Guide de relecture d'un jeu de données avant publication
(qualité de l'entrepôt de données DataSuds)

Version 3.0, septembre 2023 – Luc Decker – Service IST, MCST, IRD, France
doi:10.5281/zenodo.5884670 contact : data@ird.fr

Présentation

Dans le cadre d'une démarche Qualité, ce guide technique documente 36 points de vérification (accompagnés de recommandations) destinés aux relecteurs ou « curateurs » de jeux de données (*datasets*) avant leur publication dans DataSuds (https://dataverse.ird.fr). Entrepôt institutionnel de données scientifiques de l'IRD, DataSuds fonctionne grâce à l'application Dataverse. Le guide peut également être consulté par les déposants de données à la recherche de renseignements sur quelques points précis, de façon à finaliser plus rapidement un jeu de données.

Sous forme de tableau synthétique, le guide est divisé en 3 sections : "Métadonnées" (listées dans l'ordre du formulaire de saisie Dataverse), "Fichiers déposés" et "Conditions d'utilisation". Les principales lignes directrices sont le respect des principes FAIR, des principes Data Citation, la qualité éditoriale et les bons usages. La finalité des recommandations (comprendre le "Pourquoi") est précisée dans une colonne dédiée à cet effet. Des conseils pratiques ainsi que des liens vers des explications plus détaillées sont également fournis. Une grille pratique (*checklist*) résume ensuite la liste des critères en une page, conçue pour être imprimée.

Dans une approche d'amélioration continue, ce guide évolue au fil du temps. Certains conseils peuvent sembler évidents. Cependant tous reflètent nos observations, les cas réellement rencontrés durant la révision de 250 jeux de données. Ces éléments peuvent aussi être appliqués aux jeux de données destinés à être publiés dans d'autres entrepôts. Les critères de curation spécifiques à l'entrepôt DataSuds concernent principalement les types de données acceptés (cf. F1) et leur volume (F2).

Sommaire

- Tableaux de recommandations destinés à la vérification des métadonnées, des fichiers déposés et des conditions d'utilisation, accompagnés de conseils pratiques et de justifications.
- Grille de relecture d'un jeu de données *(checklist)* : une page à imprimer.
- Annexe 1. Conseils de mise en œuvre.
- Annexe 2. Présentation introductive : « Pourquoi et Comment publier un jeu de données dans une démarche guidée par la qualité ? ».
- Annexe 3. Ressources & liens divers.

Historique des mises à jour

Version 3.0 (09/2023)	Ajout d'une page de présentation, de 3 annexes et enfin de divers détails dans les sections F1, F13, F14, L1 et L2.
Version 2.8 (02/2023)	Informations ajoutées (surlignées en jaune) dans les sections "Description", "Choix des fichiers", "Format des fichiers", "Attribution de licences". Mise à disposition de la version Word du document afin de faciliter son réemploi (licence CC-BY-SA).
Version 2.7 (12/2022)	Précisions, clarifications et reformulations. Une attention particulière est portée à la concision du guide, de sorte que le nombre de pages reste inchangé.
Version 2.6 (08/2022)	Ajout de quelques précisions.
Version 2.5 (07/2022)	La grille (page 8) a été révisée en fonction de son expérience d'utilisation : une nouvelle colonne "Edité" aide à tracer les modifications apportées au jeu de données, afin de les rapporter ensuite au déposant ; une autre colonne est à présent destinée à répertorier les questions à poser au déposant.

Remerciements

Hanka Hensens, Caroline Doucouré et Pascal Aventurier (Service IST, MCST, IRD) ont contribué à la révision de versions antérieures (1.0 - 2.4) de ce guide.

Guide de relecture d'un jeu de données avant publication (Qualité de l'entrepôt de données DataSuds)

version 3.0 (09/2023) – Luc Decker – Service IST, MCST, IRD, France doi:10.5281/zenodo.5884670 https://dataverse.ird.fr
Contributeurs : Hanka Hensens, Caroline Doucouré, Pascal Aventurier – Service IST, MCST, IRD, France

