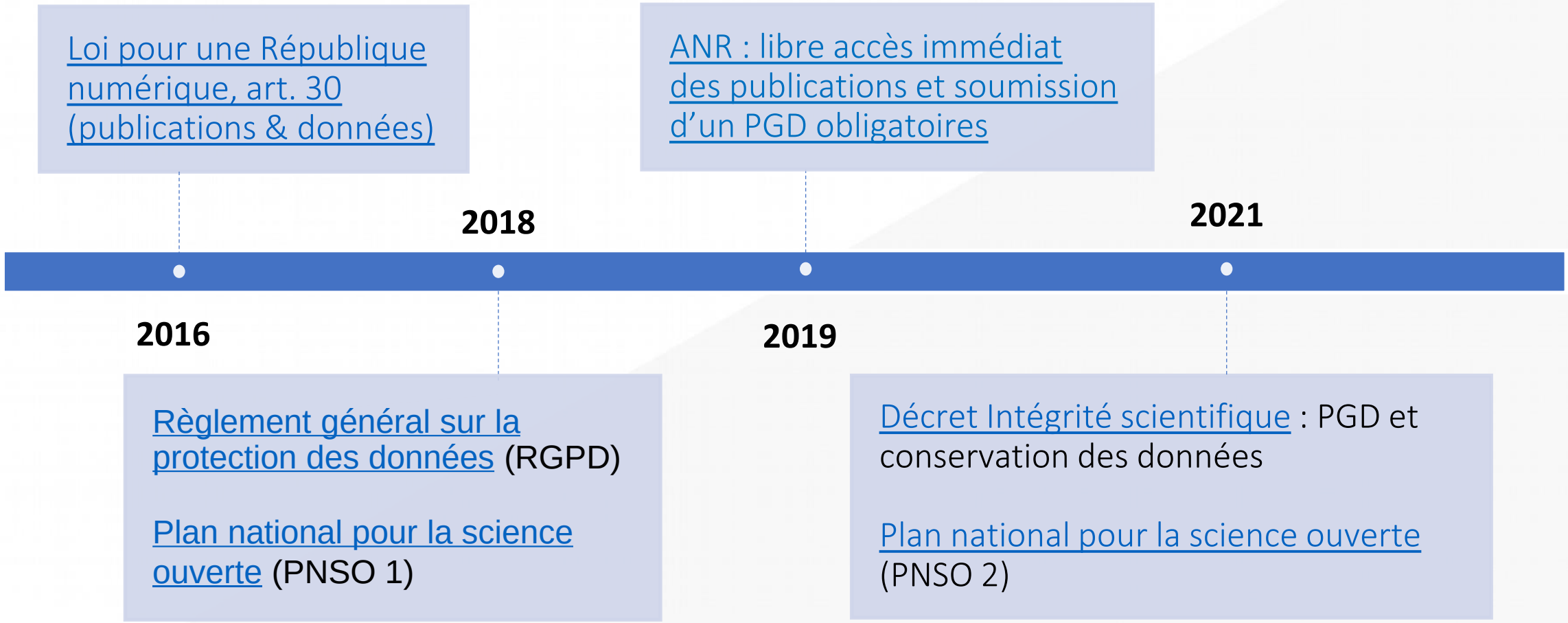
1. Planification
2. Collecte / Création
3. Traitement / Analyse
4. Accès / Partage
5. Préservation
6. Réutilisation

DoRANum (2021). Le cycle de vie des données de la recherche. DOI : 10.13143/gzj2-j593

Mais alors, on ouvre ?

Spoiler : ça dépend.

Évolution de la réglementation



Et nous voici confrontés à des injonctions qui peuvent sembler contradictoires.

Les données à caractère personnel : une exception à l'ouverture

Règlement général sur la protection des données (RGPD) : protection des personnes physiques à l'égard du traitement des données personnelles et à la libre circulation de ces données.

= toute information relative à une personne physique susceptible d'être identifiée, directement ou indirectement.

On peut parfois contourner cette difficulté :

➞ **Anonymisation** : traitement qui vise à rendre impossible, en pratique, toute identification de la personne par quelque moyen que ce soit et de manière irréversible (pas de réidentification des personnes par recoupement).

