

Informations techniques

Ce webinaire est enregistré et a vocation à être diffusé, à des fins de formation.

A ce titre et si besoin, vous pouvez couper votre caméra et renommer votre identifiant de connexion à Zoom.

Les micros doivent être coupés pendant la première partie du webinaire.

Un temps de question/réponse (non enregistré) est prévu dans la seconde partie du webinaire.

Vous pouvez poser vos questions dans le chat à tout moment.



**UNIVERSITÉ DE
MONTPELLIER**

Ouvrir les données de la recherche grâce aux principes FAIR

Webinaire du 15 février 2022

Mission "Appui à la recherche" du Service Commun de Documentation de l'Université de Montpellier



— Objectifs et plan schématique de la formation

- Comprendre le contexte et les enjeux liés à l'ouverture des données de la recherche
- Savoir appliquer les principes FAIR

Axes du webinaire :

- I. La Science Ouverte : définition et contexte (rappel)
- II. L'Open Data
- III. Données de la recherche : définition et cycle de vie
- IV. Les principes FAIR

— I. La science ouverte : définition et contexte

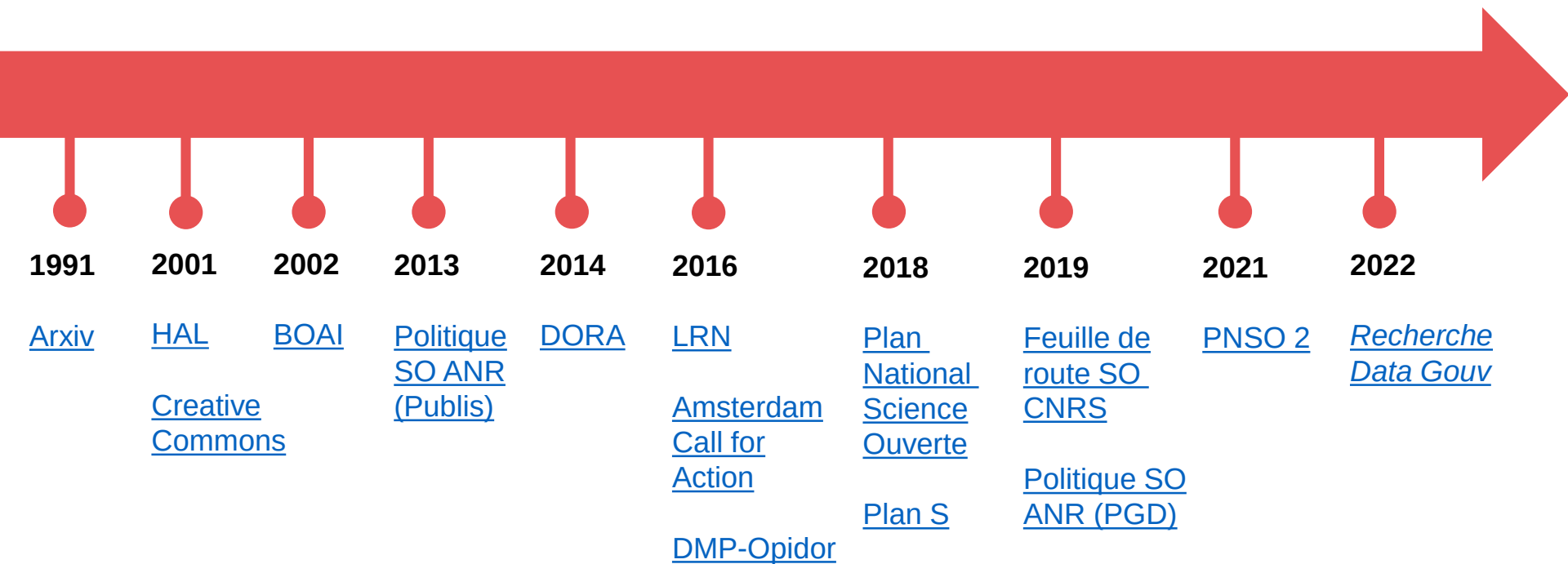
La science ouverte est la diffusion sans entrave des publications et des données de la recherche. Elle s'appuie sur l'opportunité que représente la mutation numérique.