Métadonnées

Champs	Réf.	**Préconisations,** recommandations et **conseils pratiques**	**Finalité & commentaires**
applicable à tous	L1	En fonction du domaine scientifique, il est souvent recommandé de saisir les informations en anglais ; à plus forte raison si les données accompagnent un article rédigé en anglais. Dans ce cas, l'interface en anglais facilite la saisie : avant de débuter, changer de langue à l'aide du menu en haut à droite de l'écran.	Visibilité et valorisation scientifique du jeu de données, comme pour les articles.
	L2	Ne pas mélanger différentes langues, sauf éventuellement pour le champ « Description » (ajouter une séparation entre les langues avec le code <hr>) ainsi que pour le champ « Autre titre » (possibilité de fournir une traduction : un titre dans une langue locale peut améliorer la visibilité du jeu de données) et pour les mots-clés. La langue sélectionnée dans le menu 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 des traductions dans différentes langues.
	L3	Orthographe et grammaire : vérifier tous les textes saisis dans les champs de métadonnées. Lorsque les saisies sont terminées, se rendre dans l'onglet « Métadonnées », copier/coller tout le texte dans un logiciel de traitement de texte (tel que *Word)*, sélectionner la langue utilisée et lancer la vérification automatique de l'orthographe.	Prouve que les informations ont été relues avant publication, témoigne du soin apporté dans la gestion des données. Réputation de l'entrepôt, des auteurs et de leur laboratoire.
Titre	T1	Caractérisation des données : type de données, contexte, période de collecte, localisation géographique - si applicable et pertinent. Si des données différentes pourraient recevoir le même titre, alors la précision du titre serait à améliorer. Le nom ou acronyme du projet peut être inclus mais ne constitue pas un titre à lui seul. Le cas échéant, le jeu de données et l'article associé devraient avoir des titres distincts ; une solution : « *Replication data for…* » ou « *Supplementary data for…* » […insérer le titre de l'article]. Les titres n'ont pas besoin d'inclure le terme générique « *dataset* » ou « *data* » ; quand c'est implicite, le retirer - ou idéalement, le remplacer par un terme plus précis. Exemples de titre : consulter les jeux de données publiés récemment dans DataSuds.	Selon la formulation et la précision du titre, un utilisateur de données potentiel ira - ou non - consulter plus en détails le jeu de données. Le titre doit être compréhensible de lui-même. Pour plus de conseils, consulter le site CoopIST (CIRAD)
	T2	Longueur appropriée, approximativement entre 3 et 20 mots	Suivre les usages. Un tel titre pourrait-t-il être attribué à un article, une présentation ou un rapport scientifique ?
	T3	Retirer les éléments qui n'ont en général pas leur place dans un titre : noms de fichiers, noms d'auteurs, citation complète d'un article, codes ou numéros de référence internes au projet, parenthèses inutiles, caractères spéciaux, mots en majuscules, point final.	
Auteurs	A1	Personnes qui ont contribué à la production des données, rôle scientifique, technique, logistique : collecte, traitement, vérification, calcul, conception d'outil, supervision. 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 avec le nom de la personne qui dépose (techniquement) les données : cette personne n'est pas nécessairement 1er auteur, parfois pas même auteur … à éditer si nécessaire. 3) Une approche consiste à recopier la liste des auteurs d'un article associé aux données - ou encore de modifier cette liste pour ajouter, mettre en avant des intervenants ayant joué un rôle important dans la collecte ou le traitement des données. Comme pour les articles, un champ « Contributeurs » est aussi disponible.	Procéder tout comme pour un article scientifique dans le choix et l'ordre des auteurs. Il est possible mais rare qu'un article ait un seul auteur, de même pour un jeu de données. Intégrité scientifique : reconnaitre toutes les contributions ; se référer aussi aux Data Citation Principles

	A2	Format : « Nom, Prénom », 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, présentée sur sa page d'accueil.
Affiliations des auteurs	A3	Format : Structure 1 **–** Tutelles (si applicable) **–** Pays **;** Structure 2 **–** Tutelles **–** Pays **;** … Exemple : "UMR DIADE **–** IRD**,** University of Montpellier**,** CIRAD**,** CNRS **–** France" Les informations préremplies 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 : est-elle suffisante pour le bon référencement des données dans les moteurs de recherche ? 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 ? Les formulations telles que *« Ces données accompagnent l'article … Pour tous les détails, consulter l'article »* n'atteignent pas cet objectif, elles doivent être complétées.	Principe **F**AIR : rendre les données « Faciles à trouver ».
	D2	Contexte, périmètre, typologie des données - Résumer le projet scientifique associé ou/et l'intérêt, l'objectif de ces données (« pourquoi ? ») - Résumer la liste des données déposées (« quoi ? ») ; comment, où, par qui et quand ont-elles été collectées/traitées ? Exemple : *« This dataset provides the following contents : … »* - Précisions diverses : particularités du format des données, documentation importante à lire, …	Répondre aux questions que les utilisateurs potentiels pourraient se poser avant d'aller plus loin et de télécharger les données. Participe au bon référencement du jeu de données : Principe FAI**R** « Réutilisables ».
	D3	Possibilité d'ajouter des liens Internet cliquables vers une description du projet, le site web du bailleur, un site de référence, etc… (code : texte du lien). Si la description est rédigée en anglais, fournir les liens vers les versions anglaises des contenus.	Facilite la démarche des utilisateurs et en général apprécié par les sites référencés.
	D4	Mise en page de la description : créer des paragraphes (code : <p>) ou insérer des sauts de ligne (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 termes. Précision : comme pour le champ « Description ». Utiliser le bouton [+] pour saisir un seul mot clé par case. Utiliser un vocabulaire (thesaurus) de référence dans son domaine facilite la découverte des données par les scientifiques de ce domaine.	Suivre les usages, comme pour un article scientifique.
Publication connexe (associée)	P1	Saisir les références complètes des publications associées au jeu de données. Le cas échéant, préciser *(in preparation)*, *(submitted)* ou *(accepted)*. 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 optionnel permet de citer les bailleurs et organisations qui ont soutenu 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. Citer les contributions de l'IRD : infrastructures, RH…	Bonnes relations avec les bailleurs qui ont également besoin de visibilité et de montrer l'impact des financements accordés.
Type de données, Période couverte, Période de collecte, Logiciel, Langue	M2	Remplir ces différents champs de métadonnées -lorsqu'ils sont pertinents- aide à mieux décrire les données. Le formulaire de saisie autorise cependant qu'ils soient laissés vides. Note : le formulaire qui permet d'éditer les métadonnées comprend davantage de champs que le formulaire initial utilisé pour créer un nouveau jeu de données.	Référencement des données et interopérabilité. Aide à la réutilisation des données.
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 est sans importance.	