➞ **Pseudonymisation** : traitement qui rend impossible l'attribution des données relatives à une personne physique sans information supplémentaire : remplacer les données directement identifiantes (nom, prénom, etc.) d'un jeu de données par des données indirectement identifiantes (alias, numéro séquentiel, etc.).

Contact : dpo@liste.parisnanterre.fr

Les autres exceptions à l'ouverture

- Secret **médical**
- Secret **industriel** ou commercial : influencer la position sur un marché ou révéler des procédés de fabrication
- Secret **statistique**
- Sécurité des systèmes d'information des **administrations**
- Secret **défense**, sûreté de l'État (données stratégiques), sécurité publique
- Données sur lesquelles des tiers ont **priorité d'exploitation**
- Données sur lesquelles des tiers ont des droits de **propriété intellectuelle**

Reste la question : doit on TOUTES les ouvrir dès qu'on le peut ?

Comment savoir quand ouvrir ?

Depuis la *Loi pour une République numérique*, **ouverture par défaut** des données de la recherche (publique) sauf -nombreuses- exceptions.

Ces données doivent être diffusées en **libre accès** (ou a minima être **communicables à la demande**) et ce, **gratuitement**.

Lionel Maurel (2021), Quel cadre pour l'ouverture des données de la recherche dans le contexte de la science ouverte ?

En cas de doute, il existe plusieurs outils d'aide à la décision :

- Arbre de décision du CIRAD
- Logigramme de l'École des Ponts ParisTech
- Logigramme d'INRAE
- Logigramme de l'Institut Pasteur

et pour savoir où déposer ses données : un logigramme de Recherche Data Gouv.

Une ouverture raisonnée des données

Le principe « Aussi ouvert que possible, aussi fermé que nécessaire » reste la règle dans le domaine de la recherche.

Mais à qui appartiennent vos données (de recherche financée sur fonds publics) ?

Propriété des données

Dans leur activité de recherche académique, les chercheurs (enseignants-chercheurs, chercheurs, doctorants, autres personnels de la recherche) produisent des **données « publiques »** :

- Ils/elles n'ont pas de propriété intellectuelle sur les données factuelles collectées (uniquement sur les œuvres de l'esprit, le critère étant l'originalité).
- Les établissements « producteurs » ont un droit de propriété intellectuelle sur la structuration des données qui donne lieu à une base de données, mais **pour les établissements publics, c'est l'ouverture par défaut qui prévaut** (Loi pour une République numérique, art. 11).

La question de la propriété des données n'est pas forcément le bon angle d'attaque depuis la Loi pour une République numérique (et le principe d'ouverture par défaut). Aujourd'hui, savoir qui est le propriétaire des données n'est souvent plus nécessaire. Il s'agit de **savoir si on est dans le principe de l'ouverture par défaut ou bien dans un cas d'exceptions**.

DoRANum et Tuto@Mate (2020). Lionel Maurel, « À qui appartiennent les données »

Pourquoi gérer ses données – même lorsqu'on ne peut les ouvrir ?

Pourquoi gérer ses données

Pour ne pas risquer de perdre ou d'altérer ses données → **Stabilité** des données

Pour pouvoir les conserver à long terme → **Pérennité** des données

Pour améliorer la finesse de ses données → **Qualité** des données

Pour garantir l'intégrité de la recherche → **Transparence** et **reproductibilité** des données

Pour pouvoir les partager plus facilement → **Réutilisation** et **visibilité** des données

Et (aussi)… pour aller dans le sens des orientations institutionnelles (Et augmenter les chances d'obtenir un financement).

Gestion des données, FAIRisation et bonnes pratiques

FAIRisation des données

L'objectif des principes FAIR est de favoriser la **découverte**, l'**accès**, l'**interopérabilité** et la **réutilisation** des données partagées.

Chaque principe FAIR se décline en un ensemble de caractéristiques que doivent présenter les données et les métadonnées pour faciliter leur découverte et leur utilisation par les hommes mais aussi par les machines.

