Décrire ses jeux de données dans les règles du FAIR : accompagner les chercheurs à l'utilisation des métadonnées

1. Introduction

Les métadonnées sont encore souvent **peu connues du public scientifique**, en dépit de l'existence de nombreuses ressources sur le sujet, notamment sur la plateforme DoRANum. Elles ont pourtant une importance cruciale pour la **découverte** et le **signalement** de jeux de données de recherche, à l'heure où ces derniers sont un enjeu majeur de la science ouverte comme en témoigne l'axe 2 du 2ème Plan national pour la science ouverte.

De plus en plus de **revues** demandant l'**accès aux données sous-jacentes de l'article** par le biais de leur dépôt dans un **entrepôt**, la question des métadonnées descriptives de ces données se pose, déjà, de manière récurrente. Or d'un entrepôt à un autre, les champs disponibles pour décrire les données peuvent fortement varier selon qu'il soit généraliste ou disciplinaire. Le GTSO Données de Couperin a consacré en 2022 un webinaire à ce sujet : De la pierre au joyau : les métadonnées au service de la qualité des données.

En outre, une bonne utilisation des métadonnées est une étape cruciale pour se conformer aux principes FAIR.

Comment, dès lors, accompagner au mieux les chercheurs pour associer les bonnes métadonnées à leurs jeux de données ?

2. Définitions

Quelques définitions doivent être rappelées en préambule :

Métadonnées : *"ensemble d'informations structurées qui décrit, explicite, localise une ressource informationnelle, dans le but d'en faciliter la recherche, l'usage et la gestion"* (source : DoRANum). Par exemple, le titre d'un article est une métadonnée.

Métadonnées embarquées / enrichies : les métadonnées **embarquées** sont automatiquement générées au moment de la création de la donnée. Par exemple, lorsqu'une photo est prise, la métadonnée "date de prise de vue" est souvent créée. Les métadonnées **enrichies** sont ajoutées a posteriori. Par exemple pour une photo, le titre de l'album dans lequel elle figurera.

Standards / schémas de métadonnées : ils définissent la manière dont les métadonnées sont organisées, la façon de les remplir (comme le format de la date), les métadonnées obligatoires et facultatives… Ils peuvent être généralistes, comme DataCite, ou disciplinaires. Par exemple, le Crystallographic Information Framework est spécifiquement dédié à la description des structures de cristaux. On parle en général de "standard" pour un schéma propre à une discipline et de "schéma" pour les généralistes.

Des schémas de métadonnées existent aussi selon le :

- *Type de ressources* : image (dimension, espace colorimétrique…), vidéo (nombre d'images par seconde, durée…), texte, livre, entretiens etc. ;
- *Type d'entrepôts* : généralistes ou spécialisés, privés ou publics ;
- *Public visé* : ressources pour les pairs, le grand public, le public français ou étranger…

Thésaurus ou vocabulaire contrôlé : liste fermée de mots-clefs permettant de décrire un jeu de données. Il existe de nombreux vocabulaires différents, disciplinaires comme généralistes. Quelques exemples :

- Sciences Sociales : ELSST - European Language Social Science Thesaurus du CESSDA (multilingue)
- Éducation, culture, sciences naturelles, SHS, communication et information : le thésaurus de l'Unesco (multilingue)
- Physique : Physics Subject Headings (anglais)
- Environnement : GEMET - General Multilingual Environmental (multilingue)
- Economie : thésaurus du JEL - Journal of Economic Literature (anglais)

Jeu de données ou dataset : un jeu de données est constitué d'un ensemble de fichiers de données formant une unité intellectuelle, de la documentation explicative sur ces données (type *Readme*) et de métadonnées (descriptives, administratives et structurelles), complétées par un identifiant pérenne de type DOI.

Fichier *Readme* ou Lisez-moi : document *"nécessaire pour décrire plus en détails le contexte de production et/ou les données dans les fichiers. Il vient en complément des métadonnées saisies lors du dépôt du jeu de données, du dictionnaire de données (qu'il peut aussi contenir) et/ou autres supports de documentation accessibles"* (source : Recherche Data Gouv).