Fichiers déposés

Elément	#	**Préconisations** ou/et **recommandations** - Conseils pratiques	Finalité et commentaires
Choix des fichiers de données à diffuser	F1	S'assurer de l'accord final du déposant avant de procéder à la publication du jeu de données. Vérifier qu'il s'agit bien de données : les documents tels que les articles et notes scientifiques, présentations, posters, rapports d'avancement de projets, manuels, procédures ou guides n'ont pas leur place dans DataSuds, sauf s'il s'agit de documenter des données (voir exemples dans la section F12) et en tel cas, le titre du jeu de données reste focalisé sur les données. Des données peuvent cependant se présenter sous forme de textes, donc de documents, notamment en Sciences Humaines et Sociales (transcriptions d'interviews, récits, …). La frontière entre documents et données est parfois floue ; contacter data@ird.fr pour prise en compte de cas particuliers. Si les données ont déjà été publiées dans un autre entrepôt et surtout si elles disposent déjà d'un identifiant DOI : ne pas les déposer une seconde fois, éviter toute duplication. Il n'est pas toujours autorisé de diffuser, de partager librement l'ensemble des données d'un projet de recherche. Il convient de respecter des obligations légales et contractuelles, de vérifier la propriété des données. Ces critères sont prioritaires par rapport aux principes de la Science Ouverte. Exemples : accord des partenaires impliqués ; réutilisation de données existantes : licences ou accords des fournisseurs ; droit d'auteur et copyright dans le cas de textes ou photographies ; consentement des participants. En l'absence d'autorisation, il n'est pas permis de rediffuser des extraits significatifs de bases de données commerciales, dont le *droit sui generis* protège le contenu. Les principaux points de vigilance sont décrits sur https://data.ird.fr/cadre-juridique/. En cas de réclamation (instruite par le service juridique ou par le Référent Intégrité Scientifique), l'entrepôt peut être conduit à retirer un jeu de données, à titre conservatoire ou définitivement. Apporter une attention particulière en cas de données personnelles (exemple : résultats d'enquêtes, interviews) ou de données dites « sensibles », en particulier en santé et sciences sociales. Pour des raisons réglementaires, la publication *en libre accès* de données personnelles sensibles n'est pas autorisée, même si elles ont été pseudonymisées ou dé-identifiées : elles conservent leur caractère personnel. Sous certaines conditions, une publication en accès restreint est possible. Prendre contact (data@ird.fr) pour conseil. Si les données ont été totalement anonymisées, des éléments concrets doivent prouver la qualité de ce traitement : appel à un expert reconnu, logiciel utilisé, indicateurs statistiques (k-anonymat, l-diversité, …). Vérifier de plus que le formulaire de consentement ne s'oppose pas de manière explicite au partage des données ou à leur réutilisation. Dans presque tous les cas (sauf obligation de confidentialité portant sur l'existence d'un projet), il reste possible de publier uniquement les métadonnées dans l'entrepôt, sans déposer de fichier de données. Si nécessaire, l'entrepôt permet de restreindre l'accès à certains fichiers, sous la forme d'un formulaire de demande d'accès qui devra être rempli par les utilisateurs : des autorisations d'accès leur seront accordées au cas par cas selon des conditions personnalisées que le déposant aura définies. Lorsque c'est techniquement possible, 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.	Après avoir rendu public un contenu, il n'est pas possible de « revenir 100% en arrière ». D'autres entrepôts ou plateformes de publication acceptent les documents, parfois même y sont dédiés, tels que l'archive ouverte HAL et Horizon IRD. Respecter les règles d'attribution de DOIs. Il ne s'agit pas d'ouvrir systématiquement et indistinctement l'accès à toutes les données : « aussi ouvert que possible, *aussi fermé que nécessaire* ». Les éditeurs et bailleurs diffusent parfois des consignes générales qui ignorent les règles propres à certains types de données et peuvent donc induire en erreur, inciter à ouvrir les données alors que cela n'est pas autorisé. Des fichiers de données peuvent être ajoutés ultérieurement, après vérifications complémentaires. Un jeu de données peut être facilement mis à jour, sous la forme de nouvelles versions (V2, V3, …). Attention : les utilisateurs peuvent consulter l'historique des versions et consulter/télécharger les versions antérieures. Conserver la possibilité de retraiter les données brutes pour (1) appliquer, tester, comparer de nouvelles méthodes (statistiques, intelligence artificielle, …) qui pourraient être conçues dans le futur ; (2) corriger ou améliorer un traitement de données, par exemple en cas d'erreur dans un logiciel qui serait découverte plus tard. Il parait futile et risqué de ne pas préserver les données brutes par seule application d'une règle générale.

