Projet Open OPSE

Entrepôts de données : mise en pratique

Comment bien préparer un jeu de données ?

Luc DECKER
IRD (service IST – MCST)
data@ird.fr

Projet Open OPSE

Comment bien préparer un jeu de données ?

Suivre les bonnes pratiques

- Principes **FAIR**
  - ✓ Qualité des métadonnées
  - ✓ Standards de métadonnées

ⓘ **Par son fonctionnement et ses règles internes, un entrepôt aide à respecter les principes FAIR**

- Principes **Data Citation**
- Qualité éditoriale des métadonnées
  - ✓ Contenu intelligible, compréhensible
  - ✓ Précision et complétude (lieux, périodes…)
  - ✓ Mise en forme, présentation
  - ✓ Grammaire & orthographe
- Du bon sens : penser aux utilisateurs de données, à l'image de l'institut ou du laboratoire

Projet Open OPSE

Principes FAIR : augmenter le potentiel des données

*"The FAIR Guiding Principles for scientific data management and stewardship". Scientific Data. **3**: 160018. doi:10.1038/sdata.2016.18*

- Identifiants pérennes
- Bien décrire les données
  ⇨ Référencement par les moteurs de recherche

Facile à trouver

- Gérer de façon pérenne l'accès aux données
- Obtenir l'accord des producteurs

Accessible

« Aussi ouvert que possible, aussi fermé que nécessaire »

Interopérable

- Formats ouverts
- Vocabulaires partagés
- Standards de métadonnées communs

Réutilisable

- Licence d'utilisation appropriée
- (Qualité des documentations)

Projet Open OPSE

Principes « *Data Citation* » : Valoriser et encourager les efforts consacrés aux données

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: ***FORCE11; 2014***
https://doi.org/10.25490/a97f-egyk [Traduction française partielle et non officielle]

1. **Importance.** Les données doivent être considérées comme des produits de recherche légitimes et pouvant être cités. Les citations de données devraient avoir la même importance dans un dossier scientifique que les citations d'autres objets de recherche tels que les publications.
2. **Crédit et attribution.** Les citations de données devraient faciliter l'attribution d'un crédit scientifique …/… à tous les contributeurs des données …/…
3. **Preuve.** Dans la littérature scientifique, chaque fois qu'une affirmation repose sur des données, les données correspondantes doivent être citées.
4. **Identification unique.** Une citation de données doit inclure une méthode d'identification persistante qui soit exploitable informatiquement, unique au monde et largement utilisée par une communauté.
5. **Accès.** Les citations de données devraient faciliter l'accès aux données …/…
6. **Persistance.** Les identifiants uniques et les métadonnées décrivant les données et leur disposition doivent persister, même au-delà de la durée de vie des données …/…
7. **Spécificité et vérifiabilité.** Les citations de données devraient faciliter l'identification, l'accès et la vérification des données spécifiques à l'appui d'une affirmation. …/… faciliter la vérification que la tranche de temps, la version et/ou le niveau de granularité des données consultées ultérieurement sont les mêmes que celles citées à l'origine.
8. **Interopérabilité et flexibilité** …/…

Parmi ces principes, quels sont ceux qui complémentent « FAIR » ?

Projet Open OPSE

Principes « *Data Citation* » : Valoriser et encourager les efforts consacrés aux données

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: ***FORCE11; 2014***
https://doi.org/10.25490/a97f-egyk [Traduction française partielle et non officielle]