https://www.ouvrirlascience.fr/category/science_ouverte/

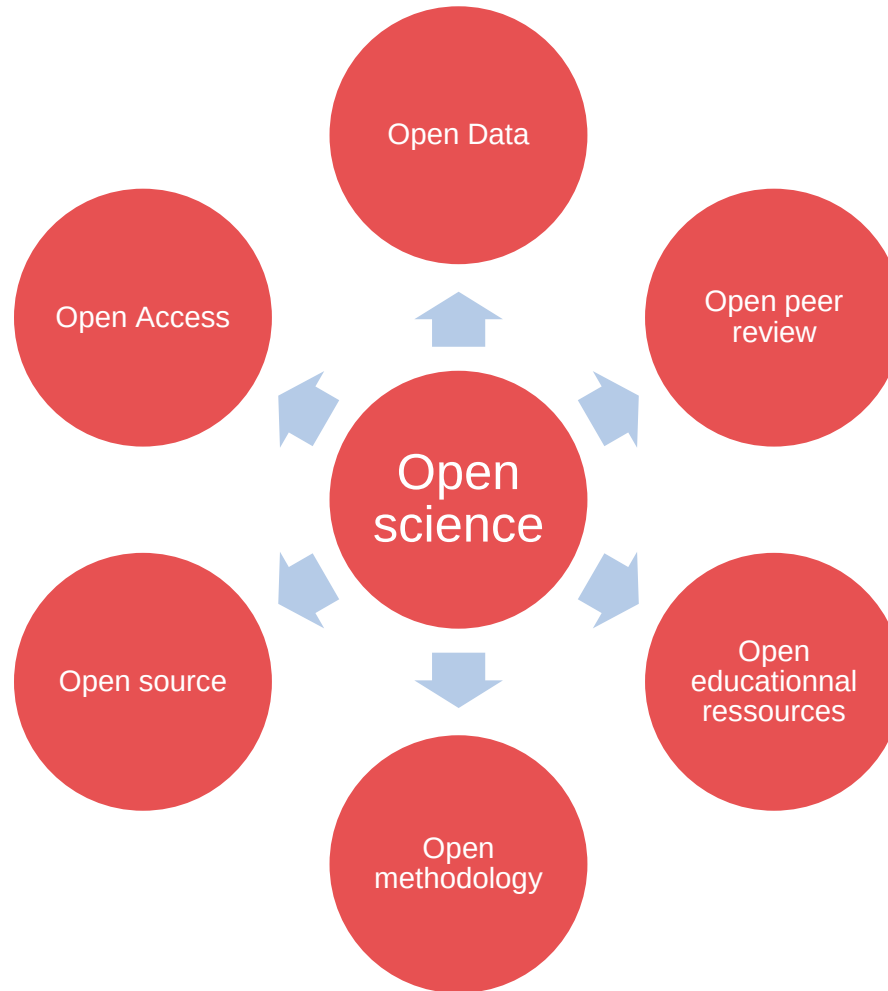
La science ouverte est la pratique de la science qui rend possible la collaboration et la contribution avec d'autres, où les données de recherche, les notes de laboratoire et autres processus de recherche sont librement disponibles, dans des conditions permettant la réutilisation, la redistribution et la reproduction de la recherche, ainsi que de ses données et méthodes sous-jacentes.