3. Exemples de questions de doctorants ou de chercheurs

- Que sont les métadonnées ?
- Comment savoir quelles métadonnées sont importantes pour ma discipline ?
- A quoi cela sert puisque toutes les informations à connaître sont dans mes données ou dans ma publication ?
- Où/comment faut-il les remplir ?
- Quand faut-il créer des métadonnées ?
- Je n'ai pas de métadonnées, je ne vais pas en inventer.
- L'entrepôt que j'utilise ne propose qu'un ensemble limité de métadonnées.
- Est-ce utile de créer des métadonnées lorsque mes données ne peuvent pas être ouvertes, pour quelles données est-il utile de créer des métadonnées ?
- Comment savoir s'il y a un format à respecter (par exemple pour les noms, l'affiliation, etc.) ?
- Quelle est la différence entre un fichier *Readme* et des métadonnées ?

4. Bonnes pratiques, bons réflexes

Titre, auteur, date, éditeur, etc. sont les métadonnées de publications. Les données de la recherche peuvent également être décrites avec des métadonnées (type de données, date de constitution du jeu, nom du manager, version, format, etc.). Toutes les données, même celles qui ne seront pas ouvertes, peuvent avoir des métadonnées associées.

C'est plus facile qu'il n'y paraît ! La création de métadonnées se fait le plus souvent lors du dépôt sur un entrepôt à la fin du projet. En effet, de manière générale, les entrepôts de données de qualité proposent un formulaire de métadonnées standard à compléter. Il est cependant recommandé de l'anticiper dès les phases de planification et de collecte des données afin de s'assurer d'avoir toutes les informations pour bien décrire les données et gagner du temps au moment du dépôt.

Lors de la planification

Rédiger un plan de gestion des données permet de réfléchir à la documentation des données et à anticiper le temps à y consacrer. C'est à cette étape que le choix de l'entrepôt adéquat doit se faire (multidisciplinaire comme Recherche Data Gouv, disciplinaire comme SEANOE par exemple). Attention, les entrepôts proposent des formulaires de saisie de métadonnées différents et donc le choix de l'entrepôt pourra avoir un impact sur la description des données ! S'il en existe, il est préférable de choisir un entrepôt thématique car le schéma de métadonnées sera plus adapté à la discipline.

Il peut donc être utile de se renseigner auprès des centres de référence thématiques de Recherche Data Gouv et des infrastructures de recherche nationales pour des renseignements sur les schémas de métadonnées et les formats préconisés selon les types de données collectées par discipline. Les ateliers de la donnée peuvent aiguiller vers les bons interlocuteurs.

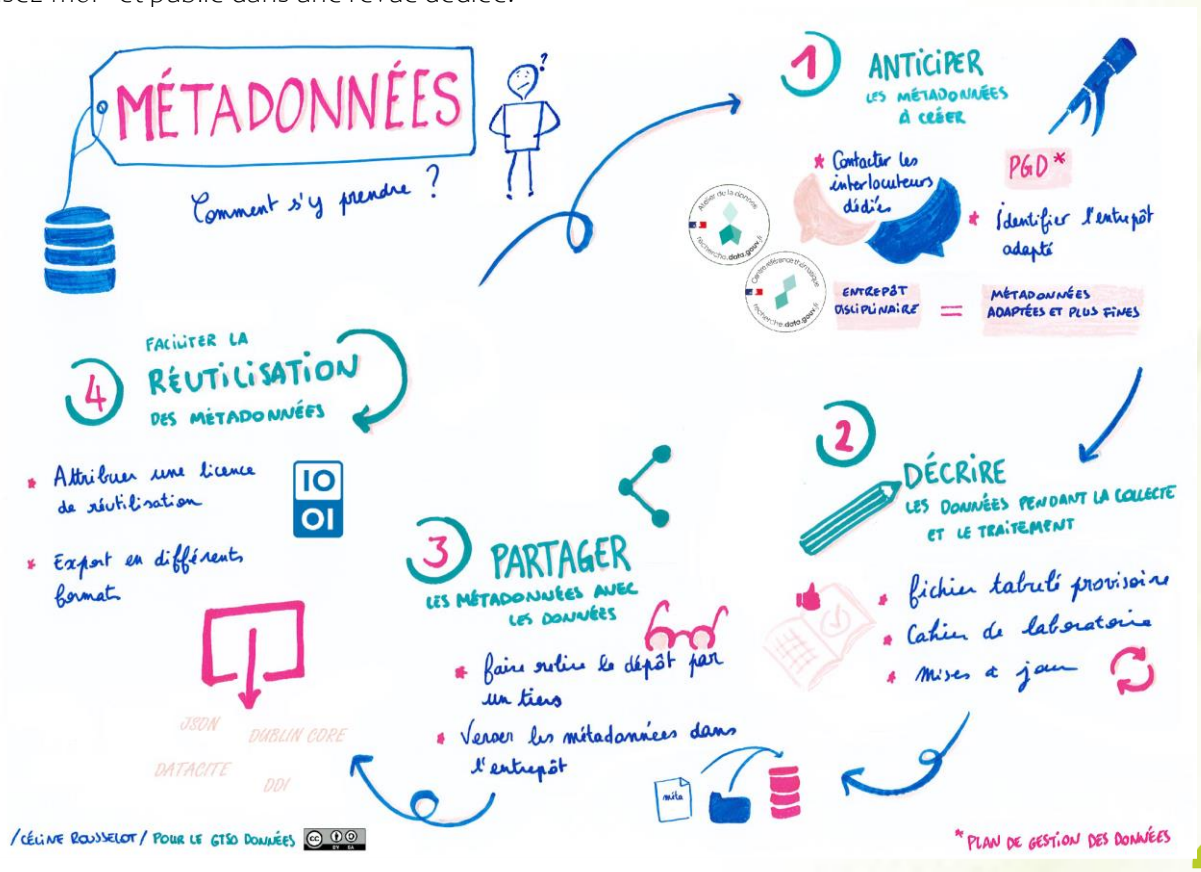
Lors de la collecte et du traitement des données