Volume de données	F2	Dans le cas de DataSuds, la taille de *chaque fichier* déposé ne doit pas dépasser 975 Mégaoctets. Cette limite est à considérer *après compression* éventuelle des fichiers. La limite est arbitraire, définie dans la configuration de DataSuds et il est possible de l'augmenter. Un jeu de données peut 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 De plus, si le volume total de données dépasse 200 Mo répartis dans de nombreux fichiers, leur téléchargement groupé ne sera pas possible. Il est alors conseillé de rassembler les fichiers dans une archive ZIP comme indiqué ci-dessous, ou peut-être de les déposer dans un autre entrepôt. Immédiatement après avoir été déposés, les fichiers ZIP sont automatiquement décompressés par l'entrepôt. Leur contenu est extrait et présenté aux utilisateurs sous forme d'une liste de fichiers qui peuvent être téléchargés individuellement ou de manière groupée. *Astuce* : si l'on souhaite diffuser directement des données au format ZIP, créer (puis déposer) un fichier ZIP temporaire qui contient lui-même le fichier ZIP qui sera présenté tel quel 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 maximum 100 fichiers par jeu de données. Au-delà et si nécessaire, diffuser les fichiers de manière regroupée dans des archives au format ZIP, suivant l'*astuce* décrite ci-dessus. Pour déposer plusieurs fichiers en une seule passe, les regrouper également dans une archive ZIP. A l'opposé, 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…	Présenter des centaines de fichiers téléchargeables individuellement est en général sans intérêt et peu pratique. Par défaut, l'entrepôt affiche 10 fichiers par page ; après un réglage manuel à répéter à chaque fois, au plus 50 fichiers par page.
Noms des fichiers	F4	Adopter des noms de fichiers spécifiques au projet et au contenu. Autant que possible, leur ajouter un préfixe en lien avec le projet (un acronyme ou un identifiant) qui sera commun à tous les fichiers déposés, tel que « PROJET_ABC_stationN_dailyflow.csv ». Eviter les noms de fichiers 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 (sauf dans le cas des fichiers Excel) : ouvrir le formulaire d'édition des métadonnées du fichier. Par défaut, Dataverse affiche la liste des fichiers triée par ordre alphabétique : prévoir l'ordre dans lequel les fichiers sont présentés aux utilisateurs et de les organiser, par exemple en insérant des numéros dans les noms de fichiers.	Eviter 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 préférable.
	F5	Dans les noms de fichiers, utiliser uniquement des lettres, chiffres et caractères séparateurs (**-** ou **_**). Remplacer les autres caractères (spéciaux, accentués, parenthèses…) et les espaces.	Principe FAIR « 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 TSV/TAB (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 De plus, afin de rendre possible leur archivage sur le long terme (un service actuellement en projet), les données devraient être déposées dans l'un des 64 formats actuellement acceptés par le CINES, dont la liste figure ici : https://facile.cines.fr. Les données déposées dans d'autres formats sont acceptées dans DataSuds mais elles ne pourront pas être archivées sur le long terme, à moins de disposer ultérieurement des ressources et outils nécessaires à la conversion des fichiers. Dans le cas de fichiers au format texte, utiliser de préférence l'encodage de caractères UTF-8 (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 ». Attention : dans la version française de Windows, les fichiers CSV enregistrés avec Excel utilisent le caractère point-virgule comme délimiteur de colonnes. L'application Dataverse s'attend à ce que le caractère délimiteur soit une virgule, conformément au standard international : elle ne reconnait pas correctement le contenu du fichier. Pour remédier à ce problème, enregistrer au format « *Texte (séparateur : tabulation)* » dont le caractère délimiteur est réellement standardisé, puis renommer l'extension .TXT du fichier en .TSV ou .TAB. Autre solution : produire le fichier TSV/TAB à l'aide du logiciel gratuit LibreOffice qui permet de choisir le caractère délimiteur aussi bien à l'ouverture qu'à la sauvegarde d'un fichier, grâce à l'option « Editer les paramètres du filtre ». Au cours de la conversion d'un fichier Excel, prêter attention à la présence de plusieurs pages/onglets, produire 1 fichier TSV/TAB par page. Si la conversion des données dans un format ouvert occasionne une perte d'information, déposer également le fichier original en complément du fichier converti. Exemple d'un tableau Excel converti au format CSV : perte du formatage, des couleurs, de la mise en page, des formules et macros.	Principe FAIR « Interopérable » Formats de données pérennes : assurer la conservation et la lisibilité des données sur le long terme.
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. Solution alternative : déposer un sommaire (fichier texte) qui résume le contenu de chaque fichier ou groupe de fichiers.	Aider les utilisateurs à identifier le contenu de chaque fichier et à trouver ce qu'ils cherchent.
	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 au préalable l'interface en anglais afin de ne pas utiliser la version française des libellés.	Catégorisation des fichiers déposés, donne la possibilité de les trier par type.
Noms de dossiers	F10	Définir un nom de dossier pour chaque fichier déposé. A minima, saisir un nom de dossier racine, commun à tous les fichiers, correspondant au nom ou à l'acronyme du projet de recherche. Pour les utilisateurs, il deviendra le « dossier d'installation des données », de façon similaire à un logiciel. Les noms de dossier (« Chemin d'accès au fichier ») peuvent être saisis manuellement dans le formulaire « Métadonnées » disponible au niveau des fichiers. Il est possible de créer une arborescence complète 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 le nom de dossier associé à chaque fichier sera prérempli. Inclure également un dossier racine dans le fichier ZIP déposé.	Organisation des fichiers déposés. Lorsqu'un utilisateur sélectionne plusieurs (ou tous) fichiers à télécharger, l'entrepôt crée une archive ZIP qui inclut les différents dossiers qui ont été définis. Cela permet aux utilisateurs de recréer automatiquement l'arborescence des dossiers et de préserver l'organisation des fichiers. Cela évite 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 (TSV/*TAB*, CSV, Excel…), il est essentiel de préciser la signification de chaque variable, « champ » ou « colonne ». Cette information est en général documentée sous la forme d'un dictionnaire des données (*data dictionary*) à déposer également dans l'entrepôt. Un dictionnaire des données est constitué d'un ou plusieurs tableaux de référence au format TSV/CSV/texte avec les colonnes suivantes : « nom de la variable », « contenu/signification de la variable », « unité de mesure », « signification des différents codes utilisés » (si applicable) ; « format » (facultatif). Il peut être commun à plusieurs tables ou fichiers ; il est aussi parfois intégré dans un fichier de données. Si de nombreux codes sont utilisés, il est conseillé de documenter leur signification dans des tableaux de référence (*codebook*).	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 augmente le risque de mauvaise interprétation des données. Attacher ces informations avec les données participe à rendre les données réellement réutilisables sur le long terme.
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 de leur production. Le déposant peut décider de déposer certaines documentations en accès restreint dans le seul but de les préserver avec les données. Voici des suggestions de documentations qui peuvent accompagner des données : fiche de présentation du projet de recherche (éventuellement sous forme de diapositives) ; 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 ; protocole de l'étude ; accord d'un comité d'éthique. Une exception : les articles scientifiques (dans toutes leurs versions, *preprints* inclus) 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) Sur le long terme, améliorer le potentiel de réutilisation des données, ainsi que l'autonomie des utilisateurs : réduit les besoins de contacter le producteur des données pour demander des précisions. 2) Le cas échéant, aider les utilisateurs à comprendre pourquoi ils ne parviennent pas à « reproduire l'expérience », par exemple s'il s'avère que leurs paramètres ou conditions expérimentales sont différents.
	F13	Insérer la référence bibliographique (citation complète ou DOI) du jeu de données dans une ou plusieurs documentations téléchargeables, déposées avec les données. Le plus simple consiste à créer et déposer un fichier texte « *readme* » dans lequel sont copiés la référence du jeu de données, le contenu du champ « Description » et ses conditions d'utilisation (licences) ; attendre que ces éléments soient finalisés. Dès la préparation d'un jeu de données, sa future citation est connue à l'avance. Le DOI est pré-attribué et 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 ».	Aider les utilisateurs à conserver trace de la source des données, après téléchargement. Inciter à la citation des données, qui peuvent être réutilisées longtemps après avoir été obtenues. Faciliter les accès ultérieurs au jeu de données, par exemple dans le but de vérifier s'il existe des mises à jour.
	F14	(Facultatif) Si l'on dispose d'illustrations en lien avec le jeu de données, en choisir une et l'intégrer en tant que vignette. Exemples : logo du projet ; photo, zone extraite d'une photo ; figure.	Rendre l'entrepôt plus agréable à consulter. Donner graphiquement un aperçu de la diversité des projets de recherche.