<https://www.fosteropenscience.eu/foster-taxonomy/open-science-definition>

I. La science ouverte : développement



I. La science ouverte : périmètre



I. La science ouverte : enjeux

Enjeux académiques

- Faciliter l'accès à l'information scientifique et technique
- Faciliter les échanges entre pairs, la circulation des idées
- Garantir l'objectivité des connaissances

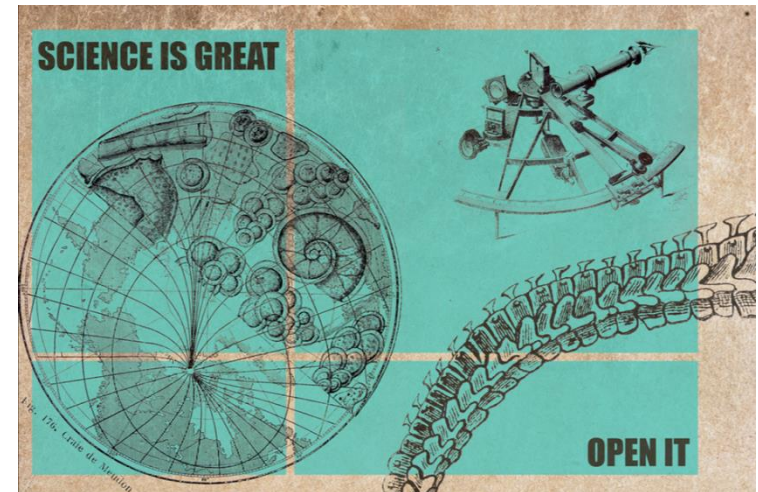
Enjeux politiques

Restaurer la confiance (mouvements anti-science)

- Science au cœur de la démocratie
 - besoin d'information citoyen
 - sciences participatives, forums hybrides
 - pilier des démocraties modernes

Enjeux économiques

- Efficience des dépenses publiques
- Economie de la connaissance
 - partage générateur de croissance, de réduction des coûts



By Martin Clavey – CC-BY-SA

— II. L'Open Data : définition

L'Open Data est une démarche de communication des documents ou données publics, afin qu'ils soient diffusés de manière structurée selon une méthode garantissant leur libre accès et leur réutilisation par tous, sans restriction technique, juridique ou financière injustifiée.

Source : [Ouvrir la science](#) « Ouverture des données de recherche : Guide d'analyse du cadre juridique en France », 2021, MESRI

— II. L'Open Data : définition

La directive européenne « PSI » et la loi française « CADA 2 » définissent un **droit d'accès individuel aux documents administratifs et un droit de réutilisation** pour tous des informations qui y sont contenues, **sous réserve d'exception**.

Les **documents administratifs** sont tous les documents produits ou reçus, dans le cadre de leur mission de service public, par l'État, les collectivités territoriales ainsi que par les autres personnes de droit public ou les personnes de droit privé chargées d'une telle mission.

Ce qui englobe : **données brutes, données élaborées et métadonnées**.

— II. L'Open Data et données de la recherche

Les droits et principes définis en matière d'Open Data s'appliquent pleinement aux données de la recherche.

Les règles de réutilisation des données sont aujourd'hui les mêmes pour tout document administratif.

Les deux conditions préliminaires pour diffuser les données sont :

- **des données réalisées dans le cadre de la mission de service public de l'établissement**
- **des données achevées**

MAIS d'autres conditions doivent aussi être réunies, c'est pourquoi il faut également **vérifier la nature des données.**

—II. L'Open Data : définition

Les données que j'exploite sont-elles concernées par l'Open Data ?

Cela dépend de leur **nature**, la communication peut être :

1. **Obligatoire** (données géographiques ou relatives à des émissions de substances dans l'environnement)
2. **Interdite** (données présentant des risques pour la protection du secret de la défense nationale, la sécurité de l'État, la sécurité publique, la sécurité de l'établissement ou soumises au secret professionnel)
3. **ou soumise à certaines conditions** (données présentant des risques pour la protection du potentiel scientifique et technique de la nation, liées à un contrat avec un tiers non soumis à une obligation de service public, relatives aux personnes et à la vie privée, protégées par le droit d'auteur et les autres droits de propriété intellectuelle, statistiques ou encore le cas des zones à régime restrictif)

—II. L'Open Data : définition

Les données que j'exploite sont-elles concernées par l'Open Data ?