1. **Importance.** Les données doivent être considérées comme des produits de recherche légitimes et pouvant être cités. Les citations de données devraient avoir la même importance dans un dossier scientifique que les citations d'autres objets de recherche tels que les publications.
2. **Crédit et attribution.** Les citations de données devraient faciliter l'attribution d'un crédit scientifique …/… à tous les contributeurs des données …/…
3. **Preuve.** Dans la littérature scientifique, chaque fois qu'une affirmation repose sur des données, les données correspondantes doivent être citées.
4. **Identification unique.** Une citation de données doit inclure une méthode d'identification persistante qui soit exploitable informatiquement, unique au monde et largement utilisée par une communauté.
5. **Accès.** Les citations de données devraient faciliter l'accès aux données …/…
6. **Persistance.** Les identifiants uniques et les métadonnées décrivant les données et leur disposition doivent persister, même au-delà de la durée de vie des données …/…
7. **Spécificité et vérifiabilité.** Les citations de données devraient faciliter l'identification, l'accès et la vérification des données spécifiques à l'appui d'une affirmation. …/… faciliter la vérification que la tranche de temps, la version et/ou le niveau de granularité des données consultées ultérieurement sont les mêmes que celles citées à l'origine.
8. **Interopérabilité et flexibilité** …/…

ⓘ Complémentarité avec les principes FAIR

Comment bien préparer un jeu de données ?

Jeu de Rôle

Scénario

Un chercheur très occupé a préparé par lui-même un jeu de données dans un entrepôt *Dataverse*.

Il souhaite publier rapidement ce jeu de données, mais vous sollicite pour le vérifier.

Votre mission

- ✓ Trouver *ce qui ne convient pas très bien…*
- ✓ Expliquer pourquoi : convaincre le chercheur de la nécessité d'améliorer le jeu de données
- ✓ Proposer des changements : où ? comment ?

Projet Open OPSE

Démarche Qualité et dépôts de données : pour aller plus loin

Exemple : Guide de dépôt d'un jeu de données dans l'entrepôt de l'IRD (*DataSuds*)

Métadonnées

Champs	Réf.	**Préconisations** ou/et **recommandations** - Conseils pratiques	**Finalité et commentaires**
tous	L1	Saisir les informations en anglais de préférence. Afficher l'interface de DataSuds (les formulaires) en anglais peut faciliter la saisie : changer de langue dans le menu principal en haut de l'écran.	Visibilité et valorisation du jeu de données : recommandé mais pas obligatoire, comme pour les articles scientifiques
	L2	Ne pas mélanger différentes langues, sauf éventuellement pour le champ « Description » (dans ce cas, ajouter une séparation entre les langues avec le code <hr>) ainsi que les mots-clés. Possibilité de saisir une traduction du titre dans le champ « Autre titre » car un titre dans la langue du pays améliore la visibilité au niveau local.	Clarté de la présentation du jeu de données. Dans l'entrepôt, les champs de métadonnées ne sont pas multilingues, sauf exception : il n'est pas possible de saisir les informations traduites dans différentes langues. La langue sélectionnée dans le menu supérieur s'applique uniquement à l'interface utilisateur.
Titre	T1	Spécificité et caractérisation des données : type de données, contexte, période de collecte ou/et localisation géographique - si applicable et pertinent. Autre possibilité de titre : « *Replication data for*…(insérer le titre de l'article scientifique associé aux données)… » Exemples : consulter les jeux de données publiés récemment dans DataSuds. Pour davantage de conseils : https://coop-ist.cirad.fr/rediger/article-scientifique/le-titre/1-le-titre-premier-niveau-de-selection-sur-le-web	Selon la formulation et la précision du titre, un utilisateur de données potentiel ira - ou non - consulter plus en détails le jeu de données.
	T2	Longueur appropriée, approximativement entre 3 et 20 mots	Suivre les usages, comme pour un article scientifique.
	T3	Retirer les informations trop détaillées qui n'ont en général pas leur place dans un titre : noms de fichiers, noms d'auteurs, citation complète de l'article associé, parenthèses inutiles, caractères spéciaux.	
Auteurs	A1	Personnes qui ont contribué à la production des données : rôle scientifique ou technique : conception, collecte, traitement, analyse. Le responsable du projet valide la liste des auteurs. Conseils : 1) Utiliser le bouton [+] pour ajouter des lignes au formulaire. 2) Attention, ce champ est prérempli par DataSuds avec le nom de la personne qui dépose « techniquement » les données : cette personne n'est pas nécessairement 1er auteur, parfois pas même auteur … à corriger si nécessaire. 3) Une méthode consiste à reprendre la liste des auteurs d'un article associé aux données - ou encore de modifier cette liste pour ajouter ou mettre en avant des intervenants ayant joué un rôle important dans la collecte ou le traitement des données.	Procéder comme pour un article scientifique dans le choix et l'ordre des auteurs. Data Citation Principles
	A2	Format : « Nom, Prénom », avec les noms et prénoms en lettres minuscules. Exemple : Dupont, Jean	Suivre les usages, comme pour un article scientifique. Le format des auteurs se retrouve dans la citation du jeu de données.