Conditions d'utilisation

Elément	#	**Préconisations** ou/et **recommandations** - Conseils pratiques	Finalité et commentaires
Attribution de licences, ou de conditions d'accès et de réutilisation	L1	Si les données sont ouvertes, le formulaire en ligne https://creativecommons.org/choose/ peut aider à choisir un modèle standard de licence. Cette licence devra être acceptée et respectée par les utilisateurs qui téléchargent puis utilisent les données. Tenir compte des exigences éventuelles des partenaires et des financeurs du projet, ainsi que de la nécessité de préserver la confidentialité des données dans certaines situations : respect des réglementations, déontologie et éthique (lire également la section F1). S'il s'agit de code informatique, choisir un modèle de licence adapté, tel que GPL ou MIT. Les licences de type CC-BY-ND peuvent convenir pour un document, mais sont déconseillées pour des données : elles n'autorisent pas les travaux dérivés, donc les réutilisations. Il est possible d'attribuer une licence différente à chaque fichier ou à un groupe de fichiers, par exemple une licence CC-BY aux documentations qui accompagnent les données. Dans ce cas, préciser distinctement à quels fichiers chaque licence s'applique. Par contre, attribuer plusieurs licences aux mêmes contenus (*multi-licensing*) peut poser problème : les exigences des licences sont parfois contradictoires ; de plus, les utilisateurs peuvent alors choisir la licence qui leur convient. Par défaut, l'entrepôt DataSuds attribue une licence CC-BY (ou parfois CC0 = domaine public) aux jeux de données. Consulter l'onglet « Conditions d'utilisation » du jeu de données et vérifier que la licence correspond bien à ce qui est souhaité. La licence CC-BY demande à ce que le jeu de données soit cité lorsqu'il est utilisé ; elle est conforme aux préconisations du MESRI pour les projets sur financement public. Des exigences de citations peuvent être précisées dans un champ dédié. Si les modèles de licence standards (*Creative Commons* ou autres) ne conviennent pas, indiquer les conditions particulières à remplir pour accéder aux données et les réutiliser. Il peut s'agir de la nécessité de signer un accord. Un formulaire du type *Data Use Agreement* peut être préparé et mis à disposition dans l'entrepôt : contacter data@ird.fr pour obtenir un modèle. Fournir des informations de contact (si possible pérennes : pas une unique personne) pour la soumission des demandes d'accès aux données.	Principe FAI**R** « Réutilisable » : répondre à la question « Sous quelles conditions est-t-il possible de réutiliser ces données ? » Rappel : « Aussi ouvert que possible, aussi fermé que nécessaire », d'abord respecter les exigences réglementaires, éthiques et contractuelles. En cas de changement de licence après publication des données, la nouvelle licence ne s'appliquera pas rétroactivement aux utilisateurs qui ont déjà obtenu les données, à plus forte raison s'ils ont conservé une preuve de la licence antérieure qui leur a été accordée. Attention : les licences ouvertes *Creative Commons* sont non-révocables. La fiche de présentation du jeu de données est diffusée implicitement sous licence CC0 afin de faciliter son exploitation par les moteurs de recherche.
Formulation et cohérence des conditions d'accès et de réutilisation (licences)	L2	Si une licence est attribuée, sa formulation (dans la même langue que les autres métadonnées), son logo et les liens internet éventuels doivent être complets, reproduits dans leur intégralité. Pour réduire le risque d'erreur, copier la licence à partir de son modèle original, sans la modifier. Si des conditions particulières sont imposées pour l'usage de certaines données, la licence doit être cohérente vis-à-vis de ces dispositions. Dans le cas de la licence CC-BY, les utilisateurs qui obtiennent l'accès aux données disposent du droit de les rediffuser librement par eux-mêmes. Ce type de licence ne convient donc pas aux fichiers dont l'accès est restreint : ne pas attribuer de licence ouverte à des données fermées, sauf en cas d'embargo temporaire. De plus, il n'est pas autorisé d'ajouter des conditions supplémentaires aux licences *Creative Commons*, à moins de ne plus faire mention ou référence à « *Creative Commons* ». Si les données sont disponibles par ailleurs, tel que sur le site web du projet, la licence doit être cohérente et compatible avec les indications données par le site qui héberge les données.	Validité juridique des conditions d'utilisation.