Pour résumer, il faut impérativement vérifier que les données dont la diffusion est envisagée ne contiennent pas d'éléments pouvant :

- relever du **secret de la défense nationale**
- présenter des risques pour la **sécurité publique ou celle de l'établissement**
- relever du **secret professionnel** (médical, instruction, etc.) ou être couvert par la **confidentialité** (affaires, etc.)

Avez-vous le droit ou l'obligation de diffuser vos données ? Consultez l'arbre « [Aide à la décision sur la diffusion des données de la recherche](#) ».

— II. L'Open Data : conclusion

Les données de la recherche financées sur fonds publics sont des données publiques dont le principe est l'ouverture.

Néanmoins, des **restrictions juridiques et éthiques** s'imposent notamment pour les données personnelles, les données sensibles, les données protégées par le droit d'auteur.

En revanche, la **communication** des données relatives aux secrets professionnels, au secret de la défense nationale, à la sûreté de l'Etat et à la sécurité publique **est interdite par principe**.

Ainsi, il est nécessaire de **rester vigilant sur les données que l'on souhaite diffuser**.

— III. Données de la recherche : définition

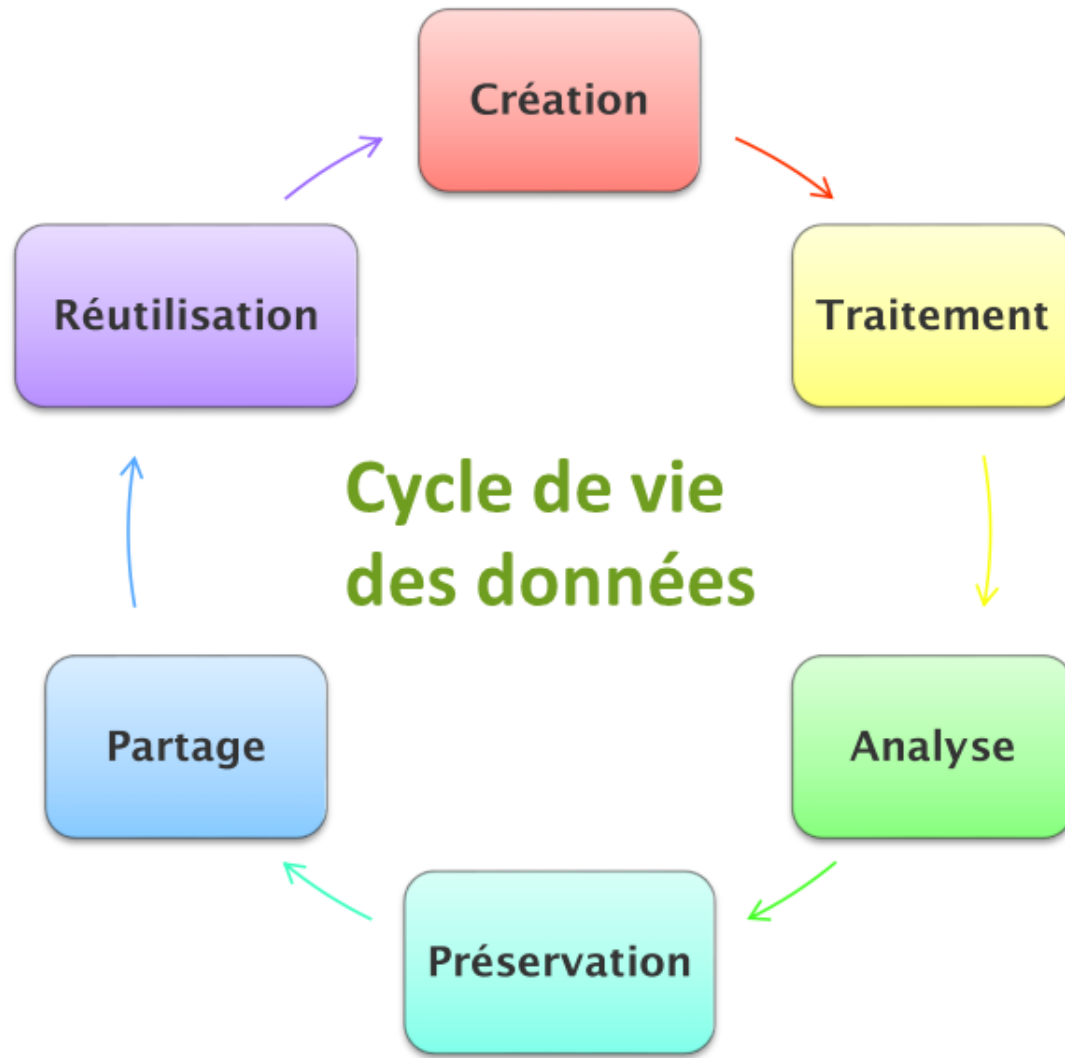
« Les données de la recherche sont définies comme des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider les résultats de la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

