Guide de dépôt d'un jeu de données : Qualité de l'entrepôt de données *DataSuds*

Version 2.3, février 2021 – Luc Decker (administrateur de DataSuds),
Hanka Hensens, Caroline Doucouré et Pascal Aventurier – Service IST/MCST, IRD.

Métadonnées

Champs	Réf.	**Préconisations** ou/et **recommandations** - Conseils pratiques	Finalité et commentaires
tous	L1	Saisir les informations en anglais de préférence. Dans ce cas, afficher l'interface (les formulaires) de DataSuds en anglais facilite la saisie : changer de langue dans le menu 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 dans le champ « Description » (dans ce cas, ajouter une séparation entre les langues avec le code <hr>) ainsi que pour les mots-clés. Possibilité de saisir une traduction du titre dans le champ « Autre titre ». Un titre dans une langue locale peut améliorer la visibilité du jeu de données. La langue sélectionnée dans le menu supérieur s'applique uniquement à l'interface utilisateur.	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.
	L3	Orthographe et grammaire : vérifier tous les textes saisis dans les métadonnées. A partir de l'onglet « métadonnées », copier l'ensemble du texte vers un logiciel de traitement de texte (tel que *Word)* et lancer la vérification.	Prouve que les informations ont été relues avant publication, témoigne du soin apporté dans la gestion des données. Réputation de l'entrepôt et des auteurs des données déposées.
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 qui n'ont en général pas leur place dans un titre : noms de fichiers, noms d'auteurs, citation complète d'un article, 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 est repris dans la citation du jeu de données.

Affiliations des auteurs	A3	Format : Structure –Tutelles (si applicable) – Pays Exemple : "UMR DIADE – IRD, University of Montpellier, CIRAD, CNRS – France" Les informations préremplies dans le formulaire doivent en général être modifiées.	Informations complètes sur les auteurs, utiles pour savoir où ils travaillent, pour les contacter, notamment en cas d'homonymie.
Identifiants ORCID	A4	Saisir les identifiants ORCID des auteurs lorsqu'ils sont connus. Ne pas retarder la publication s'il manque quelques identifiants, ils pourront être ajoutés ultérieurement.	ORCID devient un standard
Description	D1	Précision générale : est-elle suffisante pour le référencement adéquat des données par les moteurs de recherche ?	Principe **F**AIR « Facile à trouver » : les autres chercheurs potentiellement intéressés vont-t-ils trouver facilement ces données ? Quels mots clés sont-ils le plus susceptibles de saisir lorsqu'ils lancent une recherche ?
	D2	Contexte, périmètre, typologie des données - Résumer le projet scientifique associé ou/et *l'intérêt, l'objectif* de ces données - Résumer la liste des données déposées (« quoi ? ») ; comment, où et quand ont-elles été collectées/traitées ?	Répondre aux questions que les utilisateurs potentiels sont susceptibles de se poser avant d'aller plus loin et de commencer à utiliser les données. Participe au bon référencement du jeu de données. Principe FAI**R** « Réutilisable».
	D3	Possibilité d'ajouter des liens cliquables vers des pages web telles que la description du projet, le site du bailleur, etc… (code texte du lien)	Facilite la démarche des utilisateurs et en général apprécié par les sites référencés.
	D4	Mise en page de la description : créer éventuellement des paragraphes ou/et insérer des sauts de ligne à l'aide du code . Possibilité de mettre du texte en gras (code …) ou en italique (code …).	Facilite la lecture de la description. Améliore la présentation du jeu de données
Mots-clés	D5	Saisir au moins 4-5 mots-clés. Précision : comme pour le champ « Description ». Utiliser un vocabulaire (thesaurus) de référence dans son domaine facilite la découverte par les scientifiques de ce domaine.	Suivre les usages, comme pour un article scientifique.
Publication connexe	P1	Saisir les références complètes des publications liées au jeu de données. Possibilité de préciser « *(submitted)* » le cas échéant. Utiliser le bouton [+] pour ajouter des lignes au formulaire.	Les articles citent les données, avec leur identifiant DOI, et réciproquement. Augmente les citations.
	P2	Saisir l'identifiant pérenne (DOI) et/ou le lien (http…) des publications.	Facilite la navigation des utilisateurs.
Renseignements sur la subvention	M1	Ce champ de métadonnées optionnel permet de citer les bailleurs qui ont permis la réalisation du projet. Le contrat conclu avec un bailleur impose parfois que cela soit effectué systématiquement. Il est possible d'ajouter des liens vers des sites web.	Bonnes relations avec les bailleurs qui ont également besoin de visibilité et de montrer l'impact des financements accordés.
Type de données, **Période de collecte,** **Langue**	M2	Remplir ces différents champs de métadonnées (lorsqu'ils sont applicables) aide à bien décrire les données, bien que le formulaire de saisie autorise qu'ils soient laissés vides. Note : le formulaire qui permet d'éditer les métadonnées comprend davantage de champs que le formulaire initial (simplifié) utilisé pour créer un nouveau jeu de données.	Référencement des données et interopérabilité. Aide à la réutilisation.
Métadonnées géospatiales	M3	Ce champ de métadonnées est à saisir si les données ont été collectées dans un/des périmètre(s) géographique(s) déterminé(s) : pays, villes..., sauf si la localisation n'a aucune importance.	