https://data.ird.fr/wp-content/uploads/2021/02/DataSuds_qualite_depots_202102v23_LD.pdf

Démarche Qualité et dépôts de données : pour aller plus loin

Exemple : Guide de dépôt d'un jeu de données dans l'entrepôt de l'IRD (*DataSuds*)

Métadonnées

Champs	Réf.	Préconisations ou/et recommandations - Conseils pratiques	Finalité et commentaires
tous	L1	Saisir les informations en anglais de préférence. Afficher l'interface de DataSuds (les formulaires) en anglais peut faciliter la saisie : changer de langue dans le menu principal en haut de l'écran.	Visibilité et valorisation du jeu de données : recommandé mais pas obligatoire, comme pour les articles scientifiques
	L2	Ne pas mélanger différentes langues, sauf éventuellement pour le champ « Description » (dans ce cas, ajouter une séparation entre les langues avec le code <hr>) ainsi que les mots-clés. Possibilité de saisir une traduction du titre dans le champ « Autre titre » car un titre dans la langue du pays améliore la visibilité au niveau local.	Clarté de la présentation du jeu de données. Dans l'entrepôt, les champs de métadonnées ne sont pas multilingues, sauf exception : il n'est pas possible de saisir les informations traduites dans différentes langues. La langue sélectionnée dans le menu supérieur s'applique uniquement à l'interface utilisateur.
Titre	T1	Spécificité et caractérisation des données : type de données, contexte, période de collecte ou/et localisation géographique - si applicable et pertinent. Autre possibilité de titre : « *Replication data for* .. (insérer le titre de l'article scientifique associé aux données)… » Exemples : consulter les jeux de données publiés récemment dans DataSuds. Pour davantage de conseils : https://coop-ist.cirad.fr/rediger/article-scientifique/le-titre/1-le-titre-premier-niveau-de-selection-sur-le-web	Selon la formulation et la précision du titre, un utilisateur de données potentiel ira - ou non - consulter plus en détails le jeu de données.
	T2	Longueur appropriée, approximativement entre 3 et 20 mots	Suivre les usages, comme pour un article scientifique.
	T3	Retirer les informations trop détaillées qui n'ont en général pas leur place dans un titre : noms de fichiers, noms d'auteurs, citation complète de l'article associé, parenthèses inutiles, caractères spéciaux.	
Auteurs	A1	Personnes qui ont contribué à la production des données : rôle scientifique ou technique : conception, collecte, traitement, analyse. Le responsable du projet valide la liste des auteurs. Conseils : 1) Utiliser le bouton [+] pour ajouter des lignes au formulaire. 2) Attention, ce champ est prérempli par DataSuds avec le nom de la personne qui dépose « techniquement » les données : cette personne n'est pas nécessairement 1er auteur, parfois pas même auteur … à corriger si nécessaire. 3) Une méthode consiste à reprendre la liste des auteurs d'un article associé aux données - ou encore de modifier cette liste pour ajouter ou mettre en avant des intervenants ayant joué un rôle important dans la collecte ou le traitement des données.	Procéder comme pour un article scientifique dans le choix et l'ordre des auteurs. Data Citation Principles
	A2	Format : « Nom, Prénom », avec les noms et prénoms en lettres minuscules. Exemple : Dupont, Jean	Suivre les usages, comme pour un article scientifique. Le format des auteurs se retrouve dans la citation du jeu de données.