Définition de l'OCDE, rapport 2017 - *Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics*

Elles peuvent avoir différents :

- **supports** : carnets de laboratoire, documents électroniques, logiciels, programmes informatiques...
- **types** : archives, audio, vidéo, bases de données, codes sources, données géospatiales, images, photographies, numérisations, scans, qualitatives, quantitatives, statistiques...
- **natures** : brutes, expérimentales, produites ou collectées, issues d'observation ou de simulation...

— III. Données de la recherche : cycle de vie

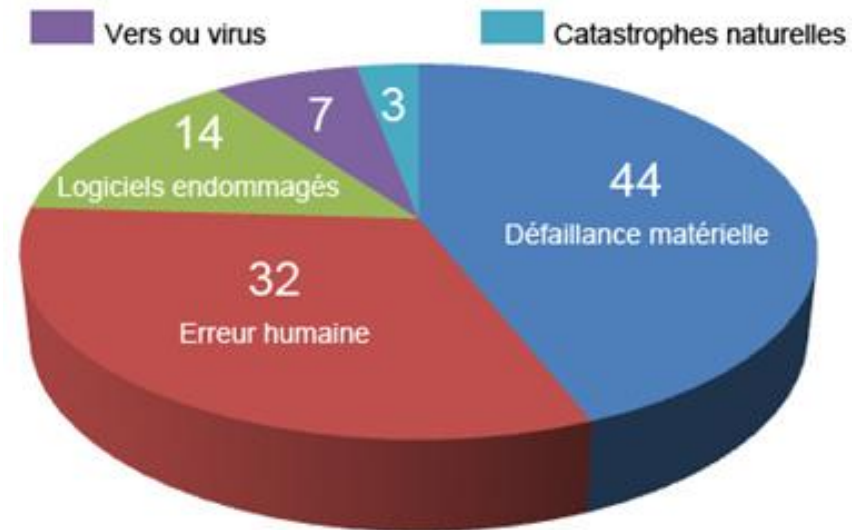


Source : <https://data.ird.fr/utiliser-des-donnees/>

— III. Données de la recherche : pourquoi les gérer ?

20 ans après publication, 80 % des données scientifiques sont perdues...

- Pour les auteurs
- Pour les équipes, organismes, institutions
- Pour la recherche mondiale
- **Causes :**
 - Destruction des supports
 - Problème de version, obsolescence matérielle ou logicielle
 - Lieu de stockage indéfini
- **Conséquences :**
 - Perte de temps, d'argent (fonds publics...)
 - Pas de vérification des résultats possible
 - Pas de comparaison des résultats dans le temps ou l'espace
 - Pas de réutilisations (différentes)