Fichiers déposés

Elément	#	**Préconisations** ou/et **recommandations** - Conseils pratiques	Finalité et commentaires
Choix des fichiers de données à diffuser	F1	Il n'est pas toujours permis de diffuser (partager) librement toutes les données d'un projet de recherche. Il convient de respecter des obligations légales et contractuelles, de vérifier également la propriété des données (exemples : partenaires impliqués ; réutilisation de données existantes : licences ou accords des fournisseurs). Les principaux points de vigilance sont décrits sur https://data.ird.fr/cadre-juridique/ Apporter une attention particulière en cas de données personnelles (exemple : résultats d'enquêtes) et de données dites « sensibles ». Dans presque tous les cas (sauf obligation de confidentialité), il reste possible de publier uniquement les métadonnées dans l'entrepôt, sans déposer de fichier de données. L'entrepôt permet enfin de restreindre l'accès à certains fichiers, sous la forme d'un formulaire de demande d'accès qui devra être rempli par les utilisateurs et d'autorisations qui leur seront accordées au cas par cas. Il est recommandé de préserver également les données brutes dans l'entrepôt, en complément des données traitées, dérivées ou d'intérêt - d'autant plus si l'espace occupé par les données brutes est négligeable par rapport aux capacités de stockage actuelles	Il ne s'agit pas d'ouvrir systématiquement et indistinctement l'accès à toutes les données : « aussi ouvert que possible, *aussi fermé que nécessaire* ». Des fichiers de données peuvent être ajoutés ultérieurement après vérifications complémentaires. Après publication, un jeu de données peut être mis à jour, ce qui produit de nouvelles versions (V2, V3, …). Conserver la possibilité de retraiter les données brutes pour 1) appliquer, tester, comparer de nouvelles méthodes mathématiques qui pourraient être conçues dans le futur, 2) corriger une erreur de traitement éventuelle (bug logiciel, etc…)
Volume de données	F2	La taille de *chaque fichier* déposé ne doit pas dépasser 975 Méga-octets. Cette limite est à considérer *après compression* éventuelle du fichier. La limite imposée sur la taille des fichiers est arbitraire, fixée par la configuration de DataSuds. Il serait possible de l'augmenter. Un jeu de données peut cependant rassembler des dizaines ou même centaines de fichiers. Avant de déposer un grand volume de données (> 5 Gb), prendre contact pour conseil auprès de l'équipe support de DataSuds data@ird.fr Immédiatement après avoir été déposés, les fichiers ZIP sont décompressés automatiquement par l'entrepôt. Leur contenu est extrait et présenté aux utilisateurs sous forme d'une liste de fichiers. Cependant et si l'on souhaite conserver les données au format ZIP, créer et déposer un fichier ZIP temporaire qui contient lui-même le fichier ZIP qui doit être présenté aux utilisateurs.	DataSuds n'a pas été conçu pour les données dites « massives », en particulier pour les résultats bruts de séquençage génomique. Des entrepôts thématiques (par exemple *GenBank*) sont spécialisés dans la préservation et la diffusion de telles données. Il n'est pas toujours réalisable de conserver l'ensemble des données brutes d'un projet lorsque leur volume est très important. L'intérêt des données doit être mis en balance avec les coûts engendrés par leur stockage à très long terme. Il est parfois moins couteux de préserver physiquement les échantillons. Même si l'utilisation de DataSuds est gratuite, la question est à considérer.
Nombre de fichiers	F3	Déposer au plus quelques centaines de fichiers par jeu de données. Au-delà et si nécessaire, regrouper et diffuser les fichiers dans des archives au format ZIP, suivant le conseil décrit ci-dessus. Inversement, il est possible de publier un jeu de données qui ne comprend aucun fichier, donc uniquement des métadonnées, et d'indiquer aux utilisateurs la procédure à suivre pour obtenir les données : personne à contacter, lien vers un site externe, attente de la fin d'un embargo, etc…	

