

L'origine et la description des données

Dans votre DMP, vous devez préciser l'origine des données. S'agit-il de collecter des données ? D'en créer de nouvelles ? Ou de réutiliser des données existantes ?

Dans tous les cas, vous devez détailler la méthodologie de travail utilisée, et décrire le plus précisément possible les données.

Origine des données

Vous devez spécifier pour chaque jeu de données :

- si vous allez collecter et utiliser des jeux de données existants ;
- ou bien créer de nouveaux jeux de données ;
- ou bien à la fois collecter des jeux de données existants et en créer de nouveaux.

Dans chacun des cas et pour chaque jeu de données, vous devez également détailler la méthodologie adoptée.

Concrètement, pour des données existantes, vous devez décrire le chemin d'accès à ces données, c'est-à-dire citer la source, les modalités d'accès et donner une description synthétique.

Pour des données nouvelles, créées ou générées, vous devez détailler le processus de création ou le mode opératoire mis en place.

Description des données

Chaque DMP doit comporter une description des données de recherche du projet. Il convient de décrire le plus précisément possible ces données pour qu'elles soient comprises et exploitées :

- par tous les acteurs liés au projet (partenaires, financeurs...)
- mais aussi par d'autres chercheurs dans le cadre du partage des données de recherche pour de nouveaux projets.

Il n'existe pas de méthodologie unique. En effet, de nombreux domaines scientifiques disposent de leur propre terminologie pour décrire leurs données de recherche.



Voici un guide interactif qui présente plusieurs pistes de classifications descriptives suffisamment généralistes pour être applicables à tout projet :

- en fonction de la source des données,
- en fonction de la forme des données,
- et en fonction de leur stabilité.

De quelle source proviennent les données ?

En fonction de leur provenance, les données peuvent être :

➤ Des données d'observation :

Les données d'observation sont capturées en temps réel, la plupart du temps en dehors du laboratoire. Elles sont habituellement uniques et donc impossibles à reproduire. C'est le cas de la neuroimagerie, de la photographie astronomique ou des données d'enquête.

➤ Des données expérimentales :

Les données expérimentales sont obtenues à partir d'équipements de laboratoire. Elles sont souvent reproductibles mais cela peut s'avérer coûteux. Les chromatogrammes, les puces à ADN et les séquences de gènes relèvent de cette deuxième catégorie.

➤ Des données computationnelles ou de simulation :

Les données computationnelles ou de simulation sont générées par des modèles informatiques ou de simulation. Elles sont souvent reproductibles à condition que le modèle soit correctement documenté. On peut citer par exemple les modèles météorologiques, de simulation sismique ou encore les modèles économiques.

➤ Des données dérivées ou compilées :



Les données dérivées ou compilées sont issues du traitement ou de la combinaison de données brutes. Elles sont souvent reproductibles mais coûteuses. C'est le cas pour la fouille de texte, les bases de données compilées ou les modèles 3D.

Sous quelle forme se présentent ces données ?

Il peut s'agir de :

- Données textuelles (notes de terrain ou de laboratoire, réponses d'enquête...),
- Données numériques (tableaux, comptes, mesures...),
- Données audiovisuelles (images, sons, vidéos...),
- Modèles, codes informatiques,
- Données spécifiques liées à une discipline (par exemple FITS – *Flexible Image Transport System* - en astronomie ou CIF – *Crystallographic Information File* - en cristallographie),
- Données spécifiques produites par certains instruments (électrocardiogramme, sismogramme...).

Quelle est la stabilité de ces données ?

On peut distinguer 3 niveaux :

- Les jeux de données fixes : les données ne changent pas après avoir été recueillies ou produites.
- Les jeux de données croissants : de nouvelles données peuvent être ajoutées, mais les anciennes données ne sont jamais modifiées ou supprimées.



- Les jeux de données révisables : de nouvelles données peuvent être ajoutées et les anciennes données peuvent être modifiées ou supprimées.

Pour cette partie du DMP, quelle que soit la méthodologie adoptée, le plus important est de décrire les données de recherche du projet de façon précise et détaillée.

L'objectif est d'optimiser la gestion des données de recherche et leur partage dans le cadre de l'Open access.



dorandum.fr