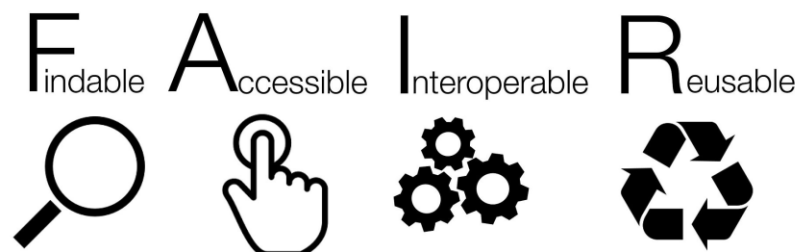


III. Données de la recherche : enjeux du partage et de l'ouverture des données

- Enjeux **patrimoniaux**
- Enjeux **économiques**
 - Valeur économique de la donnée
 - Réutilisation des données et exploitation des résultats de recherches
 - Innovation et retour sur Investissement dans la R&D
- Enjeux **scientifiques**
 - Pour la communauté : science ouverte, « data driven science », interdisciplinarité, encourager la collaboration, éviter la duplication des efforts, augmenter les possibilités de financements
 - Pour le scientifique : gain en termes de visibilité et de citations
- Enjeux **sociétaux**
 - Participation des citoyens et de la société civile : "Citizen Science"
 - Plus de transparence du processus scientifique, plus de confiance en la recherche

Source : <http://www6.inra.fr/datapartage/Partager-Publier/Les-enjeux-pour-l-Inra>

— IV. Les principes FAIR : définition



- **Faciles à trouver** : dotées d'identifiants uniques et pérennes (ex. DOI), décrites par des métadonnées précises, déposées dans un entrepôt de données
- **Accessibles** : récupérables par des protocoles de communication standardisés, libres et ouverts, avec authentification en cas de données sensibles
- **Interopérables** : décrites avec un vocabulaire contrôlé permettant l'interopérabilité et les liens entre les données
- **Réutilisables** : publiées avec une licence d'utilisation claire et accessible, avec indication de leur provenance

—IV. Les principes FAIR : Findable

1. Faciles à trouver (findable)

L'objectif est de faciliter la découverte des données.

- Les données ont un **PID**
- Les données sont décrites avec des **métadonnées**
- Ces **métadonnées indiquent les PID** des données
- Les données sont déposées dans un **entrepôt** de données

—IV. Les principes FAIR : Findable

1. Les données ont un PID

Un **PID** (Persistent IDentifier) est un mécanisme permettant d'identifier de façon stable et unique des ressources sur le Web. Le plus connu est le **DOI** (Digital Object Identifier).

Un PID permet de créer un **lien hypertexte unique et pérenne** à un jeu de données spécifique.

—IV. Les principes FAIR : Findable

2. Les données sont décrites avec des métadonnées

Une **métadonnée** est un élément qui sert à décrire une donnée.

Pour les rédiger, il est recommandé d'utiliser des **standards** spécifiques (Dublin Core, DDI...) et des **formats** ouverts (XML...).

3. Les métadonnées indiquent les PID des données

Pour être FAIR, les métadonnées doivent contenir le **PID** du jeu de données décrit.

Comme les métadonnées se retrouvent souvent dans un fichier séparé du jeu de données, il vaut mieux mentionner dans le fichier des métadonnées le PID du jeu de données en question.

— IV. Les principes FAIR : Findable

4. Les données sont déposées dans un entrepôt de données

Les entrepôts de données sont les endroits où l'on peut déposer des données, en rechercher d'autres et y accéder, en vue d'une réutilisation.

Ils offrent une indexation fine et contrôlée, utile dans la recherche scientifique.

2. Accessibles (Accessible)

Permettre l'accès aux données et leur téléchargement.

- Les données sont accessibles à travers un **protocole de communication standard**
- Ce protocole est **libre et ouvert**
- Ce protocole permet un accès par **authentification** si besoin
- Les **métadonnées restent accessibles** même si les données ne le sont pas

—IV. Les principes FAIR : Accessible

1. Les données sont accessibles à travers un protocole de communication standard

Un **protocole de communication** est un ensemble de procédures que suivent les machines pour communiquer entre elles.