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

Version 3.0 (09/2023) – Luc Decker, Service IST/MCST, IRD, France

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

Elément à vérifier	OK	Edité	Demander au déposant	Améliorer Requis	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 : précision et 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 sources de support	☐	☐	☐	☐	☐	☐
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. Description de chaque fichier, ou sommaire des fichiers	☐	☐	☐	☐	☐	☐
F9. Attribution de catégories (*tags*/libellés) aux fichiers	☐	☐	☐	☐	☐	☐
F10. Saisie de nom(s) de dossier(s)	☐	☐	☐		☐	☐
F11. Dictionnaire des données (variables, unités, codes)	☐	☐	☐	☐	☐	☐
F12. Dépôt de documentations associées	☐	☐	☐	☐	☐	☐
F13. Citation du jeu de données insérée dans la documentation	☐	☐	☐		☐	☐
F14. Vignette du jeu de données : logo, photo, …	☐	☐	☐		☐	☐
L1. Choix de licence(s) d'utilisation des données	☐	☐	☐	☐	☐	☐
L2. Formulation, cohérence des licences/conditions d'utilisation	☐	☐	☐	☐	☐	☐
	OK	Edité	Demander au déposant	Améliorer Requis	Améliorer Conseillé	Non applicable

Guide de relecture d'un jeu de données avant publication
Annexe 1. Conseils de mise en œuvre

Pour commencer…
Accuser bonne réception de la demande de relecture du jeu de données. Si le déposant l'a transmise via le bouton dans l'interface du logiciel Dataverse : mentionner que dans l'intervalle, le déposant ne pourra plus modifier le jeu de données, c'est le fonctionnement normal du logiciel. Si le déposant ne l'a pas précisé, demander s'il s'agit bien de *publier* le jeu de données et dans quel délai. En effet, la publication des données est parfois différée : dans l'attente de la publication d'un article associé ou pour d'autres raisons voulues par le déposant. Dans ce cas, une alternative est de publier les métadonnées et de conserver les fichiers sous embargo. Si un jeu de données doit être transmis à des *reviewers* ou à un éditeur à l'aide de son lien d'accès privé, il est important qu'il soit finalisé. A l'inverse, il s'agit parfois de demandes pour recueillir un premier avis.

Sauvegarder le jeu de données avant et après éditions
Par sécurité, exporter la fiche descriptive (toutes les métadonnées) dans un fichier texte : à partir de l'onglet "Métadonnées", sélectionner puis copier/coller tout le texte. En effet, le logiciel Dataverse ne permet pas de repérer/tracer les changements réalisés sur un jeu de données durant sa préparation (version *draft*). Ce fichier texte pourra être chargé dans un logiciel de comparaison de versions, tel que *WinMerge*. Si le déposant conserve la possibilité de modifier le jeu de données : lorsque l'édition des métadonnées est terminée, les sauvegarder à nouveau dans un 2ème fichier texte qui permettra alors de détecter facilement et de relire les derniers changements avant publication. Enfin, il est souvent utile de télécharger les fichiers déposés pour en examiner le contenu, en fonction de leur type, nombre et volume.

Pourquoi et comment utiliser une version papier de la grille de la relecture ?
Si l'on dispose d'un seul écran, cette méthode permet de consulter simultanément la grille et le jeu de données à relire. Le papier apporte de l'agilité manuscrite, la possibilité d'annoter, surligner ou entourer rapidement des éléments, à l'aide de différentes couleurs. Pour limiter l'impact environnemental, la grille tient sur une seule page dont le verso peut servir à prendre davantage de notes. Imprimer la grille à l'échelle 90% ou 80% offre plus d'espace pour annoter dans les marges. Conserver la grille complétée durant quelque mois à des fins de traçabilité.