- ➔ Destiné aux intervenants qui relisent/vérifient les jeux de données avant procéder à leur publication
- ➔ Liste les critères à suivre et leurs justifications (+ *checklist*)
- ➔ Pour les déposants : anticiper la relecture, aider à finaliser plus rapidement les jeux de données
- ➔ Evolution du guide à partir des situations (parfois imprévues) réellement rencontrées

Démarche Qualité et dépôts de données : pour aller plus loin

Exemple : Guide de dépôt d'un jeu de données dans l'entrepôt de l'IRD (*DataSuds*)

Métadonnées

Champs	Réf.	Préconisations ou/et recommandations - Conseils pratiques	Finalité et commentaires
tous	L1	Saisir les informations en anglais de préférence. Afficher l'interface de DataSuds (les formulaires) en anglais peut faciliter la saisie : changer de langue dans le menu principal en haut de l'écran.	Visibilité et valorisation du jeu de données : recommandé mais pas obligatoire, comme pour les articles scientifiques
	L2	Ne pas mélanger différentes langues, sauf éventuellement pour le champ « Description » (dans ce cas, ajouter une séparation entre les langues avec le code <hr>) ainsi que les mots-clés. Possibilité de saisir une traduction du titre dans le champ « Autre titre » car un titre dans la langue du pays améliore la visibilité au niveau local.	Clarté de la présentation du jeu de données. Dans l'entrepôt, les champs de métadonnées ne sont pas multilingues, sauf exception : il n'est pas possible de saisir les informations traduites dans différentes langues. La langue sélectionnée dans le menu supérieur s'applique uniquement à l'interface utilisateur.
Titre	T1	Spécificité et caractérisation des données : type de données, contexte, période de collecte ou/et localisation géographique - si applicable et pertinent. Autre possibilité de titre : « *Replication data for…* (insérer le titre de l'article scientifique associé aux données)… » Exemples : consulter les jeux de données publiés récemment dans DataSuds. Pour davantage de conseils : https://coop-ist.cirad.fr/rediger/article-scientifique/le-titre/1-le-titre-premier-niveau-de-selection-sur-le-web	Selon la formulation et la précision du titre, un utilisateur de données potentiel ira - ou non - consulter plus en détails le jeu de données.
	T2	Longueur appropriée, approximativement entre 3 et 20 mots	Suivre les usages, comme pour un article scientifique.
	T3	Retirer les informations trop détaillées qui n'ont en général pas leur place dans un titre : noms de fichiers, noms d'auteurs, citation complète de l'article associé, parenthèses inutiles, caractères spéciaux.	
Auteurs	A1	Personnes qui ont contribué à la production des données : rôle scientifique ou technique : conception, collecte, traitement, analyse. Le responsable du projet valide la liste des auteurs. Conseils : 1) Utiliser le bouton [+] pour ajouter des lignes au formulaire. 2) Attention, ce champ est prérempli par DataSuds avec le nom de la personne qui dépose « techniquement » les données : cette personne n'est pas nécessairement 1er auteur, parfois pas même auteur … à corriger si nécessaire. 3) Une méthode consiste à reprendre la liste des auteurs d'un article associé aux données - ou encore de modifier cette liste pour ajouter ou mettre en avant des intervenants ayant joué un rôle important dans la collecte ou le traitement des données.	Procéder comme pour un article scientifique dans le choix et l'ordre des auteurs. Data Citation Principles
	A2	Format : « Nom, Prénom », avec les noms et prénoms en lettres minuscules. Exemple : Dupont, Jean	Suivre les usages, comme pour un article scientifique. Le format des auteurs se retrouve dans la citation du jeu de données.

Proposition : créer votre propre guide ?

davantage personnalisé en fonction de :

- ✓ votre laboratoire
- ✓ vos thématiques de recherche
- ✓ les types de données plus fréquemment utilisés
- ✓ les standards de métadonnées dans votre domaine