Renseigner sur un tableau correspondant aux métadonnées de l'entrepôt visé les informations suivantes : les méthodes de collecte ou de production, les traitements effectués (nettoyage, fusion, codage…), la citation précise des sources utilisées, les instruments et logiciels utilisés, les variables et les unités de mesure, les rôles dans la collecte, la gestion et le traitement des données, le contrôle qualité (mesures pour réduire les risques, vérification des erreurs et adéquation des méthodes aux objectifs, calendrier).

Lors du partage / ouverture des données

1. Compléter les métadonnées avec les informations rassemblées à la phase de collecte des données. A minima, le déposant devra compléter des métadonnées de citation ;
2. Se mettre à la place du réutilisateur quand on choisit ses métadonnées. Chercher des jeux de données similaires pour avoir des modèles ;
3. Relire le document "lisez-moi" et le mettre à jour. Si le fichier Lisez-moi n'a pas été créé au moment de la collecte ou du traitement des données, créer le document et l'associer aux fichiers de données sur l'entrepôt ;
4. Faire relire à des personnes extérieures au projet pour s'assurer que le dépôt est compréhensible, complet et réutilisable ;
5. Lier le dépôt de données à la publication associée ;
6. Si l'entrepôt est modéré, transmettre le dépôt au curateur qui pourra être de bons conseils, notamment sur le format des métadonnées (noms, affiliations, etc.).

Si les données ne peuvent être ouvertes, cette documentation créée lors de la collecte et du traitement pourra être conservée ou archivée avec les données. Au moment de la publication des données, un *data paper* peut être rédigé à partir du fichier "lisez-moi" et publié dans une revue dédiée.



5. Quelques outils incontournables

Pour trouver des standards de métadonnées par discipline :

- Disciplinary metadata (Digital Curation Center)
- Metadata standards catalog (Research Data Alliance)