Il en existe plusieurs qui sont destinés à des types de communications particuliers comme le [HTTP](#) et le [FTP](#).

L'entrepôt sur lequel vos données sont déposées devrait utiliser des protocoles standards tels que le HTTP et le FTP.

2. Les protocoles sont libres et ouverts

Les protocoles de communication libres et ouverts sont librement **utilisables et interopérables**.

Il est conseillé d'utiliser les protocoles libres et ouverts (HTTP, FTP...) et d'éviter les protocoles propriétaires fermés (Skype, Microsoft exchange server...).

—IV. Les principes FAIR : Accessible

3. Les protocoles permettent un accès par authentification (si besoin)

Le principe Accessible n'est pas synonyme de libre et ouvert : certaines données ne peuvent pas être accessibles pour des raisons légitimes.

Les protocoles de communication utilisés doivent alors pouvoir **restreindre l'accès aux données** par authentification et/ou autorisation (HTTPS ou FTPS).

— IV. Les principes FAIR : Accessible

4. Les métadonnées restent accessibles même si les données ne le sont pas

Si les données disparaissent ou sont inaccessibles, les métadonnées continuent de fournir de précieuses informations pour que d'autres chercheurs puissent connaître l'existence des données, contacter les personnes ressources ou encore retrouver les articles associés aux données.

— IV. Les principes FAIR : Interoperable

3. Interopérables (Interoperable)

Permettre l'exploitation des données quel que soit l'environnement informatique utilisé.

- Les données sont **décrites avec un vocabulaire contrôlé**
- Le vocabulaire utilisé **respecte les principes FAIR**
- Les **métadonnées sont reliées à d'autres données**

—IV. Les principes FAIR : Interoperable

1. Les données sont décrites avec un vocabulaire contrôlé

Afin que les données restent automatiquement accessibles et compréhensibles, leur description doit suivre un **vocabulaire contrôlé** et un **modèle de représentation** servant à le structurer.

Cela fournit des ressources compréhensibles par les humains et par les machines. L'**échange** de données peut ainsi se faire automatiquement d'un système à l'autre.

— IV. Les principes FAIR : Interoperable

2. Le vocabulaire utilisé respecte les principes FAIR

Il s'agit de vocabulaires contrôlés dont on peut retrouver la documentation grâce à un PID.

Il doit être **documenté** (décrit par des métadonnées) et lisible par les machines.

Par exemple, le vocabulaire ANDO respecte les principes FAIR.

— IV. Les principes FAIR : Interoperable

3. Les métadonnées sont reliées à d'autres données

Les principes FAIR s'appuient sur les technologies liées au Web de données. En ce sens, il est fortement recommandé de s'en servir afin de constituer un **réseau global d'informations scientifiques**.

- En créant des **liens significatifs entre les données**, vous mettez en avant d'autres données en lien avec la recherche initiale. La recherche des données devient alors plus efficace et permet de découvrir de nouvelles données pertinentes.
- L'objectif est de créer un réseau de données liées entre elles, afin d'enrichir la connaissance contextuelle des données. Pour cela, il est recommandé de **déposer ses données dans des entrepôts**.

4. Réutilisables (Reusable)

Permettre la réutilisation des données pour de futures recherches :

- Les métadonnées ont une **pluralité d'attributs**
- Une **licence de réutilisation** est attribuée aux données
- La description des données indique leur **provenance**
- Le partage des données suit les **standards de la communauté scientifique**

—IV. Les principes FAIR : Reusable

1. Les métadonnées ont une pluralité d'attributs

Plus on connaît le contexte dans lequel des données ont été créées, plus on peut en tirer profit.

Ce principe incite à offrir **beaucoup d'informations contextuelles sur les données**, même celles qui peuvent paraître inutiles, car on ne sait pas quels seront les besoins des futurs ré-utilisateurs.