Comment appliquer la grille de relecture ?
1. Lorsqu'un élément convient déjà entièrement, cocher la case "OK".
2. Lorsqu'un élément n'est pas pertinent dans le cas présent, cocher la case "Non applicable".
3. Lorsqu'un élément semble à améliorer :
    - [Facultatif] Pour prioriser, cocher la case "Requis" ou "Conseillé" selon que cela semble essentiel ou non.
    - Si le relecteur/curateur dispose des informations nécessaires, il édite par lui-même le jeu de données dans DataSuds, coche la case "Edité" - et enfin la case "OK" si à présent tout lui semble satisfaisant.
      Exemples de tels changements : modifications qui portent sur la forme ou l'apparence des informations ; insertion de balises HTML ; saisie ou mise à jour de champs de métadonnées à partir de détails déjà présents dans d'autres champs ; insertion de liens vers des sites web (projets, sources de financement, origine des données, …) ; corrections typographiques et orthographiques ; autres corrections évidentes ; déplacement d'information entre 2 champs de métadonnées ; conversions simples de format de fichier ; création et ajout d'un fichier "readme".
    - Dans le cas contraire, cocher la case "Demander au déposant". Après réception et prise en compte de sa réponse, le relecteur/curateur coche ultérieurement les cases "Edité" ou/et "OK".

- De manière générale, les différentes cases ne sont pas exclusives. Par exemple, un même élément peut être tout à la fois édité, à vérifier auprès du déposant et requis.
- On commence souvent par parcourir rapidement un nouveau jeu de données. Si certains points à améliorer sont décelés immédiatement : comme aide-mémoire, cercler les cases correspondantes (« à faire ») avant de les traiter dans un second temps.

Communiquer avec le déposant après relecture et éditions ?
A partir des informations saisies dans la grille, rédiger un compte-rendu synthétique à l'intention du déposant : d'une part lister chaque élément édité et résumer la nature du changement, ainsi que la raison si cela parait nécessaire ; d'autre part, énumérer les questions, suggestions et autres points restant à vérifier. Faire en sorte que le compte-rendu soit concis et structuré (sous-titres, listes) ; l'envoyer par email au déposant. Communiquer de manière positive, parler d'amélioration au lieu de "problème à corriger". Pour son prochain jeu de données, le déposant sera-t-il enclin à utiliser à nouveau l'entrepôt ?

Voici un exemple de compte-rendu, inspiré de cas réels :

Comme annoncé, j'ai complété la relecture et l'édition du jeu de données https://dataverse.....
Je vous invite à consulter le compte-rendu ci-dessous.

Edition des métadonnées
- Titre : ajout d'une précision sur la période de collecte des données "(1952-1992)"
- Auteurs : noms en minuscules ; affiliations complétées suivant le format recommandé ; ajout d'un identifiant ORCID pour….
- Description : mise en page ; insertion d'un sous-titre "…." ; corrections orthographiques : "…." ; liens rendus cliquables.
- Période de collecte : saisie à partir des informations disponibles dans la description.
- Langue : saisie de "English".

Fichiers déposés
- Création et ajout d'un document texte « readme » : une copie téléchargeable de la citation et de la description des données.
- Ajout de labels « Data » et « Documentation ».
- Saisie d'une description pour le fichier "…."
- Conversion du fichier Excel "…." au format TAB (texte tabulé), ajout de cette version en complément du fichier original.

Conditions d'utilisation
- Licence par défaut CC-BY remplacée par CC-BY-SA, suite à nos échanges et selon votre choix.
- Mention de conditions particulières pour le fichier "…." dont l'accès est restreint.

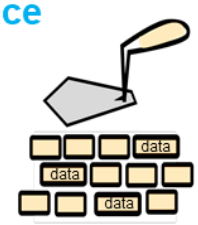
Recommandations, suggestions d'amélioration
1) Dans le champ "Publication associée", saisir le titre de l'article en préparation, ou sa future référence bibliographique, même incomplète, avec la mention [submitted]
2) Comme pour un article, remplir le champ "Renseignement sur la subvention" : citer/remercier les sources de financement et de support de l'étude. Si vous me transmettez l'article associé, je peux me charger de saisir l'information.
3) Comme pour un article, mentionner l'accord de comité(s) d'éthique ou/et les modalités de consentement des participants, dans le champ "Description" ou le champ "Remarques".
4) Possibilité d'ajouter des liens vers des pages web en lien avec le projet.

Adapter la méthode en cas d'urgence à publier un jeu de données ?
Proposer d'éditer, finaliser et publier le jeu de données en présence du déposant : organiser un rendez-vous ou une visioconférence. Cette façon de procéder rend les échanges plus fluides (le déposant peut immédiatement répondre aux questions et valider les changements) et évite de devoir rédiger un compte-rendu. De façon pragmatique, sélectionner ce qui peut être amélioré au mieux dans le temps limité dont on dispose. Se concentrer sur les éléments les plus importants : titre et description du jeu de données, liste des auteurs, fichiers à mettre en accès restreint ou à retirer pour des raisons légales ou éthiques, conditions d'utilisation et licences (rappel : les licences Creative Commons ne sont pas révocables). Considérer les améliorations faciles à réaliser, qui ne demandent que quelques secondes ou minutes supplémentaires, par exemple les labels "Data"/"Documentation" attribués aux fichiers, la saisie de certains champs de métadonnées tels que "Période couverte", "Renseignement sur la subvention", le format des noms des auteurs.