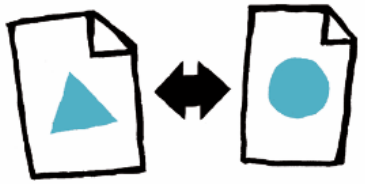
INRAE (2023). [Produire des données FAIR]

Ministère de l'Enseignement supérieur et de la Recherche (2024).

[Passeport pour la science ouverte]

Le principe ***Findable*** (Facile à trouver) a pour objectif de faciliter la découverte des données par les humains et les systèmes informatiques et requiert une description et une indexation des données et des métadonnées.

Le principe ***Accessible*** encourage à stocker durablement les données et les métadonnées et à faciliter leur accès et/ou leur téléchargement, en spécifiant les conditions d'accès (accès ouvert ou restreint) et d'utilisation (licence).



Le principe ***Interoperable*** peut se décomposer en : téléchargeable, utilisable, intelligible, et combinable avec d'autres données, par des humains et des machines.

Le principe ***Reusable*** (Réutilisable) met en avant les caractéristiques qui rendent les données réutilisables pour de futures recherches ou d'autres finalités (enseignement, innovation, reproduction/transparence de la science).

Quels outils pour bien gérer (et FAIRiser) ses données ?

Les licences ouvertes

Une licence ouverte garantit à tous le droit d'utiliser, de partager et d'accéder à vos données avec la sécurité juridique nécessaire aux producteurs et aux ré-utilisateurs des données.

DoRANum (2022), Guide des licences ouvertes DOI : 10.13143/tv6f-sv31

Principales licences ouvertes :

- Licences Creative Commons
- Licence Etalab
- Licences logicielles (GNU, BSD, Apache, MIT, CeCILL)
- Licences bases de données (ODbL, PDDL, OCD-By)

CopIST, Dedieu, L. (2015), Rendre publics ses jeux de données scientifiques

Les identifiants pérennes

Numéro ou une entité alphanumérique, permettant de désigner et de **retrouver de manière univoque et pérenne un objet, un document, une personne, un lieu, un organisme**, ou toute entité, dans le monde physique ou numérique.

Des identifiants ouverts pour la science ouverte : note d'orientation du Comité pour la science ouverte (2019)

Un identifiant pérenne ou PID (persistent identifier) est :

- Unique : il se réfère à une seule ressource
- Pérenne : sa durée est garantie dans le temps
- Résolvable : il permet d'accéder à la ressource ou au moins à une description de celle-ci

Les identifiants *objet*, pour identifier et retrouver les productions scientifiques

Plusieurs identifiants objet existent dans le périmètre science ouverte mais c'est le DOI qui est recommandé.

