Comment décrire ses données pour les rendre compréhensibles et réutilisables ?

openscience.pasteur.fr/2024/06/17/comment-decrire-ses-donnees-pour-les-rendre-comprehensibles-et-reutilisables/

CeRIS - Institut Pasteur 17 juin 2024

[Mise à jour d'un article initialement publié en février 2022]

Mettre des données de recherche à disposition de la communauté scientifique ou les préserver sur le long terme : des bonnes pratiques qui ne servent à rien si les données ne sont pas compréhensibles. **Décrire les données et documenter leur contexte de création** est donc une étape indispensable.

Que doit-on inclure dans la description des données ?

Pour que les données soient compréhensibles et réutilisables par votre « futur vous-même », vos collègues ou d'autres scientifiques, la description devrait se faire à deux niveaux :

- une description de l'étude ou du projet : titre, description du projet et des objectifs, contributeurs, institutions impliquées, financement, méthodes employées, lien vers les autres jeux de données…
- une description des données elles-mêmes, incluant notamment la signification des termes utilisés, les types de variables, les facteurs expérimentaux, etc.

Quelle forme peut prendre la description des données ?

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**. Elle permet de comprendre ce qui a été fait et pourquoi.

Plusieurs documents peuvent être associés aux données :

- extrait de cahier de laboratoire électronique décrivant la méthode employée et les résultats,
- dictionnaire de données expliquant la signification des noms de variables et des valeurs d'un tableur ou d'une base de données,
- plan de gestion des données,
- fichier README décrivant notamment l'organisation hiérarchique des dossiers…

L'entrepôt Recherche Data Gouv propose par exemple un modèle de README que vous pouvez télécharger et adapter.

– **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 disciplinaires (ex : organisme étudié, pathologie associée, technique de mesure…).

La description des données par des métadonnées se fait généralement au moment du dépôt des données dans un entrepôt. Mais il est possible d'**anticiper** : si vous avez trouvé l'entrepôt adapté à vos besoins, vous pouvez chercher quel standard de métadonnées il utilise. Vous pourrez ainsi utiliser ce standard pour décrire vos données le plus tôt possible sous forme d'un fichier CSV, JSON, XML ou RDF que vous conservez à proximité des données.

Quelques exemples de standards de métadonnées utilisés par des entrepôts :

- L'entrepôt FlowRepository suit le standard de métadonnées MIFlowCyt (*Minimum Information about Flow Cytometry*).

- L'entrepôt BioImage Archive suit le standard REMBI (*REcommended Metadata for Biological Images*).
- L'entrepôt ArrayExpress suit le standard MIAME (*Minimum Information About a Microarray Experiment*).

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.
- La fonctionnalité Annotate Experiment du logiciel FlowJo : permet d'annoter son expérience de **cytométrie en flux** suivant le standard MIFlowCyt.
- 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.

Si vos données ne peuvent pas être déposées dans un entrepôt de données, vous pouvez tout de même annoter vos données par des métadonnées structurées. 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