Guide de relecture d'un jeu de données avant publication
Annexe 2. Présentation introductive : **« Pourquoi et comment publier un jeu de données dans une démarche guidée par la qualité ? »**

Pourquoi publier un jeu de données dans une démarche guidée par la qualité ?

- **Rendre les données « Faciles à trouver, Accessibles, Interopérables, Réutilisables »** *(principes FAIR)*
  - ► **La démarche de partage des données devient pleinement utile et efficace, pour la Science**
- **Contribuer à une meilleure prise en compte des publications de données dans les évaluations de la recherche** *(principes Data Citation)*



- **Préserver & renforcer la réputation** des chercheurs, des laboratoires, de l'institut …et de l'entrepôt

1

Comment publier un jeu de données dans une démarche guidée par la qualité ?

- **Données FAIR : Faciles à trouver, Accessibles, Interopérables, Réutilisables**
  - ► Se mettre à la place d'un utilisateur des données
    - ✓ Qualité des métadonnées
    - ✓ Documentations
    - ✓ Fichiers : formats ouverts

L'entrepôt de données s'occupe déjà du respect d'autres exigences : identifiants uniques, stockage pérenne, standards

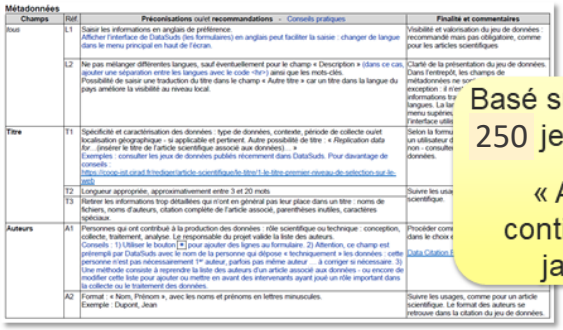
- **Réputation** (image) de l'institut, du laboratoire, des chercheurs et de l'entrepôt
  - ► Se mettre à la place d'un visiteur de l'entrepôt

Qualité éditoriale des métadonnées
- ✓ Contenu intelligible
- ✓ Contenu complet et précis (lieux, périodes…)
- ✓ Mise en forme, présentation, illustrations
- ✓ Grammaire, orthographe, anglais

Respect des règlementations ; déontologie et éthique

2

Guide de dépôt d'un jeu de données dans l'entrepôt *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 so... exception : il n'en... informations tra... langues. La lan... menu supérieu... l'interface utilis...
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 forma... un utilisateur d... non - consulte... données.
	T2	Longueur appropriée, approximativement entre 3 et 20 mots	Suivre les usa... 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 com... dans le choix d... Data Citation E...
	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.

Basé sur la révision de 250 jeux de données

« Amélioration continue » depuis janvier 2021

https://doi.org/10.5281/zenodo.5884670

- ➔ Destiné aux curateurs, qui relisent, vérifient et parfois éditent les jeux de données avant procéder à leur publication
- ➔ Tableaux : critères et justifications, *checklist*
- ➔ Egalement proposé aux déposants, pour anticiper la relecture, finaliser plus rapidement les jeux de données
- ➔ Evolution du guide à partir des situations rencontrées, parfois inattendues

Comment publier un jeu de données dans une démarche guidée par la qualité ?

Questions pratiques

- Combien de temps pour relire & éditer un jeu de données ?
- Comment procéder si le besoin de publier est urgent ?
- Formation initiale et expérience acquise avant de recevoir le droit de publier de manière autonome ?
- Cas particulier des données sensibles ?
- Où demander conseil ? **data@ird.fr**

Guide de relecture d'un jeu de données avant publication
Annexe 3. Ressources & liens divers

- **Mode d'emploi officiel de *Dataverse* en anglais**, référence complète : https://guides.dataverse.org/

- **Mode d'emploi de *Dataverse* en français : site IRD Data** ; présentation de l'API Dataverse

- **Cadre juridique pour la publication de données** : https://data.ird.fr/cadre-juridique

- **Supports de formation "Entrepôts de données"** (Open-OPSE, 2022 : 5 présentations PowerPoint, 7h de formation, 90 diapos : https://doi.org/10.5281/zenodo.5724006

- **Logiciels gratuits** pour aider à la préparation et la curation des données

Notepad++	Visualiser/éditer le contenu de tout fichier texte, convertir en UTF-8
LibreOffice Calc	Données tabulées : convertir des fichiers CSV, modifier les séparateurs,
WinMerge	Comparer des textes, tels que 2 versions d'une fiche de métadonnées
PDFsam Basic	Assembler des documents PDF
FastStone Image Viewer	Redimensionner/recadrer des images pour illustrer collections et jeux de données
ARX	Anonymiser des données à caractère personnel (après formation)

- Auto-évaluation de la connaissance des principes FAIR : outil **FAIR-Aware**

- Moissonnage de l'entrepôt DataSuds par l'entrepôt national *Recherche Data Gouv :* collection IRD