DOI = Digital Object Identifier

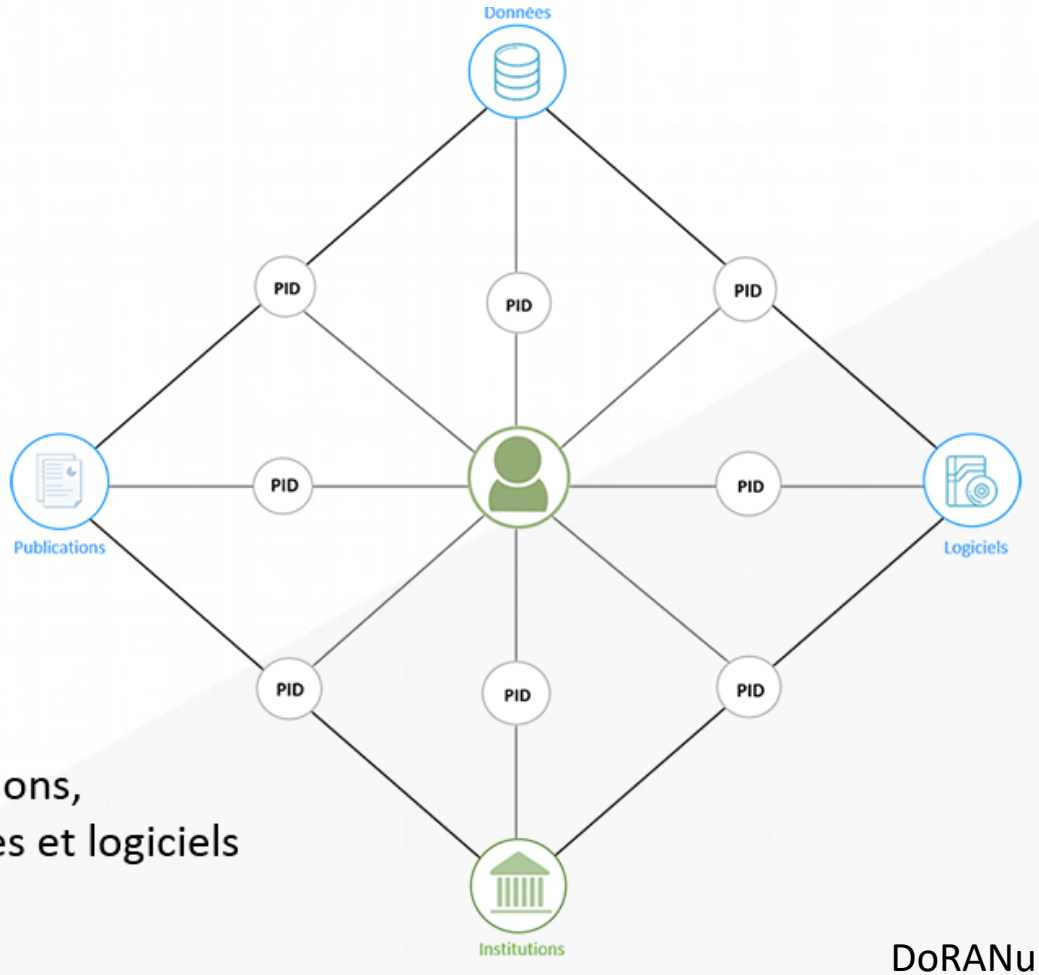
Certains entrepôts de données attribuent automatiquement un identifiant pérenne aux données.

Par exemple nakala ID : 10.34847/nkl.7dcdw459 Ou RÉPUBLIQUE FRANÇAISE recherche.data.gouv.fr doi:10.57745/NTSGDF

Les identifiants *contributeur,* pour identifier les auteurs...

Plusieurs identifiants contributeur existent dans le périmètre science ouverte mais c'est l'OrcID qui est le plus utilisé à l'international.
On peut « lier » un OrcID avec d'autres identifiants auteur, comme l'IdHAL ou IdRef.

...Et faire le lien entre les produits de la recherche et leurs auteurs.



Lier auteurs, institutions, publications, données et logiciels

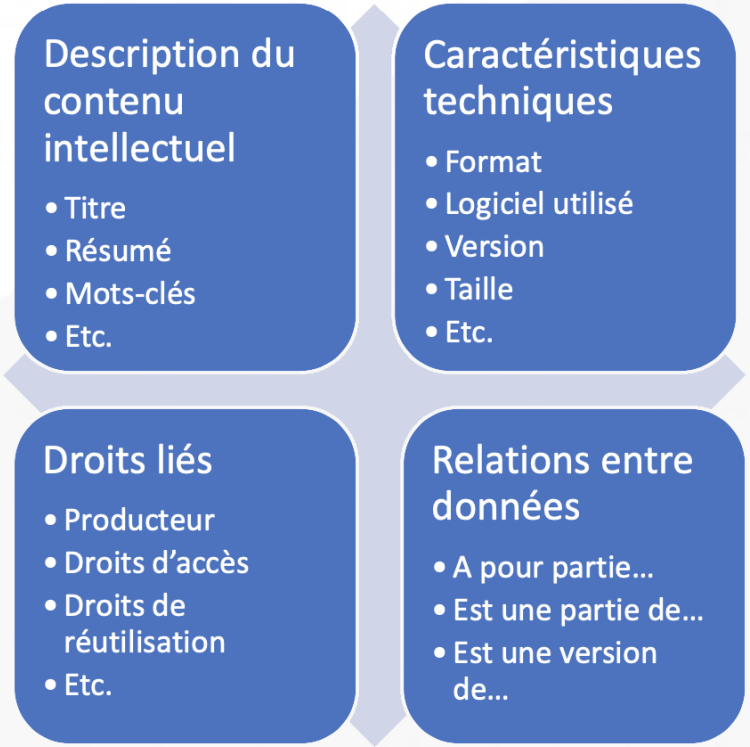
ORCID est l'identifiant le plus utilisé dans le monde.
On peut le lier à d'autres identifiants auteur, comme l'IDHAL.

