Comment décrire ses données pour les rendre compréhensibles et réutilisables ?

openscience.pasteur.fr/2022/02/21/comment-decrire-ses-donnees-pour-les-rendre-comprehensibles-et-reutilisables/

CeRIS - Institut Pasteur 21 février 2022

Vous êtes-vous déjà retrouvé dans la situation d'être incapable de **décrire précisément** un jeu de données, par exemple au moment où vous souhaitez le partager ou le publier sur un entrepôt ?

Il est tout à fait normal de ne pas se souvenir des détails sur des données, plusieurs mois ou années après les avoir générées ou collectées. C'est pourquoi, **décrire ses données au moment où elles sont créées** – le moment où on les connait le mieux – fait partie des bonnes pratiques pour **rendre ses données compréhensibles et réutilisables**. Mais comment faire concrètement ?

Des données scientifiques peuvent être décrites de deux façons :

> Par de la **documentation**, un texte qui décrit les données, les contextualise et donne toute information nécessaire à leur compréhension. La documentation n'est pas structurée et est lisible uniquement par les humains.

Il peut s'agir par exemple d'un fichier README associé aux données ou d'une description des résultats dans un cahier de laboratoire électronique (pensez à bien faire le lien entre cette description et le lieu de stockage des données).

> Par des **métadonnées**, des informations structurées et lisibles par machine ("machine-readable") qui décrivent les données. Elles peuvent être généralistes (ex : titre, auteur, format, date de création…) ou plus scientifiques (ex : organisme étudié, pathologie associée, technique de mesure…).

Ces métadonnées structurées prennent généralement la forme d'un fichier CSV, JSON, XML ou RDF associé aux données. Voir un exemple de description au format XML (source : DDI).

Quelques outils qui pourraient vous être utiles pour générer des métadonnées structurées :

- ISA tools : une suite logicielle open source vous permettant de décrire précisément vos **données omiques** en suivant le standard ISA (Investigation/Study/Assay).
- MethodsJ2 : un logiciel open source basé sur ImageJ/Fiji qui capture automatiquement les métadonnées des **images de microscopie** à partir de plusieurs sources et guide l'utilisateur pour la saisie des métadonnées expérimentales spécifiques.

- DDI (Data Documentation Initiative) propose une liste d'outils pour documenter des **données issues d'enquêtes** ou d'autres méthodes d'observation dans le domaine des sciences sociales, comportementales, économiques et de la santé.

Que doit-on inclure dans la description des données ?

Pour que les données soient réutilisables et reproductibles (par votre "futur vous-même" ou un autre utilisateur), la description devrait se faire à deux niveaux :

- **une description de l'étude ou du projet** : titre, résumé du projet, auteurs, objectifs de l'étude, institutions impliquées, financement, méthodes, lien vers les autres jeux de données… Le plan de gestion des données peut notamment être utilisé comme documentation.
- **une description des données elles-mêmes**, incluant notamment la signification des termes utilisés, les types de variables, les facteurs expérimentaux, etc.

En biologie, les standards MIBBI (Minimum Information for Biological and Biomedical Investigations) peuvent être utilisés comme guides pour décrire des données dans différentes disciplines.

Pour aller plus loin : RDMkit – Documentation and metadata