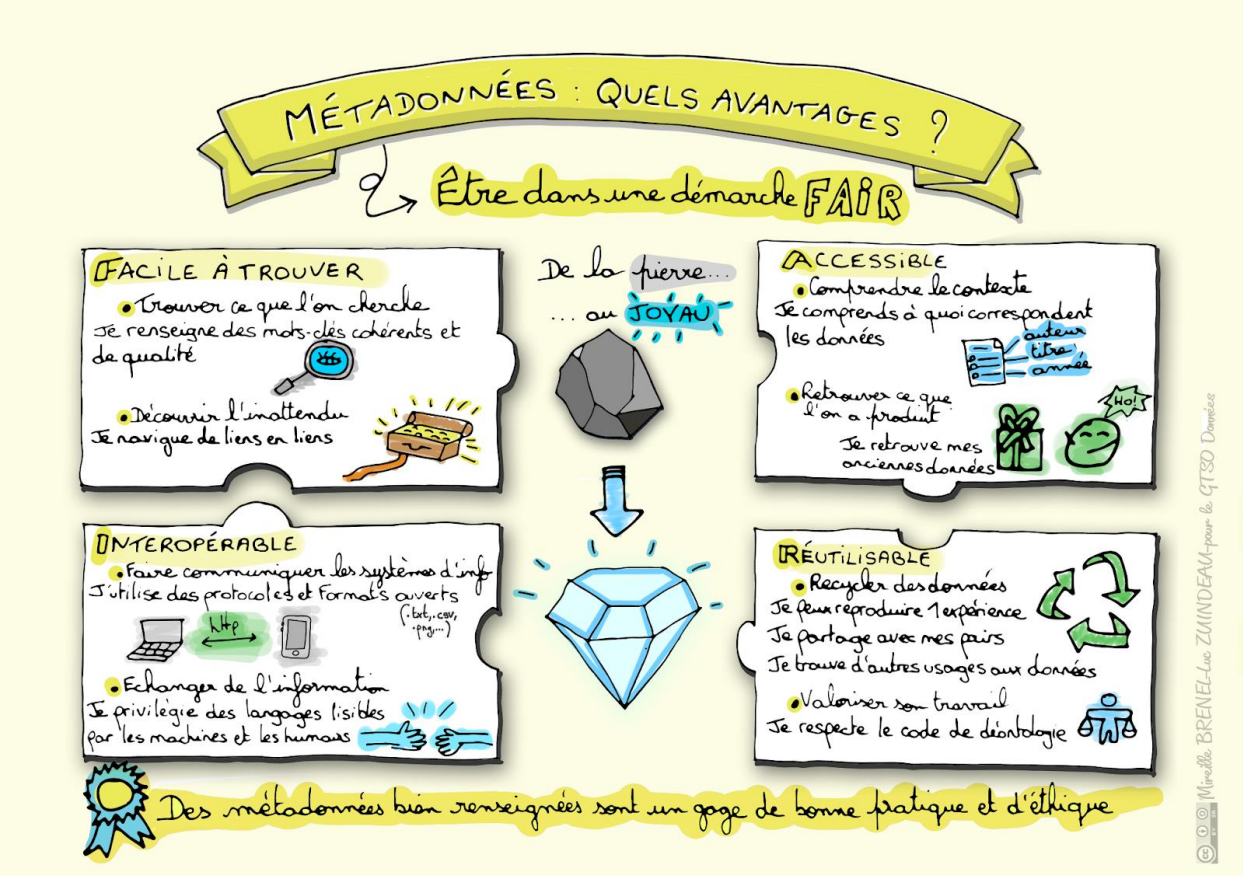
Pour trouver des vocabulaires / thésaurus :

- Loterre, Linked open terminology resources (Inist-CNRS)
- FAIR Sharing, rubrique "thesaurus"

Pour écrire facilement un fichier *Readme* pour accompagner un jeu de données : Modèle proposé par Recherche Data Gouv
Pour générer des métadonnées selon le schéma Datacite : DataCite Metadata Generator
Pour s'assurer de la bonne saisie des métadonnées dans Recherche Data Gouv : Modèle de rapport de curation

6. Avantages

Qu'elles soient descriptives, techniques, administratives ou de provenance, des métadonnées complètes, détaillées et normées sont un atout pour vos données. Travailler sur ses métadonnées peut paraître chronophage et fastidieux. Cependant, pour que les données soient découvertes, accessibles puis réutilisées, elles sont indispensables.



7. Liens utiles

En français

- De la pierre au joyau : les métadonnées au service de la qualité des données - webinaire organisé par le GTSO Données de Couperin
- DoRANum Métadonnées, standards, formats : comment décrire les données ? : on trouve sur cette page des ressources sur le cycle des métadonnées, les standards ou encore les outils de création de métadonnées
- François Ehrenmann, Philippe Chaumeil, Daniel Jacob, Edouard Guitton. Gestion de métadonnées pour les espaces de stockage de données. INRAE. 2022, <hal-03952340>
- Guide de saisie des métadonnées de Recherche Data Gouv
- Guide Etalab : Préparer les données à l'ouverture et la circulation/Documenter les données

En anglais

- Section "metadata and documentation" dans le *Data Management expert Guide* du CESSDA
- *Metadata basics*, Libguide de l'Université du Texas
- LEGO® Metadata for Reproducibility game pack