DoRANum (2020). Les identifiants pérennes
DOI : 10.13143/t427-f432

Les métadonnées

C’est un ensemble structuré d'informations décrivant une ressource quelconque, numérique ou non.

=> C’est **une donnée sur une donnée**.



Ancelin, J. Urfist de Paris (2021). Introduction aux données de la recherche

Métadonnées – Pourquoi faire ?

- Comprendre l'origine des données et leur contexte de création ou de collecte
- Améliorer le moissonnage par les machines (moteur de recherche)
- Garantir l'interopérabilité
- Connaître les conditions de réutilisation et de partage des données
- Fournir des informations très utiles quand les données ne peuvent pas être partagées (embargo, accès restreint) ou lors du retrait de données (données obsolètes, etc.)

DoRANum (2022), Cours introductif sur les métadonnées

DOI : 10.13143/VWCE-G965

Bonnes pratiques de gestion des données

- Réfléchir à l'**organisation** de ses fichiers de données
- **Décrire** et **documenter** ses données de façon à les identifier sans erreur (nommage, versionnage...)
- Choisir des **formats** de fichier adaptés pour la conservation à long terme
- Opter pour une solution de **stockage** adaptée
- Planifier la **sauvegarde** de ses données
- Anticiper un éventuel **partage des données** (quelle licence, quel entrepôt…)
- Réfléchir aux **aspects éthiques, par exemple à l'accessibilité et la consommation énergétique.et juridiques.**

Stockage, sauvegarde, archivage...Mais où ranger ses données ?

Trois phases de sauvegardes

Données chaudes **Données tièdes** **Données froides**

DoRANum (2021). Stockage, partage et archivage : quelles différences ? 2021 DOI : 10.13143/5dax-qp58

Qu’est-ce qu’un plan de gestion de données ?

C’est LA réponse à toutes ces questions !

Le plan de gestion de données

Le plan de gestion de données (PGD) est un document formel et évolutif précisant la manière dont les données seront produites, traitées, décrites, partagées ou protégées et conservées au cours et à l'issue du projet.