— IV. Les principes FAIR : Reusable

2. Une licence de réutilisation est attribuée aux données

Attribuer une licence permet de **définir explicitement les droits des ré-utilisateurs concernant vos données.**

Le choix de la licence va donc dépendre des droits que vous pouvez accorder en fonction de la nature juridique des données.

Elles peuvent par exemple être déjà protégées par le droit d'auteur ou faire l'objet d'exceptions à leur libre accès.

Il existe des **licences gratuites adaptées** pour l'Open Data : [Creative Commons](#), [Open Licence](#) et [Open Database Licence](#).



LICENCE OUVERTE
OPEN LICENCE



— IV. Les principes FAIR : Reusable

3. La description des données indique leur provenance

Parmi les informations à fournir avec les données, celles concernant leur provenance est essentielle pour prouver leur **crédibilité** et leur **fiabilité**.

Elles doivent aider à **déterminer la qualité des données**, permettre de reproduire l'expérience et de réutiliser les données.

Ces informations peuvent **être demandées lors du dépôt dans un entrepôt** (comme les auteurs, leur institution, la date de création des données, leur relation avec d'autres données déjà publiées, etc).

— IV. Les principes FAIR : Reusable

4. Le partage des données suit les standards de la communauté scientifique

Il est plus aisé de réutiliser des données qui sont organisées de la même façon, ont les mêmes formats, sont décrites avec le même schéma... en somme, des données qui suivent un **modèle commun pour leur partage**.

S'il en existe un, il faut **suivre le modèle de sa communauté scientifique** afin de faciliter la réutilisation des données.

S'il n'existe pas de standard disciplinaire, il faut trouver les moyens les plus adaptés pour partager ses données (format ouvert...).

— IV. Les principes FAIR : FAIR-Aware

FAIR-Aware est un outil en ligne qui aide les chercheurs à évaluer leur connaissance des pratiques FAIR pour rendre leurs jeux de données FAIR, avant de les déposer dans un entrepôt de données.



Your first step towards your FAIR data(set)

Concrètement, il comporte **10 questions auxquelles il faut répondre en ayant en tête un jeu de données précis.**

Webinaires à venir

- **Déposer sa production scientifique dans HAL**

8 mars 2022 de 13h à 14h

- **Pourquoi rédiger un plan de gestion de données ?**

15 mars 2022 de 13h à 14h

- **Créer son IdHAL et son CV dans HAL en lien avec ORCID**

05 avril 2022

Pour se tenir informé des webinaires thématiques bimensuels, c'est [ici](#) !

Merci pour votre écoute !

Gardons contact pour la suite :

- Pour des questions autour des publications ou de l'utilisation de HAL : hal-assistance@umontpellier.fr
- Pour des questions autour de la gestion des données de la recherche ou pour toute relecture de PGD : donnees-recherche@umontpellier.fr
- Pour toute autre question : laure.lefrancois@umontpellier.fr

Chargée de mission Appui à la recherche du SCD

Webographie

- [Deuxième Plan national pour la science ouverte](#) (Ouvrir la science - MESRI)
- [La feuille de route 2021-2024 du MESRI sur la politique des données, des algorithmes et des codes sources](#)
- [Guide sur l'ouverture des données de la recherche](#) (Ouvrir la science – MESRI)
- [Logigramme de communicabilité des données](#) (Dorandum)
- [Parcours interactif sur la gestion des données de la recherche](#) (Dorandum)
- [Conseils pour l'organisation des données](#) (Dorandum)
- [Comprendre les données Fair](#) (Université de Paris Saclay)
- [Enjeux et bénéfices des données FAIR](#) (Dorandum)
- [L'Open Data](#) (Ouvrir la science – MESRI)
- [Les métadonnées](#) (Dorandum)
- [Les entrepôts de données](#) (Dorandum)
- [FAIR-Aware](#)
- [Les licences](#)