Noms des fichiers	F4	Adopter des noms de fichiers spécifiques au projet et au contenu. Autant que possible, leur ajouter un préfixe propre au projet (un acronyme ou un identifiant) qui sera commun à tous les fichiers déposés, tel que « PROJETABC_stationN_dailyflow.csv ». Eviter les noms génériques tels que « data.csv », « datasuds_table1.csv », « tableau4.xls », « documentation.pdf » … L'interface de DataSuds permet de renommer un fichier sans avoir à le redéposer : ouvrir le formulaire d'édition des métadonnées du fichier.	Evite que les utilisateurs ne mélangent ou confondent des fichiers aux noms identiques provenant de différentes sources. Ajouter un identifiant à la fin des noms de fichiers ne permet pas de les trier convenablement : un préfixe est donc préférable. Par défaut, Dataverse trie les fichiers par leur nom. Cela permet de prévoir l'ordre dans lequel les fichiers seront présentés aux utilisateurs et de les organiser.
	F5	Dans les noms de fichiers, utiliser uniquement des lettres, chiffres et caractères séparateurs (**-** ou _). Remplacer les autres caractères (spéciaux ou accentués) et les espaces.	Principe FA**I**R « Interopérable »
	F6	Limiter les noms de fichiers à une longueur raisonnable : au plus 40 caractères sauf besoin particulier.	
Format des fichiers	F7	Autant que possible, déposer les fichiers dans un format « ouvert » tel que CSV ou texte… se référer à https://fr.wikipedia.org/wiki/Format_ouvert et https://doranum.fr/wp-content/uploads/FS2_liste_indicative_formats_V1.pdf Dans le cas de fichiers au format texte, utiliser l'encodage de caractères UTF-8 de préférence (un outil gratuit pour éditer et convertir les fichiers texte : Notepad++). Dans le cas de documents PDF, les enregistrer avec l'option « format PDF/A » Si la conversion des données dans un format ouvert occasionne une perte d'information (dans le cas d'un tableau Excel converti au format CSV : la mise en page et le formatage), déposer également le fichier original en plus du fichier converti. Attention : les fichiers CSV enregistrés à l'aide d'Excel avec la version française de Windows utilisent le caractère point-virgule comme séparateur de champs. Dataverse s'attend à ce que le caractère séparateur soit une virgule, selon le standard anglophone : il ne parvient pas à analyser correctement le contenu du fichier. Pour remédier à ce problème, créer ou convertir le fichier CSV à l'aide du logiciel gratuit LibreOffice, qui permet de choisir le caractère séparateur aussi bien à l'ouverture qu'à la sauvegarde d'un fichier (option « Editer les paramètres du filtre »).	Principe FA**I**R « Interopérable »
Description des fichiers	F8	Remplir le champ « Description » attaché à chaque fichier, dans le formulaire d'édition des métadonnées des fichiers : résumer le contenu en quelques mots ou davantage.	Aider les utilisateurs à identifier le contenu de chaque fichier
	F9	Attribuer un ou plusieurs libellés (*tags*) à chaque fichier, comme le permet l'entrepôt : « Data », « Documentation », « Code ». Il est possible de créer des libellés personnalisés. Passer l'interface en anglais avant d'effectuer cette opération afin d'éviter d'utiliser le libellé « Données ».	Organisation, catégorisation des fichiers déposés