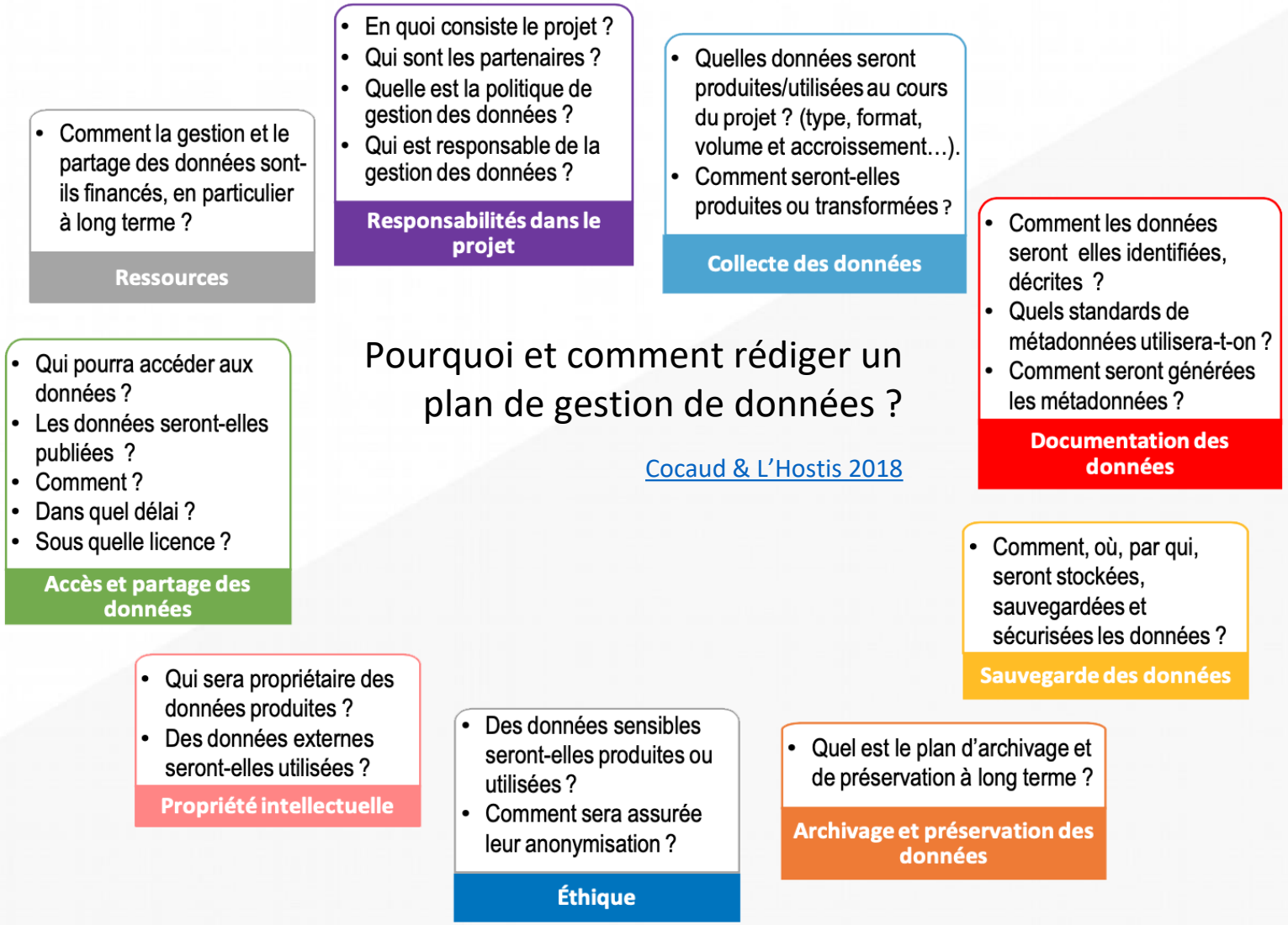
Objectifs

- Décrire le cycle de vie des données produites ou collectées au cours du projet de recherche.
- Anticiper les questions de gestion qui surviennent au cours d'une recherche et les conditions d'une diffusion et d'une conservation futures des données.

Réaliser un plan de gestion de données « FAIR » : modèle

Reymonet *et al.* 2018

Les grandes rubriques d'un PGD



Voir aussi : DMP OPIDoR et DoRANum, La minute plan de gestion de données

Ressources

- CIRAD (2016), Le cycle de vie des données. Intégrer la gestion de données scientifiques aux activités de recherche (poster)
- Deboin M.-C. (2020). S'initier en ligne aux données de la recherche et à leur gestion (guide du CIRAD)
- DoRANum, modules de formation, quiz et tutoriels sur la gestion et le partage des données de recherche, notamment Le parcours interactif sur la gestion des données de la recherche
- Huma-Num (2021), Gérer et diffuser ses données avec les outils et services proposés par Huma-Num (supports de formation et vidéos)
- Lionel Maurel (2021), Quel cadre pour l'ouverture des données de la recherche dans le contexte de la science ouverte ? (séminaire Datasuds 2021)
- Ouvrir la science (2024), Passeport pour la science ouverte. Guide pratiques à l'usage des doctorantes et doctorants (guide)
- Ouvrir la science (2024), Données de la recherche (livret)
- Urfist de Paris (2021), Introduction aux données de la recherche (support de formation)
- DMP OPIDoR, plateforme d'accompagnement à l'élaboration et la mise en pratique de plans de gestion de données et de logiciels.

MERCI !