Noms de dossiers	F10	Définir un nom de dossier pour chaque fichier déposé. Pour le moins, il peut s'agir d'un nom de dossier commun à tous les fichiers, correspondant au nom du projet de recherche. Le nom de dossier (« Chemin d'accès au fichier ») peut être saisi manuellement dans le formulaire d'édition des métadonnées des fichiers. Il est possible de créer une arborescence avec plusieurs niveaux de dossiers (tel que dossier1/dossier2). Si un fichier ZIP contenant des dossiers est déposé dans DataSuds, son contenu est extrait automatiquement et les noms de dossiers sont préremplis.	Organisation, catégorisation des fichiers déposés. Lorsqu'un utilisateur sélectionne plusieurs fichiers à télécharger –--ou bien l'ensemble des fichiers-, l'entrepôt crée une archive ZIP qui reprend les différents dossiers qui auront été définis. Cela permettra aux utilisateurs de recréer automatiquement l'arborescence des dossiers et de préserver l'organisation des fichiers. Cela évitera aussi que les fichiers extraits par les utilisateurs ne se mélangent à d'autres par erreur.
Dictionnaire des données	F11	Si les données sont déposées sous la forme de tables (CSV, Excel…), il est essentiel de préciser la signification de chaque variable, « champ » ou « colonne ». Cette information est souvent documentée sous la forme d'un dictionnaire des données (*data dictionary*). Un dictionnaire des données est constitué d'un ou plusieurs tableaux de référence au format CSV (ou texte) avec les colonnes suivantes : « nom de la variable », « contenu/signification de la variable », « unité de mesure », « signification des différents codes utilisés » (si applicables) ; format (optionnel). Il peut être commun à plusieurs tables ou fichiers ; il peut aussi être intégré dans un fichier de données. Si de nombreux codes sont utilisés, il est pratique de les documenter dans des tableaux séparés (*code dictionary*).	Principe FAI**R** « Réutilisable » : en l'absence de la signification précise des variables et des codes utilisés, la réutilisation des données s'avère difficile ou même impossible ; elle conduit à un risque important de mauvaise interprétation des données.
Documentations annexes	F12	En complément des fichiers de données, l'entrepôt accepte aussi toutes les documentations qui vont aider à comprendre les données, à préserver leur histoire, les conditions de leur collecte ou/et de leur production. Voici des suggestions de documentations qui peuvent accompagner des données : fiche de présentation du projet de recherche (éventuellement une présentation Powerpoint) ; figures, schémas, cartes, photographies… ; formulaires vierges de collecte de données ou/et de recueil du consentement des participants ; guide de l'enquêteur ; guide de traitement des données ; procédure ou algorithme de traitement des données, code informatique ; notes techniques ; Plan de Gestion de Données. Une exception : les articles scientifiques (dans toutes leurs versions, dont les *preprint*) ne doivent pas être déposés dans DataSuds. Certaines documentations sont considérées comme des « œuvres de l'esprit », telles que les textes rédigés : demander l'accord de leurs auteurs avant diffusion.	Principe FAI**R** « Réutilisable » : 1) améliorer le potentiel de réutilisation des données, 2) aider à comprendre pourquoi la reproduction de l'expérience pourrait donner un résultat différent, si certains paramètres ou conditions ont changé.
	F13	Insérer la citation du jeu de données dans une ou plusieurs documentations déposées avec les données. La citation est connue à l'avance, alors que le jeu de données est encore en préparation (le DOI est pré-attribué mais sera mis en service lorsque le jeu de données sera publié). Le titre du jeu de données et la liste des auteurs doivent être finalisés, puis copier la citation à partir de la page d'accueil du jeu de données et remplacer « version provisoire » par « V1 ».	Faciliter la citation des données, qui peuvent être ré-utilisées longtemps après avoir été téléchargées : aider à conserver trace de la source des données.

Conditions d'utilisation

Elément	#	**Préconisations** ou/et **recommandations** - Conseils pratiques	Finalité et commentaires
Attribution d'une licence	L1	Le formulaire en ligne https://creativecommons.org/choose/?lang=fr peut aider à choisir une licence applicable aux utilisateurs qui téléchargeront les données. Par défaut, l'entrepôt DataSuds attribue une licence « CC-BY » (ou même dans certains cas « CC0 » = domaine public) aux jeux de données, voir onglet « Conditions d'utilisation » du jeu de données. A minima, vérifier que cette licence correspond à ce qui est souhaité. La licence CC-BY exige que le jeu de données soit cité lorsqu'il est utilisé ; elle est conforme aux préconisations du MESRI pour les projets sur financement public. Si les modèles de licence standard (*Creative Commons* ou autres) ne conviennent pas, indiquer alors les conditions particulières à remplir pour l'accès aux données. Il peut s'agir par exemple de la nécessité de signer un accord (*data transfer agreement*). Préciser les informations de contact pour la soumission des demandes d'accès aux données. Note : il est possible d'attribuer une licence distincte aux documentations qui accompagnent les données.	Principe FAI**R** « Réutilisable » : répond à la question des utilisateurs « dans quelles conditions est-t-il possible de réutiliser ces données ? »

Qualité de l'entrepôt de données *DataSuds* : Grille de révision d'un jeu de données

Version 2.3, février 2021 – Luc Decker (administrateur de DataSuds, service IST/MCST, IRD)

Collection (dataverse) Cliquez ou appuyez ici pour entrer du texte. Relecteur .....

Jeu de données Cliquez ou appuyez ici pour entrer du texte. Date Jour / Mois / Année

Critère	OK	partiellement OK	à améliorer, requis	à améliorer, conseillé	non applicable
L1. Métadonnées en anglais de préférence	☐	☐	☐	☐	☐
L2. Langue unique (sauf description, autre titre et mots-clés)	☐	☐	☐	☐	☐
L3. Vérification de l'orthographe et de la grammaire	☐	☐	☐	☐	☐
T1. Titre : précision, spécificité	☐	☐	☐	☐	☐
T2. Titre : longueur	☐	☐	☐	☐	☐
T3. Titre : formulation et format	☐	☐	☐	☐	☐
A1. Auteurs : choix et nombre	☐	☐	☐	☐	☐
A2. Auteurs : format des noms	☐	☐	☐	☐	☐
A3. Auteurs : format des affiliations	☐	☐	☐	☐	☐
A4. Auteurs : identifiants ORCID	☐	☐	☐	☐	☐
D1. Description : précision, pour bon référencement	☐	☐	☐	☐	☐
D2. Description : contexte, périmètre, typologie des données	☐	☐	☐	☐	☐
D3. Description : liens vers d'autres contenus	☐	☐	☐	☐	☐
D4. Description : mise en page	☐	☐	☐	☐	☐
D5. Mots-clés : nombre, précision	☐	☐	☐	☐	☐
P1. Publications associées : citations	☐	☐	☐	☐	☐
P2. Publications associées : DOI et/ou liens	☐	☐	☐	☐	☐
M1. Citation des bailleurs de fonds	☐	☐	☐	☐	☐
M2. Saisie de Type de données, Période de collecte, Langue	☐	☐	☐	☐	☐
M3. Saisie des métadonnées géospatiales	☐	☐	☐	☐	☐
F1. Fichiers : permission de diffuser, droits & réglementations	☐	☐	☐	☐	☐
F2. Volume de données déposées	☐	☐	☐	☐	☐
F3. Nombre de fichiers déposés	☐	☐	☐	☐	☐
F4. Noms de fichiers spécifiques au projet et à leur contenu	☐	☐	☐	☐	☐
F5. Caractères utilisés dans les noms des fichiers	☐	☐	☐	☐	☐
F6. Longueur des noms des fichiers	☐	☐	☐	☐	☐
F7. Format des fichiers de données	☐	☐	☐	☐	☐
F8. Saisie des descriptions des fichiers	☐	☐	☐	☐	☐
F9. Attribution de catégories (*tags*/libellés) aux fichiers	☐	☐	☐	☐	☐
F10. Saisie de nom(s) de dossier(s)	☐	☐	☐	☐	☐
F11. Dictionnaire des données (variables, codes, unités)	☐	☐	☐	☐	☐
F12. Dépôt de documentations associées	☐	☐	☐	☐	☐
F13. Citation du jeu de données insérée dans la documentation	☐	☐	☐	☐	☐
L1. Choix d'une licence d'utilisation des données	☐	☐	☐	☐	☐
	OK	partiellement OK	à améliorer, requis	à améliorer, conseillé	non applicable